



UNIVERSIDAD DE MURCIA

DEPARTAMENTO DE INFORMÁTICA Y SISTEMAS

**Minería de Opiniones Basada en Características
Guiada por Ontologías.**

D. Isidro Peñalver Martínez
2015

Dedicatoria

A mi padre, Pedro, por sus sabios consejos y por apoyarme siempre.

A mi madre, Teresa, por su paciencia infinita y facilitarme todo.

A mis hermanos, Mayte, Ismael y Rubén, por alegrarme la vida.

A mis amigos, por estar siempre a mi lado, en los buenos y malos momentos.

A Rafa y Paco, mis directores, por mostrarme el camino y ayudarme.

Agradecimientos

Cuando me decidí a llevar a cabo esta tesis doctoral, uno de los principales problemas a los que debía enfrentarme fue la falta de tiempo para llevar a cabo todas las investigaciones que tenía en mente, ya que tenía que compaginar el trabajo de investigación y experimentación necesarios con mi vida laboral fuera del ámbito de la Universidad.

En los primeros años, tras hacer los cursos de doctorado y superar los exámenes y la suficiencia investigadora, no tenía muy claro hacia que rama de la ciencia enfocar mis investigaciones, ya que existen distintas ramas de la Inteligencia Artificial que son de mi interés.

Tras unos primeros sondeos en el estado del arte de algunas de las ramas de la Inteligencia Artificial, junto a los conocimientos adquiridos durante los cursos de doctorado y una conversación mantenida con Rafa, mi director de tesis, me decidí por enfocar todos mis esfuerzos de investigación en la rama de la minería de opiniones y el estudio de las ontologías.

Agradezco a Rafa que en aquella conversación mantenida en su despacho tuviese la habilidad y el conocimiento necesarios para despertar en mi la curiosidad y motivación imprescindibles para abordar con éxito un trabajo de esta envergadura. Desde el principio me mostró con claridad el problema a resolver y me facilitó las guías de información necesarias para enfrentarme al problema con éxito y poder proponer una solución de calidad.

A partir de este momento mis esfuerzos se centraron en el estudio del estado del arte de las tecnologías implicadas, la lectura de artículos relacionados con el problema que yo pretendía resolver, y la redacción de artículos propios en la línea de investigación elegida.

Agradezco a mis padres, y a mi familia en general, todo el apoyo recibido. Por estar a mi lado en todo momento y por animarme a aventurarme en esta empresa. Gracias a ellos, en buena medida, he podido culminar con éxito mis investigaciones y redactar la presente tesis doctoral.

Agradezco también a Paco, mi segundo director de tesis, que junto con Rafa haya trabajado concienzudamente en las revisiones de los artículos y en las correcciones a realizar en la redacción de la presente tesis doctoral. Sin la dedicación y eficacia mostrada por ambos, junto con sus profundos conocimientos sobre el estado del arte de las tecnologías implicadas en la solución del problema, la redacción de la presente tesis doctoral no hubiese sido posible.

Agradezco a mis amigos de Murcia, y en esta última etapa a mis amigos de Madrid, el apoyo que he recibido por su parte en todo momento. Gracias por estar a mi lado en los buenos y en los malos momentos, sin pedirme nada a cambio, y ofreciéndome lo mejor de cada uno de vosotros.

Agradezco a Mari Ángeles, licenciada en periodismo, su esfuerzo desinteresado para ayudarme con la revisión final de la tesis y aportar un punto de vista distinto del de las personas que llevamos tiempo trabajando en este campo.

Por último me dirijo a todas aquellas personas que estén pensando en llevar a cabo un trabajo de estas características, o estén inmersos en él y vean difícil el poder llegar a cumplir los objetivos marcados. Me gustaría decirles que no existe “fórmula mágica” y que solamente con disciplina y trabajo se puede llegar al éxito. Que pongan toda su energía y entusiasmo en culminar su trabajo si realmente eso les hace felices, porque, al final, esto produce una satisfacción personal inmensa.

Con estas palabras doy comienzo a esta tesis doctoral, esperando que ésta sea de utilidad a una gran cantidad de investigadores que, como yo, han tenido las ganas y el interés en crear algo que pueda ayudar a los demás.

Resumen

Resumen general de toda la tesis

La participación de los usuarios es el principal objetivo de la Web 2.0. Esta idea tan simple ha tenido un tremendo impacto en la forma en la que los usuarios interactúan en la red. Mientras que en los sitios Web 1.0 las compañías publicaban contenidos y los usuarios eran meros consumidores de información, en la era 2.0 los usuarios juegan un papel más activo en las interacciones con la Web. Ya no son sólo consumidores de información sino que también la producen. El número de opiniones online en las que se expresan pensamientos acerca de una gran variedad de temas se encuentra en constante crecimiento. Los contenidos producidos por la participación de los usuarios en sitios Web como *Amazon*¹, *booking*², o *IMDb*³ son útiles tanto para los proveedores de información como para los consumidores.

La **minería de opiniones** es una reciente sub-disciplina entre la recuperación de información y la lingüística computacional. Esta sub-disciplina no trata de detectar el tema abordado en un texto sino las opiniones y sentimientos expresados en el mismo (Esuli & Sebastiani, 2005). Determina si los comentarios en foros online, blogs o redes sociales sobre un tema en particular (producto, libro, película, etc.) son positivos, negativos o neutros. Las opiniones son muy importantes cuando alguien desea conocer la opinión de los otros antes de tomar una decisión.

Tradicionalmente se pueden distinguir dos metodologías principales en minería de opiniones. Por un lado hay enfoques para minería de opiniones que tienen como objetivo clasificar documentos enteros con polaridad positiva o negativa. Por otro lado hay enfoques que se centran en distinguir las frases con contenido de sentimiento dentro de los textos de opinión de los usuarios. El inconveniente de ambas metodologías es que un mismo documento o una sola frase podrían contener diferentes opiniones sobre las distintas características de un producto o servicio (Cambria et al., 2013). De hecho la clasificación de opiniones a nivel de documento o a nivel de frase no indica lo que le gusta o le disgusta al usuario. Un documento positivo sobre un objeto no significa que el usuario tenga una opinión positiva sobre todos los aspectos o características de ese objeto. Asimismo, es imposible asegurar que un documento clasificado como negativo significa que al usuario no le gusta nada de ese objeto. En un documento (por ejemplo, una opinión sobre un producto) el usuario normalmente escribe sobre los aspectos positivos y negativos del objeto, aunque el sentimiento general hacia ese objeto puede ser positivo o negativo (Ahmad & Doja, 2012). Con el objetivo de identificar las características contenidas en una opinión y clasificar los sentimientos de esas

¹ www.amazon.com, sitio Web para la compra de productos online.

² www.booking.com, sitio Web para realizar reservas de hoteles.

³ www.imdb.com, *Internet Movie Database*, sitio Web con información sobre películas cinematográficas.

características surge el concepto de **minería de opiniones basada en características** (Feldman, 2013). La minería de opiniones basada en características sobre opiniones de productos es una tarea difícil porque existe una gran variabilidad semántica de las opiniones expresadas y una enorme diversidad de características que describen a los productos. Además, existe una gran cantidad de palabras que se pueden utilizar para hablar sobre los productos en una opinión (Balahur & Montoyo, 2010).

En los últimos años han ido apareciendo nuevos enfoques basados tanto en corpus dependientes del dominio como en tecnologías de la Web Semántica para la minería de opiniones basada en características (Cambria et al., 2013).

En la actualidad las tecnologías de la Web Semántica están alcanzando un cierto grado de madurez (Shadbolt et al., 2006). La Web Semántica fue concebida con el objetivo de añadir semántica a los datos publicados en la Web, es decir, establecer el significado de los datos permitiendo así que las máquinas sean capaces de procesarlos de una manera similar a la de los seres humanos (Berners Lee & Hendler, 2001). Las tecnologías semánticas se han aplicado con éxito en varias tareas de procesamiento del lenguaje natural como la recuperación de información (Gladun et al., 2013), la adquisición de conocimientos a partir de textos (Ochoa et al., 2013) o la anotación semántica (Rodríguez-García et al., 2014). Una vez que las tecnologías de la Web Semántica han madurado pueden ser una adición valiosa a los enfoques tradicionales de minería de opiniones. Las ontologías son un componente clave dentro de la Web Semántica que permiten representar el conocimiento subyacente en un dominio concreto. La semántica formal que subyace en los lenguajes ontológicos permite el procesamiento automático de la información representada en forma de ontologías y habilita el uso de razonadores semánticos para inferir nuevo conocimiento. En este trabajo una ontología se define como *“una especificación formal y explícita de una conceptualización compartida”* (Studer et al., 1998). Las ontologías proporcionan una representación del conocimiento estructurado y formal, con la ventaja de ser reutilizable y compartible. También proporcionan un vocabulario común dentro de un dominio y definen, con diferentes niveles de formalidad, el significado de los atributos de los términos y las relaciones entre ellos. En este trabajo se ha utilizado el lenguaje de ontologías OWL (*Web Ontology Language*), el estándar propuesto por la W3C (*World Wide Web Consortium*) para la representación de ontologías en la Web Semántica con vistas a representar los conceptos y características del dominio. El conocimiento en ontologías OWL se formaliza principalmente mediante el uso de cinco tipos de componentes: clases, propiedades de tipo de datos, propiedades de objeto, axiomas e instancias.

Resumiendo, tal y como se ha visto, actualmente la minería de opiniones es un campo muy activo que engloba el procesamiento del lenguaje natural, la lingüística computacional y las técnicas de análisis de textos con el objetivo de extraer valor añadido y elementos de información de las opiniones de los usuarios. Sin embargo los enfoques existentes actualmente poseen una serie de inconvenientes derivados de la ausencia de relaciones semánticas en el proceso de búsqueda de características, o de la falta de empleo de métodos matemáticos avanzados en el proceso de análisis de sentimientos en las opiniones.

El objetivo de esta tesis es describir un nuevo método de minería de opiniones basado en características que utiliza una ontología del dominio para la identificación de características. Se aporta un nuevo algoritmo basado en análisis vectorial en el espacio R^3 para llevar a cabo el proceso de análisis de sentimientos de las opiniones. El método ha sido validado en el dominio de las opiniones de los usuarios sobre películas cinematográficas y en el dominio de las finanzas, obteniendo resultados muy prometedores.

La tesis está organizada de la siguiente manera: en el Capítulo I se realiza un estudio del estado del arte y se presentan los trabajos relacionados. En el Capítulo II se explica en detalle el método propuesto con figuras y ejemplos que facilitan su comprensión. El Capítulo III expone las herramientas utilizadas para construir la solución y un manual de uso del software construido. El Capítulo IV muestra la validación de dicho método en el dominio de las películas cinematográficas y en el dominio de las finanzas. Por último las conclusiones y trabajos futuros se exponen en el Capítulo V.

Introducción

Introducción general de toda la tesis

Actualmente la sociedad está experimentando profundos cambios como resultado del rápido desarrollo de las Tecnologías de la Información y la Comunicación (*TIC*). El modo de vida de las sociedades de todo el mundo está cambiando debido a la facilidad con la que cualquier persona, con una conexión a Internet, puede acceder a las grandes cantidades de información que se producen en la Web (Balahur & Montoyo, 2010). De hecho, la era en la que vivimos ha sido bautizada como la “*sociedad del conocimiento*” o la “*era de la información*”. Esta sociedad de la información propicia que las compañías e individuos utilicen mecanismos de búsqueda, recuperación y procesamiento de datos relevantes para obtener conocimiento que puedan utilizar en su beneficio. Es en este punto donde juegan un papel fundamental las técnicas de Inteligencia Artificial (*IA*) que proporcionan mecanismos de extracción de conocimiento. La exactitud y fiabilidad del conocimiento extraído será un aspecto clave en todos los ámbitos de la actividad humana. Esta es la meta que ha motivado a una gran cantidad de investigadores en distintas disciplinas a desarrollar un conjunto de metodologías, herramientas y sistemas que faciliten esta labor.

Sin embargo, en la mayoría de ocasiones, toda esta cantidad de información no se encuentra en fuentes estructuradas (tablas o bases de datos) sino en documentos no estructurados escritos en lenguaje natural. El presente está marcado por los deseos de las personas por compartir sus experiencias, conocimientos y pensamientos a través de redes sociales, blogs, foros o sitios Web que dan la oportunidad de difundir opiniones sobre los productos o servicios que ofertan. La gente está generando continuamente opiniones sobre noticias, productos o servicios que aparecen a diario en redes sociales. En este contexto, actualmente se dispone de un gran volumen de datos en la Web que contienen opiniones y emociones de las personas en los distintos ámbitos que componen la sociedad, ya sea el ámbito político, económico o social. Por lo que se hace indispensable el uso de técnicas de tratamiento automático de la información contenida en estos medios para poder extraer conclusiones relevantes que ayuden a las empresas y a las personas a tomar mejores decisiones en su día a día.

La disciplina que se ha ocupado del tratamiento automático del lenguaje natural o lenguaje humano se llama Procesamiento del Lenguaje Natural (*PLN*). El *PLN* es la parte de la *IA* que se define como “*la ciencia y la ingeniería de hacer máquinas inteligentes*” (McCarthy, 1959). Tradicionalmente las áreas de aplicación del *PLN* se diseñaron para el tratamiento de textos que describen hechos que se pueden observar y comprobar en la realidad. Sin embargo en la actualidad la información sobre hechos ya no es la principal fuente de la que extraer conocimiento, sino que la mayoría de textos

en la Web contiene expresiones con sentimientos y llenas de subjetividad. En este contexto, han aparecido nuevos campos de investigación dentro del *PLN* con el objetivo de analizar la subjetividad en los textos y/o extraer y clasificar los sentimientos de las opiniones expresadas por las personas en el ámbito social, político y económico. El área de investigación dentro del *PLN* que se encarga de la detección automática de los sentimientos expresados en los textos y su clasificación según la polaridad que tienen (positiva, negativa o neutra) es el área de ***análisis de sentimientos***, también llamada ***minería de opiniones***.

Por otro lado, la Web Semántica es una extensión más inteligente de la Web actual donde el significado de la información que contienen los recursos en la Web puede ser representados formalmente. Esta representación formal y explícita consigue que los contenidos sean comprensibles por sistemas de computación (Malik et al., 2010). En este sentido las ontologías proporcionan esta forma de representación del conocimiento constituyendo uno de los pilares fundamentales de la Web Semántica (Davis, 2013).

Las razones expuestas en párrafos anteriores con respecto a la necesidad de extraer conocimiento desde fuentes de información no estructuradas, de identificar y clasificar la polaridad de las expresiones lingüísticas cargadas de sentimientos en los textos de opinión y la ingente aparición de la Web Semántica, unida a la potencia semántica que ofrecen las ontologías, han sido las principales motivaciones para la realización de la investigación que se describe en esta tesis doctoral. La solución que se plantea en este trabajo de investigación se sustenta en el desarrollo de una nueva metodología de minería de opiniones basada en características y guiada por ontologías. Esta innovadora metodología cubre todo el ciclo de vida del análisis de sentimientos de opiniones desde el procesamiento del texto a nivel morfológico y léxico, la detección de las frases cargadas de contenido que expresa sentimientos o emociones, la identificación de las características relevantes en el texto, la clasificación de la polaridad de las características, y la clasificación general de las opiniones realizadas por los individuos según su polaridad (positiva, negativa o neutra). Además, la metodología propuesta facilita la adaptación continua del dominio objeto de estudio mediante la evolución continua de la ontología subyacente.

Para lograr los objetivos descritos en el párrafo anterior se ha seguido la siguiente metodología:

- **Análisis del estado del arte** en Web Semántica, Ontologías, Procesamiento del Lenguaje Natural, Espacios Vectoriales y Minería de Opiniones. Este estudio del

estado del arte ha implicado el estudio y análisis de los antecedentes de las tecnologías a incorporar en el proyecto.

- **Definición y formalización de una arquitectura general** para la minería de opiniones basada en características guiada por ontologías. Esta tarea se ha realizado mediante la definición e interconexión de módulos independientes que se comunican entre ellos para alcanzar el objetivo general del sistema. Los módulos definidos son los siguientes: (i) módulo de procesamiento de lenguaje natural, (ii) módulo de identificación de características, (iii) módulo de cálculo de polaridades de características y (iv) módulo de análisis de sentimientos de opiniones.
- **Diseño de una aplicación software** para realizar la minería de opiniones basada en características de manera automática. Se ha desarrollado una interfaz Web para facilitar la clasificación de sentimientos de opiniones de usuarios en un dominio cualquiera.
- **Validación de la metodología** de minería de opiniones basada en características guiada por ontologías en dos dominios bien diferenciados: el dominio de las películas cinematográficas y el de las finanzas. Se ha llevado a cabo la validación del módulo de identificación de características, del módulo de cálculo de polaridades de características y, por último, del módulo de análisis de sentimientos de opiniones. Para cada uno de estos procesos de validación se aplicaron las métricas de precisión, exhaustividad y medida-F. Ampliamente utilizadas en el análisis de los sistemas de procesamiento de lenguaje natural y los sistemas de minería de opiniones. Tomando como base los resultados obtenidos de la aplicación de estos procesos de validación se ha realizado todo el estudio de evaluación para cada uno de los módulos analizados.

Los objetivos que se establecieron en el desarrollo de la metodología propuesta en esta tesis doctoral se han alcanzado con éxito y los resultados obtenidos se presentan en esta memoria con la siguiente organización:

En el **Capítulo I** se realiza un estudio detallado del estado del arte de las tecnologías relacionadas con la investigación aportada en este trabajo. El estudio comienza con la Web Semántica y las ontologías, que constituyen una parte fundamental dentro de la metodología aportada en esta tesis doctoral. Se define el concepto de Web Semántica y se enumeran los fundamentos tecnológicos sobre los que se sustenta. Seguidamente se comentan los problemas actuales con los que se está encontrando la Web Semántica para poder llegar a su implantación definitiva, y se acaba esta sección enumerando y explicando los componentes que conforman la Web Semántica. Uno de estos

componentes, el más importante para la metodología que expone esta tesis, es la ontología. Por tanto, se define el concepto de ontología y se muestran los elementos que las conforman. Además, se lleva a cabo un estudio sobre los distintos lenguajes de los que disponemos actualmente para representar a las ontologías. Finaliza esta parte de Web Semántica y ontologías con una mención a los trabajos más relevantes en minería de opiniones que utilizan estas tecnologías. Además, se explica la importancia de estas tecnologías en la solución aportada en esta tesis doctoral.

El capítulo continúa con un estudio del estado del arte en tareas de procesamiento del lenguaje natural (*PLN*). Se define el concepto de *PLN* y se habla sobre las características que posee la información expresada en lenguaje natural. Seguidamente se enumeran los niveles en los que se puede dividir el lenguaje natural. Tras esto, se aporta un estudio sobre los tipos de recursos lingüísticos que se pueden utilizar actualmente en tareas de *PLN*. Finaliza esta parte de *PLN* hablando sobre las tareas de *PLN* que se aplican usualmente en problemas de minería de opiniones, destacando el hecho de que son dos sub-disciplinas íntimamente relacionadas.

El estado del arte continúa con el estudio de los conceptos fundamentales de análisis vectorial y geometría espacial que se necesitan para llevar a cabo la implementación de la metodología de análisis de sentimientos propuesta en esta tesis. Se define el concepto de espacio vectorial y se muestra el conjunto de operaciones vectoriales que pueden realizarse dentro de él. También se define el concepto de sistema de referencia en el espacio, que será un concepto clave en la solución propuesta en esta tesis. Tras esto se muestran los conceptos de sub-espacio vectorial y sistema de generadores, ya que ambos conceptos jugarán un papel decisivo a la hora de calcular la polaridad de las características identificadas en las opiniones de los usuarios. Finalmente, termina esta parte de geometría espacial enumerando los trabajos más significativos en minería de opiniones que utilizan vectores para llevar a cabo sus tareas, y se comenta la manera en la que se han utilizado todos los conceptos descritos en esta sección dentro de la solución aportada.

El capítulo continúa mostrando el estado del arte del concepto fundamental en torno al que gira este trabajo, es decir, la minería de opiniones. Para ello se hace un recorrido histórico sobre los trabajos más relevantes a lo largo de los tiempos en esta sub-disciplina. Se muestra el concepto de minería de opiniones basada en características y, por último, se realiza una comparativa exhaustiva sobre los trabajos más significativos aparecidos en los últimos años en el campo de la minería de opiniones.

El capítulo finaliza enumerando los objetivos a cumplir durante el desarrollo de esta tesis doctoral. El principal objetivo de esta tesis es definir una metodología innovadora de minería de opiniones basada en características guiada por ontologías. La metodología es independiente del dominio e idioma en los que se aplique, sin embargo la ausencia de recursos lingüísticos de calidad para idiomas distintos del inglés hará que para la implementación y validación de la misma utilicemos un corpus de opiniones y una base de datos de sentimientos en inglés. Además, como objetivo secundario, se desarrollará una aplicación de software que implementa esta metodología cumpliendo con sus requisitos, lo cual permitirá validarla de manera fácil y rápida en distintos dominios.

En el **Capítulo II** se presenta la arquitectura general del sistema. Esta arquitectura se ha diseñado para cumplir el principal objetivo de esta tesis doctoral: desarrollar un nuevo método para la minería de opiniones basada en características y guiada por ontologías. Esta arquitectura es la base sobre la que se sustenta la metodología propuesta en este trabajo. Se compone de cuatro módulos principales: (i) módulo de procesamiento del lenguaje natural, (ii) módulo de identificación de características, (iii) módulo de cálculo de polaridades de características y (iv) módulo de análisis de sentimientos.

El módulo de procesamiento del lenguaje natural lleva a cabo el análisis morfológico y léxico de las opiniones de los usuarios contenidas en el corpus. El módulo de identificación de características tiene como objetivo la extracción de todas las características contenidas en las opiniones utilizando una ontología del dominio. El módulo de cálculo de la polaridad de características lleva a cabo el cálculo de los vectores de polaridad asociados a las características identificadas en las opiniones. Por último, el módulo de análisis de sentimientos lleva a cabo el proceso de clasificación de sentimientos de las opiniones tomando como punto de partida los vectores de polaridad de las características identificadas en ellas.

El capítulo finaliza mostrando un diagrama general de procesos donde se explica en detalle el funcionamiento de cada uno de ellos. Además, se aporta un ejemplo práctico en el que se describe paso a paso cómo funciona la aplicación de la metodología en el dominio de las películas cinematográficas con opiniones escritas en inglés.

En el **Capítulo III** se presentan los fundamentos tecnológicos necesarios para crear un sistema de software que lleve a cabo la implementación de la metodología de minería de opiniones basada en características propuesta en esta tesis doctoral. Las partes más destacadas de este sistema son las siguientes: acceso al sistema y control de usuarios, configuración del sistema, proceso de análisis morfológico y léxico, proceso de

identificación de características, proceso de cálculo de la polaridad de características y proceso de análisis de sentimientos de opiniones.

El capítulo presenta un manual de usuario del sistema implementado que explica de manera detallada cómo llevar a cabo los procesos indicados en el párrafo anterior. Además se capturan las pantallas del sistema para mostrar de forma gráfica cómo navegar a través de las distintas opciones que nos ofrece el mismo.

El capítulo finaliza mostrando las bondades del sistema que presenta un modo de trabajo individualizado para cada usuario. Esto permite que cada uno se identifique al inicio y a partir de ahí pueda trabajar con la herramienta definiendo sus propios proyectos y configuraciones. Los idiomas de la interfaz de usuario de la herramienta son el español y el inglés. La aplicación es multiplataforma y ha sido implementada en un entorno cliente/servidor que proporciona muchas ventajas.

En el **Capítulo IV** se presentan las medidas de evaluación estándar que serán empleadas para evaluar la metodología y comprobar el correcto funcionamiento del sistema implementado. Las métricas de evaluación elegidas son las siguientes: precisión, exhaustividad y medida-F. Validar los resultados producidos por el sistema que implementa la metodología propuesta en esta tesis de manera manual puede ser una labor costosa, por lo que para aplicar estas métricas de evaluación se ha implementado un módulo de validación de resultados que realiza esta labor de manera automática.

El capítulo comienza mostrando el diagrama de procesos que se ejecutan en el módulo de validación de resultados construido para validar los resultados producidos por el sistema. En el mismo se explican detalladamente las condiciones bajo las cuales se han llevado a cabo los experimentos.

A continuación se realiza la evaluación de la metodología en dos dominios distintos: el cinematográfico y el financiero. Se han escogido dos dominios tan divergentes para comprobar el comportamiento de la metodología en situaciones totalmente distintas.

Para realizar los experimentos en el dominio de las películas cinematográficas se ha utilizado un corpus existente en la Web que se ha empleado en numerosos trabajos de investigación. Se trata de un corpus con opiniones de usuarios que utilizan un lenguaje informal. Los resultados obtenidos en este dominio han sido bastante satisfactorios tanto para el cálculo de las polaridades de las características como para el análisis de sentimientos de las opiniones.

Para analizar el dominio financiero se han llevado a cabo los mismos experimentos. En este caso se ha elaborado un corpus propio para los experimentos debido a la dificultad de encontrar uno en la Web disponible para usar. Se trata de un corpus con opiniones de usuarios expertos en el dominio que utilizan un lenguaje periodístico o formal. Los resultados obtenidos para este dominio también han sido bastante prometedores. A pesar de las características tan particulares de este dominio se han obtenido buenos resultados en el análisis de sentimientos de las opiniones de los usuarios.

El capítulo finaliza con una comparativa exhaustiva de los resultados obtenidos en cada dominio. Se han analizado cada una de las variables y parámetros que condicionan el funcionamiento de la metodología y se han estudiado las posibles causas de las diferencias de los resultados obtenidos en ambos dominios.

Por último, en el **Capítulo V** se muestran las conclusiones obtenidas durante la elaboración de este trabajo en base a los experimentos realizados. Así como un análisis de las posibles líneas futuras de investigación que podrían seguirse tomando como base la presente tesis doctoral.

Contenido General

Dedicatoria	iii
Agradecimientos	v
Resumen	viii
Introducción	xiii
Contenido General	xxii
Contenido Detallado	xxiv
Índice de Figuras	xxix
Índice de Tablas	xxxiv
Lista de Acrónimos	38
Capítulo I	41
Capítulo II	110
Capítulo III	172
Capítulo IV	201
Capítulo V	284
Capítulo VI	309
Referencias	312

Contenido Detallado

Capítulo I. Estado del arte

1.1	Introducción	42
1.2	Web Semántica y ontologías	43
1.2.1	Introducción	43
1.2.2	Web Semántica.....	44
1.2.2.1	Definición.....	44
1.2.2.2	Fundamentos de la Web Semántica	46
1.2.2.3	Problemática actual de implantación de la Web Semántica.....	47
1.2.2.4	Componentes de la Web Semántica	48
1.2.3	Ontologías	51
1.2.3.1	Definición.....	51
1.2.3.2	Elementos de las ontologías	52
1.2.3.3	Lenguajes para la representación de Ontologías	54
1.2.3.3.1	Resource Description Framework (RDF).....	54
1.2.3.3.2	Resource Description Framework Schema (RDFS).....	56
1.2.3.3.3	Web Ontology Language (OWL).....	58
1.2.4	Web Semántica y ontologías aplicadas en minería de opiniones.....	63
1.3	Procesamiento del lenguaje natural.....	64
1.3.1	Introducción	64
1.3.2	Información expresada en lenguaje natural.....	66
1.3.3	Niveles del lenguaje natural	67
1.3.4	Recursos lingüísticos en sistemas de PLN	71
1.3.5	PLN aplicado en minería de opiniones	74
1.4	Espacios Vectoriales	75
1.4.1	Introducción	75
1.4.2	Sistema de referencia en el espacio.....	77
1.4.3	Sub-espacios vectoriales	78
1.4.4	Sistemas de generadores.....	79
1.4.5	Espacios vectoriales en minería de opiniones	80
1.5	Minería de opiniones	83
1.5.1	Introducción	83
1.5.2	Evolución histórica de la minería de opiniones	85
1.5.3	Minería de opiniones basada en características.....	88
1.5.4	Recursos lingüísticos actuales en minería de opiniones.....	89

1.5.5	Comparativa de soluciones actuales en minería de opiniones.....	94
1.5.5.1	Recursos iniciales	96
1.5.5.2	Técnicas de clasificación de sentimientos	97
1.5.5.3	Grado de automatización	99
1.5.5.4	Identificación de características.....	99
1.5.5.5	Dominios.....	100
1.5.5.6	Idiomas.....	101
1.5.5.7	Nivel de clasificación de sentimientos.....	102
1.5.5.8	Conclusiones de la comparativa	103
1.6	Motivación y objetivos	104
1.7	Resumen.....	106

Capítulo II. Sistema basado en tecnologías de la Web Semántica para la minería de opiniones basada en características

2.1	Introducción	111
2.2	Objetivos del sistema	112
2.3	Arquitectura del Sistema.....	113
2.3.1	Capas de la arquitectura	113
2.3.2	Esquema general de la arquitectura	115
2.4	Módulo de procesamiento del lenguaje natural	117
2.5	Módulo de identificación de características.....	122
2.5.1	Búsqueda de características	124
2.5.2	Búsqueda de sinónimos	127
2.5.3	Cálculo del peso de las características	128
2.6	Módulo de cálculo de polaridades de características.....	131
2.6.1	Cálculo de la polaridad de las características identificadas	132
2.6.2	Agregación de polaridades de características	143
2.7	Módulo de análisis de sentimientos de opiniones.....	145
2.8	Diagrama general de procesos	147
2.9	Ejemplo de funcionamiento del sistema	149
2.9.1	Estructura de la ontología del dominio y proceso de identificación de términos	151
2.9.2	Configuración inicial del sistema	153
2.9.3	Análisis morfológico del corpus de opiniones.....	154
2.9.4	Identificación de características.....	158
2.9.4.1	Cálculo del peso de las características identificadas	160
2.9.5	Cálculo de la polaridad de las características	161
2.9.5.1	Agregación de polaridades de características	165
2.9.6	Análisis de sentimientos	167
2.10	Resumen.....	169

Capítulo III. Un sistema de software para la minería de opiniones de usuarios a partir de texto en lenguaje natural

3.1	Introducción	173
3.2	Fundamentos tecnológicos.....	173
3.2.1	Herramientas	174
3.2.1.1	Java	174

3.2.1.2	Eclipse	175
3.2.1.3	Stanford Log-linear Part-Of-Speech Tagger	175
3.2.1.4	Etiquetas Penn Treebank	177
3.2.1.5	WordNet	178
3.2.1.6	SentiWordNet	180
3.2.1.7	Protégé 3.4.1	181
3.2.2	Librerías	182
3.2.2.1	JENA	182
3.2.2.2	JAWS	183
3.3	Herramienta software para la clasificación de sentimientos de opiniones de usuarios expresadas en lenguaje natural	184
3.3.1	Introducción	184
3.3.2	Una sesión con el sistema implementado en esta tesis	185
3.3.2.1	Acceso al sistema e interfaz inicial	185
3.3.2.2	Configuración del sistema	187
3.3.2.3	Procesos del sistema	190
3.3.2.4	Idiomas de la interfaz del sistema	197
3.3.2.5	Ayuda del sistema	197
3.4	Resumen	198

Capítulo IV. Validación de la metodología de minería de opiniones basada en características guiada por ontologías

4.1	Introducción	202
4.2	Medidas de evaluación	203
4.2.1	Definición de las medidas de evaluación estándar	203
4.2.2	Medidas de evaluación para la fase de identificación de características... ..	205
4.2.3	Medidas de evaluación para la fase de cálculo de la polaridad de las características	206
4.2.4	Medidas de evaluación para la fase de cálculo de la polaridad global de las opiniones	206
4.3	Módulo de validación de resultados	207
4.4	Validación en el dominio de las películas cinematográficas	208
4.4.1	Introducción	208
4.4.2	Corpus	210
4.4.3	Ontología del dominio cinematográfico	211
4.4.4	Resultados para la búsqueda de características	211
4.4.5	Resultados para la clasificación de sentimientos de características	214
4.4.6	Resultados para la clasificación de sentimientos de opiniones	222
4.4.7	Conclusión	236
4.5	Validación en el dominio de las finanzas	237
4.5.1	Introducción	237
4.5.2	Corpus	239
4.5.3	Ontología del dominio financiero	240
4.5.4	Resultados para la búsqueda de características	242
4.5.5	Resultados para la clasificación de sentimientos de características	244
4.5.6	Resultados para la clasificación de sentimientos de opiniones	253
4.5.7	Conclusión	266
4.6	Conclusión global del proceso de validación	267

4.7	Resumen.....	281

Capítulo V. Conclusiones y trabajo futuro

5.1	Introducción	285
5.2	Conclusiones	285
5.2.1	Procesamiento del lenguaje natural	290
5.2.2	Identificación de características.....	290
5.2.3	Análisis de sentimientos de características	291
5.2.4	Análisis de sentimientos de opiniones	292
5.2.5	Validación de resultados	293
5.3	Líneas futuras.....	294
5.3.1	Procesamiento del lenguaje natural	294
5.3.2	Identificación de características.....	296
5.3.3	Análisis de sentimientos de características.....	298
5.3.4	Análisis de sentimientos de opiniones	300
5.3.5	Aplicación de la lógica difusa dentro de la metodología propuesta en esta tesis doctoral	303

Capítulo VI. Contribuciones científicas

6.1	Publicaciones JCR	310
6.2	Publicaciones en revistas de otros índices de impacto	310
6.3	Publicaciones en congresos internacionales	310

Referencias

Referencias.....	313

Índice de Figuras

Capítulo I. Estado del arte

Figura 1.1. Tecnologías principales que forman el núcleo de la solución propuesta en esta tesis.	43
Figura 1.2. Arquitectura en capas para la Web Semántica.	48
Figura 1.3. Ejemplo de una sentencia en <i>RDF</i> (tripleta).	55
Figura 1.4. Ejemplo de un grafo en <i>RDF</i>	56
Figura 1.5. Ejemplo de definición en <i>RDF/XML</i>	56
Figura 1.6. Versiones del lenguaje <i>OWL</i>	59
Figura 1.7. Perfiles del lenguaje <i>OWL2</i>	62
Figura 1.8. Ejemplo de red semántica.	72
Figura 1.9. Recursos lingüísticos en <i>PLN</i> frente a ontologías.	74
Figura 1.10. Representación artística de un espacio vectorial.	75
Figura 1.11. Representación matemática de las operaciones definidas en un espacio vectorial.	76
Figura 1.12. Sistema de referencia en el espacio.	78
Figura 1.13. Expresiones de un sub-espacio vectorial.	79
Figura 1.14. Representación espacial de un sistema de generadores.	80

Capítulo II. Sistema basado en tecnologías de la Web Semántica para la minería de opiniones basada en características

Figura 2.1. Capas de software de la arquitectura del sistema.	115
Figura 2.2. Arquitectura general del sistema.	116
Figura 2.3. Módulo de procesamiento del lenguaje natural.	118
Figura 2.4. Extracto de la opinión de un usuario en lenguaje natural e idioma inglés.	121
Figura 2.5. Análisis morfológico llevado a cabo por el módulo de <i>PLN</i>	122
Figura 2.6. Módulo de identificación de características.	123
Figura 2.7. Extracto de la ontología del dominio ' <i>The Movie Ontology</i> '.	125
Figura 2.8. Características identificadas por el módulo de identificación de características en una opinión.	126
Figura 2.9. Estructura semántica para la ontología <i>The Movie Ontology</i>	127
Figura 2.10. Áreas de influencia de la posición relativa de características dentro de una opinión.	129
Figura 2.11. Módulo de cálculo de la polaridad de características.	132

Figura 2.12. Cálculo de la polaridad de características con el método <i>N_GRAM Before</i> cuando <i>N_GRAM</i> es igual a tres.....	134
Figura 2.13. Cálculo de la polaridad de características con el método <i>N_GRAM After</i> cuando <i>N_GRAM</i> es igual a tres.....	134
Figura 2.14. Cálculo de la polaridad de características con el método <i>N_GRAM Around</i> cuando <i>N_GRAM</i> es igual a tres.....	135
Figura 2.15. Cálculo de la polaridad de características con el método <i>All Phrase</i>	135
Figura 2.16. Vector de posición que representa la polaridad de una característica.	137
Figura 2.17. Vectores de posición que representan las polaridades de las características A, B, C y D.....	138
Figura 2.18. Pirámides geométricas de Polaridad en R^3	139
Figura 2.19. Representación gráfica de una <i>N-bola</i> en R^3 con centro en <i>P</i> y radio <i>r</i> ...	139
Figura 2.20. Representación gráfica del volumen de una pirámide geométrica.	140
Figura 2.21. Algoritmo que muestra la implementación de la función <i>estaContenido(P,P_x)</i>	140
Figura 2.22. Algoritmo que muestra la implementación de la función <i>puntoInterior(P,P_x)</i>	141
Figura 2.23. Algoritmo que muestra la implementación de la función <i>N-Bola(P,P_x)</i>	141
Figura 2.24. Módulo de análisis de sentimientos de opiniones.....	145
Figura 2.25. Diagrama general de procesos del sistema.	148
Figura 2.26. Arquitectura específica del sistema para un dominio e idioma concretos con las herramientas externas necesarias.	151
Figura 2.27. Ejemplo de definición en OWL de una clase en una ontología del dominio de las finanzas.	152

Capítulo III. Un sistema de software para la minería de opiniones de usuarios a partir de texto en lenguaje natural

Figura 3.1. Ejemplo de visualización de un <i>synset</i> con puntuaciones en <i>SentiWordNet</i>	181
Figura 3.2. Extracto de código Java que usa el API <i>JAWS</i>	183
Figura 3.3. Pantalla de acceso al sistema software de clasificación de sentimientos de opiniones.	186
Figura 3.4. Pantalla de inicio del sistema software de clasificación de sentimientos de opiniones.	187
Figura 3.5. Pantalla del menú de configuración del sistema software de clasificación de sentimientos de opiniones.	188
Figura 3.6. Pantalla del menú de configuración de proyectos para añadir un nuevo proyecto.	189
Figura 3.7. Pantalla del menú de configuración de proyectos para listar los proyectos de un usuario.	189
Figura 3.8. Pantalla del menú de configuración de proyectos para establecer las opciones generales de configuración asociadas a cada proyecto.	190
Figura 3.9. Pantalla del menú de procesos del sistema software de clasificación de sentimientos de opiniones.	191
Figura 3.10. Pantalla del menú de procesos del sistema para establecer el valor de los parámetros comunes a los procesos del sistema.....	192
Figura 3.11. Pantalla del menú de procesos del sistema para llevar a cabo el proceso de análisis morfológico.	193

Figura 3.12. Pantalla del menú de procesos del sistema para llevar a cabo el proceso de identificación de características.	193
Figura 3.13. Pantalla del menú de procesos del sistema para llevar a cabo el proceso de clasificación de sentimientos de características.	194
Figura 3.14. Pantalla del menú de procesos del sistema para consultar los resultados obtenidos en la clasificación de sentimientos de características.	196
Figura 3.15. Pantalla del menú de procesos del sistema para llevar a cabo el proceso de análisis de sentimientos de opiniones.	196
Figura 3.16. Pantalla del menú de procesos del sistema para consultar los resultados obtenidos en la clasificación de sentimientos de opiniones de usuarios.	197

Capítulo IV. Validación de la metodología de minería de opiniones basada en características guiada por ontologías

Figura 4.1. Diagrama de procesos de la validación del sistema.	208
Figura 4.2. Comparativa visual de identificación de características en el dominio cinematográfico.	214
Figura 4.3. Comparativa visual de precisión en el cálculo de la polaridad de las características en el dominio cinematográfico.	222
Figura 4.4. Efecto de los parámetros Z_1 , Z_2 y Z_3 en el análisis de sentimientos de opiniones con el método <i>N_GRAM Before</i>	226
Figura 4.5. Efecto de los parámetros Z_1 , Z_2 y Z_3 en el análisis de sentimientos de opiniones con el método <i>N_GRAM After</i>	229
Figura 4.6. Efecto de los parámetros Z_1 , Z_2 y Z_3 en el análisis de sentimientos de opiniones con el método <i>N_GRAM Around</i>	232
Figura 4.7. Efecto de los parámetros Z_1 , Z_2 y Z_3 en el análisis de sentimientos de opiniones con el método <i>All Phrase</i>	234
Figura 4.8. Comparativa visual de precisión en el cálculo del análisis de sentimientos de opiniones.	235
Figura 4.9. Un extracto de la ontología financiera.	242
Figura 4.10. Comparativa visual de identificación de características en el dominio de las finanzas.	244
Figura 4.11. Comparativa visual de precisión en el cálculo de la polaridad de las características en el dominio financiero.	252
Figura 4.12. Efecto de los parámetros Z_1 , Z_2 y Z_3 en el análisis de sentimientos de opiniones con el método <i>N_GRAM Before</i>	256
Figura 4.13. Efecto de los parámetros Z_1 , Z_2 y Z_3 en el análisis de sentimientos de opiniones con el método <i>N_GRAM After</i>	259
Figura 4.14. Efecto de los parámetros Z_1 , Z_2 y Z_3 en el análisis de sentimientos de opiniones con el método <i>N_GRAM Around</i>	262
Figura 4.15. Efecto de los parámetros Z_1 , Z_2 y Z_3 en el análisis de sentimientos de opiniones con el método <i>All Phrase</i>	264
Figura 4.16. Comparativa visual de precisión en el cálculo del análisis de sentimientos de opiniones.	265
Figura 4.17. Comparativa visual de resultados de exhaustividad obtenidos en el proceso de identificación de características entre los dominios cinematográfico y financiero.	269
Figura 4.18. Comparativa visual de resultados de medida-F obtenidos en el proceso de identificación de características entre los dominios cinematográfico y financiero.	269

Figura 4.19. Comparativa visual de precisión en el cálculo de polaridad de características con el método <i>N_GRAM Before</i> entre distintos dominios.....	271
Figura 4.20. Comparativa visual de precisión en el cálculo de polaridad de características con el método <i>N_GRAM After</i> entre distintos dominios.	272
Figura 4.21. Comparativa visual de precisión en el cálculo de polaridad de características con el método <i>N_GRAM Around</i> entre distintos dominios.	273
Figura 4.22. Comparativa visual de precisión en el cálculo de polaridad de características con el método <i>All Phrase</i> entre distintos dominios.	274
Figura 4.23. Comparativa visual de precisión en el cálculo del análisis de sentimientos de opiniones con el método <i>N_GRAM Before</i> entre distintos dominios.....	275
Figura 4.24. Comparativa visual de precisión en el cálculo del análisis de sentimientos de opiniones con el método <i>N_GRAM After</i> entre distintos dominios.....	276
Figura 4.25. Comparativa visual de precisión en el cálculo del análisis de sentimientos de opiniones con el método <i>N_GRAM Around</i> entre distintos dominios.....	277
Figura 4.26. Comparativa visual de precisión en el cálculo del análisis de sentimientos de opiniones con el método <i>All Phrase</i> entre distintos dominios.	277

Capítulo V. Conclusiones y trabajo futuro

Figura 5.1. Esquema general detallado de los procesos que componen la metodología propuesta.	289
Figura 5.2. Arquitectura general del sistema con la inclusión del módulo <i>fuzzy</i>	307

Índice de Tablas

Capítulo I. Estado del arte

Tabla 1.1. Niveles del lenguaje natural y herramientas de <i>PLN</i>	68
Tabla 1.2. Ejemplo de <i>synset</i> con <i>gloss</i> dentro de <i>WordNet</i>	90
Tabla 1.3. Puntuaciones de algunos <i>synsets</i> en <i>SentiWordNet</i>	93
Tabla 1.4. Sistemas de minería de opiniones actuales.	98

Capítulo II. Sistema basado en tecnologías de la Web Semántica para la minería de opiniones basada en características

Tabla 2.1. Modelos de etiquetado de la herramienta <i>Stanford Log-linear Part-Of-Speech</i>	118
Tabla 2.2. Etiquetas <i>Penn Treebank</i>	119
Tabla 2.3. Definición de los elementos empleados en los diagramas de procesos del sistema.	147
Tabla 2.4. Opinión de un usuario en lenguaje natural.	155
Tabla 2.5. Opinión etiquetada de un usuario después de analizarla.	156
Tabla 2.6. Términos candidatos a características identificados por el sistema.	158
Tabla 2.7. Resultados de las características identificados por el sistema.	159
Tabla 2.8. Peso asociado a cada una de las características identificadas.	161
Tabla 2.9. Extracto del proceso de cálculo de la polaridad de cada una de las características identificadas.	162
Tabla 2.10. Resultados de las polaridades asignada a las características identificadas.	165
Tabla 2.11. Resultados de la agregación de polaridades de características sinónimas.	166
Tabla 2.12. Resultados del análisis de sentimientos de la opinión del usuario y de sus características identificadas.	168

Capítulo III. Un sistema de software para la minería de opiniones de usuarios a partir de texto en lenguaje natural

Tabla 3.1. Ejemplo de análisis sintáctico utilizando notación <i>Penn Treebank</i>	177
Tabla 3.2. Ejemplo de <i>synset</i> con <i>gloss</i> dentro de <i>WordNet</i>	179
Tabla 3.3. Ejemplo de jerarquía <i>hypernym</i> en <i>WordNet</i>	179

Capítulo IV. Validación de la metodología de minería de opiniones basada en características guiada por ontologías

Tabla 4.1. Resultados experimentales para la identificación de características en opiniones positivas en el dominio cinematográfico.....	212
Tabla 4.2. Resultados experimentales para la identificación de características en opiniones negativas en el dominio cinematográfico.....	212
Tabla 4.3. Resultados experimentales totales para la identificación de características en el dominio cinematográfico.	213
Tabla 4.4. Extracto del fichero de entrada para validación en el dominio cinematográfico.	214
Tabla 4.5. Valor de la precisión en el cálculo de la polaridad de las características con el método <i>N_GRAM Before</i>	217
Tabla 4.6. Valor de la precisión en el cálculo de la polaridad de las características con el método <i>N_GRAM After</i>	218
Tabla 4.7. Valor de la precisión en el cálculo de la polaridad de las características con el método <i>N_GRAM Around</i>	219
Tabla 4.8. Valor de la precisión en el cálculo de la polaridad de las características con el método <i>All Phrase</i>	220
Tabla 4.9. Valor de la precisión en el análisis de sentimientos de opiniones con el método <i>N_GRAM Before</i>	225
Tabla 4.10. Valor de la precisión en el análisis de sentimientos de opiniones con el método <i>N_GRAM After</i>	228
Tabla 4.11. Valor de la precisión en el análisis de sentimientos de opiniones con el método <i>N_GRAM Around</i>	231
Tabla 4.12. Valor de la precisión en el análisis de sentimientos de opiniones con el método <i>All Phrase</i>	233
Tabla 4.13. Sitios Web del dominio financiero consultados para la elaboración del corpus.....	240
Tabla 4.14. Resultados experimentales para la identificación de características en opiniones positivas en el dominio financiero.	242
Tabla 4.15. Resultados experimentales para la identificación de características en opiniones negativas en el dominio financiero.....	243
Tabla 4.16. Resultados experimentales totales para la identificación de características en el dominio financiero.	243
Tabla 4.17. Extracto del fichero de entrada para la validación en el dominio financiero.	245
Tabla 4.18. Valor de la precisión en el cálculo de la polaridad de las características con el método <i>N_GRAM Before</i>	247
Tabla 4.19. Valor de la precisión en el cálculo de la polaridad de las características con el método <i>N_GRAM After</i>	248
Tabla 4.20. Valor de la precisión en el cálculo de la polaridad de las características con el método <i>N_GRAM Around</i>	248
Tabla 4.21. Valor de la precisión en el cálculo de la polaridad de las características con el método <i>All Phrase</i>	250
Tabla 4.22. Valor de la precisión en el análisis de sentimientos de opiniones con el método <i>N_GRAM Before</i>	255
Tabla 4.23. Valor de la precisión en el análisis de sentimientos de opiniones con el método <i>N_GRAM After</i>	258

Tabla 4.24. Valor de la precisión en el análisis de sentimientos de opiniones con el método <i>N-GRAM Around</i> .	261
Tabla 4.25. Valor de la precisión en el análisis de sentimientos de opiniones con el método <i>All Phrase</i> .	263

Capítulo V. Conclusiones y trabajo futuro

Tabla 5.1. Extracto de la opinión de un usuario en el dominio de las películas cinematográficas.	299
Tabla 5.2. Lista de características identificadas con sus vectores de polaridad asociados	299
Tabla 5.3. Agrupación de polaridades de características identificadas en torno a conceptos ontológicos	300
Tabla 5.4. Extracto del lexicón <i>FullStrengthLexicon</i> construido para el idioma español	301
Tabla 5.5. Extracto del lexicón <i>ML-Senticon</i> construido para el idioma español	302
Tabla 5.6. Comparación de lexicones construidos para el idioma español frente a <i>SentiWordNet</i> .	302
Tabla 5.7. Ejemplo de clasificación de modificadores lingüísticos difusos para el idioma inglés.	306

Lista de Acrónimos

Término	Significado
AC	Aprendizaje Computacional
API	Application Programming Interface
ANN	Artificial Neural Networks
BSD	Berkeley Software Distribution
BORO	Business Object Reference Ontology
CRF	Conditional Random Fields
CERN	European Organization for Nuclear Research
CSE	Computer Science and Engineering
CSS	Cascading Style Sheets
CVS	Concurrent Versions System
DAML	DARPA Agent Markup Language
DARPA	Defense Advanced Research Project Agency
DB	DataBase
DIP	Data Information and Process Integration
DL	Description Logic
GATE	General Architecture for Text Engineering
HTML	HyperText Markup Language
HTTP	Hypertext Transfer Protocol
IA	Inteligencia Artificial
IDE	Integrated Development Environment
IE	Information Extraction
IEEE	Institute of Electrical and Electronics Engineers
IMDb	Internet Movie Database
JAWS	Java API for WordNet
JVM	Java Virtual Machine
JEE	Java Enterprise Edition
IR	Information Retrieval
KB	Knowledge base
MO	The Movie Ontology
NS	NameSpace
OIL	Ontology Inference Layer
OKBC	Open Knowledge Base Connectivity
OWL	Web Ontology Language
PLN	Procesamiento del Lenguaje Natural
POS-Taggers	Part-Of-Speech Taggers

RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
SAD	Sistema de Ayuda a la Decisión
SHOE	Simple HTML Ontology Extensions
SPARQL	SPARQL Protocol and RDF Query Language
SVM	Support Vector Machines
SWN	SentiWordNet
TCP/IP	Transmission Control Protocol/Internet Protocol
TIC	Tecnologías de la Información y la Comunicación
TOVE	Toronto Virtual Enterprise
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
URN	Uniform Resource Name
W3C	World Wide Web Consortium
WN	WordNet
WSD	Word Sense Disambiguation
WWW	World Wide Web
XBRL	Extensible Business Reporting Language
XHTML	eXtensible HyperText Markup Language
XML	eXtensible Markup Language

Capítulo I

Estado del arte

1.1 Introducción

En este capítulo se describe el estado actual de las tecnologías que forman el núcleo central de esta tesis doctoral: Web Semántica y ontologías, procesamiento del lenguaje natural (*PLN*), espacios vectoriales y minería de opiniones.

En primer lugar se comienza con un recorrido a través de los aspectos más relevantes de la Web Semántica con el objetivo de introducir sus componentes y conocer su estado actual. Las ontologías constituyen el componente central de la Web Semántica permitiendo representar el conocimiento subyacente en un dominio concreto. A continuación se introducen los elementos que componen las ontologías y los distintos lenguajes para su representación, así como el papel fundamental de las ontologías en la solución propuesta en esta tesis doctoral. Finalmente se presentan los trabajos más relevantes en minería de opiniones que hacen uso de técnicas de la Web Semántica y las ontologías.

En segundo lugar se lleva a cabo un estudio del estado del arte en tareas de procesamiento del lenguaje natural (*PLN*). El objetivo de esta sección es definir el concepto de *PLN* y hablar sobre las características que posee la información expresada en lenguaje natural mostrando los niveles en los que se puede clasificar. Además se aporta un estudio sobre los tipos de recursos lingüísticos que se pueden utilizar actualmente en tareas de *PLN*. Para finalizar se habla sobre las tareas de *PLN* que se aplican con asiduidad en problemas de minería de opiniones.

En tercer lugar se introducen los conceptos fundamentales sobre espacios vectoriales que son necesarios para explicar la metodología de análisis de sentimientos de opiniones que se propone en esta tesis doctoral. En esta sección se describen los conceptos que se manejan en la descripción de la metodología: los conceptos de espacio vectorial y sistema de referencia de espacio. También se incluye una introducción a los sub-espacios vectoriales y los sistemas de generadores con el objeto de llegar a la aplicación de estos conceptos en el campo de la minería de opiniones. Finalmente se muestran los trabajos más importantes en el campo de la minería de opiniones utilizando vectores. Además, se realiza una comparativa destacando similitudes y diferencias entre estos trabajos y la solución aportada en esta tesis doctoral.

El capítulo concluye mostrando el estado actual de los trabajos realizados en el campo de la minería de opiniones. En esta sección se define el concepto de minería de opiniones y se hace un recorrido histórico sobre los trabajos más importantes a lo largo

de la historia en esta sub-disciplina. Seguidamente se introduce el concepto de minería de opiniones basada en características y, por último, se realiza una comparativa exhaustiva sobre los trabajos más significativos aparecidos en los últimos años en el campo de la minería de opiniones.

La figura 1.1 muestra un esquema global de las tecnologías más importantes que se han utilizado para desarrollar la metodología de minería de opiniones basada en características que se propone en este trabajo.



Figura 1.1. Tecnologías principales que forman el núcleo de la solución propuesta en esta tesis.

1.2 Web Semántica y ontologías

1.2.1 Introducción

El objetivo de esta sección es introducir los componentes de la Web Semántica empleados a lo largo de este trabajo y conocer su estado actual. Las ontologías son un componente clave de la Web Semántica que permiten representar el conocimiento

subyacente en un dominio concreto. Por esta razón, en esta sección se muestra una introducción a los elementos constituyentes de las ontologías y los distintos lenguajes para su representación, así como el papel fundamental de las ontologías en la solución propuesta en esta tesis doctoral. La sección finaliza comentando los trabajos más relevantes aparecidos en el campo de la minería de opiniones que utilizan técnicas de la Web Semántica y ontologías.

1.2.2 Web Semántica

1.2.2.1 Definición

“La Web Semántica es una red de datos que puede ser procesada directa o indirectamente por máquinas. Es una Web extendida que permitirá a humanos y máquinas trabajar en cooperación mutua” (Berners-Lee et al., 2001). Esta definición de la Web Semántica fue aportada por Tim Berners-Lee, James Hendler y Ora Lassila en su artículo *“The Semantic Web”* publicado en 2001. En ese artículo se presenta la Web Semántica como una nueva Web cuyo objetivo es la incorporación de información estructurada en la Web actual para facilitar las tareas de procesamiento de información por parte de las máquinas. Por lo que no presenta a la Web Semántica como una versión independiente sino como una evolución que dota de nuevas propiedades al formato actual.

Por otro lado, el W3C (*World Wide Web Consortium*) define la Web Semántica como *“una Web extendida, dotada de mayor significado, en la que cualquier usuario en Internet puede encontrar respuestas a sus preguntas de forma más rápida y sencilla gracias a una información mejor definida”*⁴. Según esta definición, al dotar a la Web de más semántica, se pueden obtener soluciones a problemas habituales en la búsqueda de información gracias a la utilización de una infraestructura común mediante la cual es posible compartir, procesar y transferir información de forma sencilla.

Ambas definiciones de la Web Semántica presentan a ésta como una extensión sobre la Web actual que tiene como objetivo dotar de mayor inteligencia a la misma a partir de información mejor definida. Por lo que es necesario contar con mecanismos de definición de la información que permitan la adición de descripciones explícitas en los

⁴ W3C Web Semántica. <http://www.w3.org/standards/semanticweb/>, última consulta realizada en Abril de 2015.

sitios Web actuales. Esta información añadida ayudará a las máquinas a entender la información y proporcionará a los usuarios respuestas rápidas y de mayor calidad.

Por otro lado, la Web ha cambiado profundamente la forma en la que nos comunicamos, hacemos negocios y realizamos nuestro trabajo. La comunicación prácticamente con todo el mundo en cualquier momento y a bajo coste es posible hoy en día. Podemos realizar transacciones económicas a través de Internet. Tenemos acceso a millones de recursos independientemente de nuestra situación geográfica e idioma. Todos estos factores han contribuido al éxito de la Web. Sin embargo, al mismo tiempo, estos factores que han propiciado el éxito de la Web también han originado sus principales problemas: sobrecarga de información y heterogeneidad de fuentes de información con el consiguiente problema de interoperabilidad. La Web Semántica ayuda a resolver estos dos importantes problemas permitiendo a los usuarios delegar tareas en software. Gracias a la semántica en la Web, el software es capaz de procesar su contenido, razonar con este, combinarlo y realizar deducciones lógicas para resolver problemas cotidianos automáticamente.

Dicho de otra manera, la carencia de un modelo bien definido de representación de la información en la Web actual ha traído consigo diversos problemas relacionados con su procesamiento. Para intentar solucionarlos, el W3C, organismo encargado de guiar la evolución de la Web, ha propuesto su transformación hacia una nueva Web, la Web Semántica (Pedraza-Jiménez et al., 2007). El Dr. Berners-Lee en (Berners-Lee et al., 2001) publicó que era necesario cambiar la forma de la Web actual por una en la que se pudiera expresar el contenido de las páginas de una forma en que los ordenadores pudieran “entenderlas”. Dicho proyecto se denominó “Web Semántica”. Otra de las características de esta Web es que aportará, mediante conjuntos de reglas de inferencia, la posibilidad de llevar a cabo razonamientos automáticos sobre las páginas Web, lo que permitirá a los ordenadores desarrollar tareas inteligentes (Codina & Rovira, 2006).

Por tanto, la Web Semántica surge como un conjunto de actividades desarrolladas en el seno del W3C con el objetivo de crear tecnologías para publicar en Web datos legibles por aplicaciones informáticas. La idea subyacente es añadir metadatos semánticos y ontológicos a la *World Wide Web*. Estas informaciones adicionales describen el contenido, el significado y la relación de los datos. Además, se deben proporcionar de manera formal, para que así sea posible evaluarlas automáticamente por máquinas de procesamiento. El objetivo es mejorar Internet ampliando la interoperabilidad entre los sistemas informáticos usando "agentes inteligentes", que básicamente son programas que buscan información sin operadores humanos.

El precursor de la idea, Tim Berners-Lee, intentó desde el principio incluir información semántica en su creación, la *World Wide Web*, pero por diferentes causas no fue posible. Para subsanar esta omisión introdujo el concepto de semántica en la Web.

1.2.2.2 Fundamentos de la Web Semántica

En la actualidad, la *World Wide Web* está basada principalmente en documentos escritos en *HTML*, un lenguaje de etiquetado que sirve principalmente para crear hipertexto en Internet. El lenguaje *HTML* es válido para adecuar el aspecto visual de un documento e incluir objetos multimedia en el texto, pero ofrece pocas posibilidades para categorizar los elementos que configuran el texto más allá de las típicas funciones estructurales. *HTML* permite mediante una herramienta de visualización (como un navegador) mostrar, por ejemplo, un catálogo de objetos en venta. El código *HTML* de este catálogo puede explicitar aspectos como "el título del documento es 'Carpintería Peñalver'", pero no hay forma de precisar dentro del código *HTML* si el producto MP1202 es una "estantería para libros", con un "precio de venta al público" de 150 €, o si es otro tipo de producto de consumo. Lo único que *HTML* permite es alinear el precio en la misma fila que el nombre del producto. No hay forma de indicar "esto es un catálogo", "estantería para libros" es un producto, o "150 €" es el precio. Tampoco hay forma de relacionar ambos datos para describir un elemento específico que se pudiese comparar con otros elementos similares en el mismo catálogo.

La Web Semántica se ocupa de resolver estas deficiencias. Para ello dispone de tecnologías de descripción de los contenidos, como *RDF*⁵ y *OWL*⁶, además de *XML*⁷, el lenguaje de marcado diseñado para describir los datos. Estas tecnologías se combinan para aportar descripciones explícitas de los recursos de la Web. De esta forma el contenido del documento estará formado por los datos de una base de datos accesibles por Web, y las etiquetas embebidas en el documento (normalmente en *XHTML*⁸ o *XML*). Esas etiquetas permiten a las aplicaciones software interpretar los documentos y realizar procesos inteligentes de captura y tratamiento de información.

Resumiendo, el proceso de cambio de la Web actual hacia la Web Semántica requiere de la adición de elementos que den soporte a la nueva forma de estructurar los

⁵ W3C RDF. <http://www.w3.org/RDF/>, última consulta realizada en Abril de 2015.

⁶ W3C OWL. <http://www.w3.org/2001/sw/wiki/OWL>, última consulta realizada en Abril de 2015.

⁷ W3C XML. <http://www.w3.org/XML/>, última consulta realizada en Abril de 2015.

⁸ W3C XHTML. <http://www.w3.org/TR/xhtml1/>, última consulta realizada en Abril de 2015.

contenidos y aportar significado. Ante este problema de representación del conocimiento, las ontologías proporcionan un vocabulario que permite describir las relaciones entre diferentes términos de manera flexible y sin ambigüedades, facilitando su interpretación por las máquinas y los humanos (Horrocks et al., 2003). En este contexto de adición de nuevos elementos, en el siguiente apartado se muestran algunas de las dificultades a las que se tiene que enfrentar la Web Semántica antes de su implantación definitiva.

1.2.2.3 Problemática actual de implantación de la Web Semántica

Actualmente, el desarrollo y puesta en marcha de la Web Semántica tiene algunas dificultades que no han podido superarse todavía. Una de ellas es tecnológica y la otra está relacionada con la falta de interés de los propietarios de las páginas Web (Peis et al., 2003). Las tecnologías para expresar la información en el formato que requiere una Web Semántica existen desde hace años. El componente más especializado es el lenguaje *OWL (Web Ontology Language)*, que existe como estándar del W3C desde 2004 (McGuinness & Van Harmelen, 2004). Sin embargo, el componente tecnológico que falta es el que permita convertir de forma automática el abundante contenido de texto de las páginas Web en etiquetas *OWL*.

La Web Semántica requiere que los creadores de las páginas Web traduzcan "a mano" su contenido en etiquetas *OWL* para que pueda ser interpretado por agentes semánticos. Pero esto es inviable debido al gran volumen que tiene la Web y su imparable crecimiento, razón por la que hay que plantear procesos de anotación semiautomáticos (Durley et al., 2013; Martí et al., 2007; Criado, 2009). Afortunadamente muchas páginas Web tienen información formateada que recogen de su base de datos y pueden realizar esta conversión de manera automática, aunque no representan un porcentaje muy alto. Por ejemplo, un sitio Web con miles de fichas de películas tiene sus datos bien ordenados y clasificados en una base de datos, lo cual permite elaborar un traductor de "ficha de película" a *OWL* que sirva para todas las fichas, sin necesidad de tener que realizar manualmente la traducción para cada una.

La otra barrera que encuentra la implantación definitiva de la Web Semántica es el modelo de negocio de gran cantidad de páginas Web, ya que éstas obtienen ingresos de la publicidad. Estos ingresos son posibles únicamente si sus páginas son visitadas por una persona, y dejarían de existir si los datos quedan disponibles para que los interprete un proceso automático. El siguiente ejemplo ilustra este concepto: un estudiante debe

realizar un trabajo de investigación para la escuela sobre la vida de un personaje; sin embargo, un sistema semántico realiza la investigación y presenta en pantalla el resultado (fecha de nacimiento y defunción, batallas en las que participó, hechos destacados, frases célebres, etc.), y todo esto sin necesidad de que el estudiante haya tenido que acceder a ninguna página Web específica y, por lo tanto, sin consumir la publicidad de los sitios que pusieron a disposición esa información.

Es muy probable que, de surgir una Web Semántica masiva, el modelo de negocios se adapte (Bernal et al., 2014). No obstante, mientras no haya indicios de que esto vaya a suceder de manera inevitable e inminente, no es lógico suponer que los sitios Web basados en publicidad apoyen su desarrollo poniendo su información en el formato necesario.

1.2.2.4 Componentes de la Web Semántica

El modelo de componentes o arquitectura de la Web Semántica sigue el modelo de capas definido por Tim Berners-Lee (véase la figura 1.2) (Berners-Lee et al., 2001). El objetivo de esta arquitectura es proporcionar a los usuarios y aplicaciones software capacidades de procesamiento e inferencia sobre los contenidos de la Web.

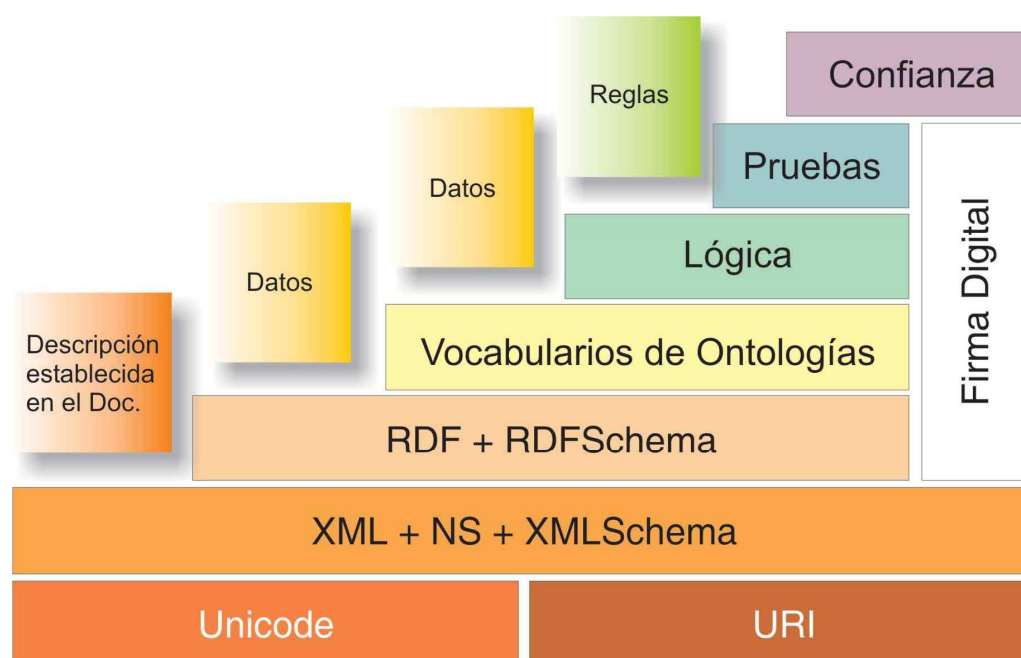


Figura 1.2. Arquitectura en capas para la Web Semántica.

En la figura 1.2 se muestra la arquitectura tecnológica de la Web Semántica. A continuación se describe la función que desempeña cada una de las capas que aparecen en la figura:

- **Capa Unicode.** La capa Unicode permite que toda la información pueda expresarse en cualquier idioma. Esta capa se encarga de proporcionar una codificación de caracteres estándar que permita utilizar distintos alfabetos.
- **Capa URI.** La capa *URI (Uniform Resource Identifier)* proporciona un identificador único compuesto por una *URL (Uniform Resource Locator)* y una *URN (Uniform Resource Name)*. La *URL* describe la localización del recurso y la *URN* describe el espacio de nombres del recurso.
- **Capa XML, NS y XML Schema.** Esta capa ofrece un formato común de los documentos que hace posible que los agentes puedan entenderse entre ellos independientemente de las fuentes de información que se utilicen para crear estos documentos. El lenguaje de etiquetas *XML* proporciona un formato común para el intercambio de documentos. Este lenguaje viene frecuentemente acompañado de la tecnología *XML Query*, también conocida como *XQuery*, que define un lenguaje de consultas estándar basado en la notación *XML* para definir consultas y manejar los resultados. El espacio de nombres (*NS*, del inglés *Namespace*) sirve para cualificar elementos y atributos de nombres, y asociarlos con los espacios de nombres identificados a través de referencias *URI*. Finalmente, los esquemas *XML (XML Schema)* proporcionan plantillas que facilitan la elaboración de documentos estándar.
- **Capa RDF y RDF Schema.** La capa *RDF* y *RDF Schema* define el lenguaje universal con el que se expresa el conocimiento en la Web Semántica. *RDF (Resource Description Framework)* es una familia de especificaciones del W3C originalmente diseñadas como un modelo de datos para metadatos. Ha llegado a ser usado como un método general para la descripción conceptual o modelado de la información que se implementa en los recursos Web. Por su parte, *RDF Schema* proporciona un vocabulario para describir las propiedades y las clases de los recursos *RDF* con una semántica claramente definida. Con esta semántica se pueden establecer jerarquías de generalización entre propiedades y clases. En las subsecciones 1.2.3.3.1 y 1.2.3.3.2 de la presente tesis doctoral se describen con mayor nivel de detalle los lenguajes para la representación de ontologías *RDF* y *RDF Schema* respectivamente.
- **Capa de Vocabulario de Ontologías.** La capa de vocabulario de ontologías incluye más vocabulario para facilitar la descripción precisa de conceptos,

relaciones y propiedades con los que conceptualizar un dominio específico. Añade expresividad a la capa anterior. Tanto esta capa como la anterior serán las que utilizaremos de manera intensiva en la implementación de la solución propuesta en este trabajo.

- **Capa Lógica.** La capa lógica está compuesta por un conjunto de reglas de inferencia que permiten a los agentes software procesar y relacionar información. Estas reglas permiten procesar de manera automática la información a nivel semántico.
- **Capa de Pruebas.** La capa de pruebas tiene como objetivo el intercambio de hechos y reglas estándar para facilitar la interoperabilidad entre recursos de la Web Semántica.
- **Capa de Confianza.** La capa de confianza evalúa las pruebas ofrecidas por la capa anterior para comprobar de forma exhaustiva si las fuentes de información son confiables.
- **Capa de Firma Digital.** La capa de firma digital define el ámbito de confianza para las capas de pruebas y confianza de manera que permita a los ordenadores y agentes software verificar la seguridad de la información adjuntada y que ésta se envió por una fuente confiable.

En la implementación de la metodología propuesta en esta tesis doctoral se ha hecho uso intensivo de la tecnología más importante de la Web Semántica: las ontologías. Las ontologías constituyen el medio principal para lograr el objetivo de la Web Semántica, ya que facilitan la definición formal de las entidades y conceptos presentes en los diferentes dominios, la jerarquía que les sustenta y las diferentes relaciones que los unen entre sí (Chang-Shing et al., 2007). De esta manera se garantiza una representación del conocimiento consensuada y reutilizable que puede ser compartida y utilizada automáticamente por cualquier sistema informático. En la siguiente sección se describe en profundidad el concepto de ontología, sus elementos y los lenguajes utilizados para la representación de las mismas.

1.2.3 Ontologías

1.2.3.1 Definición

El término ontología (del griego *οντος*, 'del ente', y *λόγος*, 'ciencia, estudio, teoría') hace referencia a una rama de la metafísica que estudia la naturaleza de la existencia (Hofweber, 2009).

A lo largo de la historia han aparecido multitud de definiciones para el término ontología (Quine, 1961; Neches et al., 1991), pero la definición más extendida en el área de inteligencia artificial es la de Thomas R. Gruber (Gruber, 1993), quién dice que: *“Una ontología es una especificación explícita de una conceptualización”*. Este autor se refiere a que es posible representar cualquier objeto del mundo actual en el que vivimos mediante conceptos y relaciones, profundizando lo necesario según nuestras necesidades.

Nicola Guarino en (Guarino, 1995) propuso una definición alternativa de ontología: *“En el sentido filosófico, podemos referirnos a una ontología como un sistema particular de categorías que representa una cierta visión del mundo. Como tal, este sistema no depende de un lenguaje particular: la ontología de Aristóteles es siempre la misma, independientemente del lenguaje usado para describirla. Por otro lado, en su uso más típico en IA, una ontología es un artefacto ingenieril constituido por un vocabulario específico para describir una cierta realidad, más un conjunto de supuestos explícitos concernientes al significado pretendido de las palabras del vocabulario. Este conjunto de supuestos tiene generalmente la forma de teorías lógicas de primer orden, donde las palabras del vocabulario aparecen como predicados unarios o binarios, respectivamente llamados conceptos y relaciones. En el caso más simple, una ontología describe una jerarquía de conceptos relacionados por relaciones de subsunción; en los casos más sofisticados, se añaden axiomas para expresar otras relaciones entre conceptos y restringir la posible interpretación”*.

Nico Borst en (Borst, 1997) refina la definición dada por Gruber, quedando así: *“Una ontología es una especificación formal de una conceptualización compartida”*.

Rudi Studer y sus colegas (Studer et al., 1998) se encargaron algo más tarde de fusionar las definiciones de Gruber y Borst, estableciendo la definición que, a su vez, ha sido adoptada para desarrollar esta investigación: *“Una ontología es una especificación formal y explícita de una conceptualización compartida”*.

Sin embargo, a día de hoy, podemos afirmar que sigue sin existir una definición consensuada del término ontología. Autores como Tom Gruber, William Nico Borst o Rudi Studer, que han proporcionado las definiciones más reconocidas y aceptadas, siguen replanteándose las mismas. Este es el caso de Tom Gruber, que afirmó en octubre de 2004 que modificaría la definición que proporcionó de ontología (Gruber, 2004).

Actualmente las ontologías conforman una parte esencial de la Web Semántica, tal y como aparece en (Chang-Shing et al., 2007). Además, son utilizadas en multitud de aplicaciones, como la instanciación automática de ontologías (Ruiz-Martínez, 2012; Valencia-García et al., 2008), traducción automática (Negacy et al., 2014; Lehmann et al., 2012; Lu et al., 2002), recuperación de información (Koza, 2015; Li et al., 2009), mantenimiento de software (Edgar-Serna et al., 2014; Ruiz et al., 2004), diseño de servicios Web semánticos (Vega & Umaña, 2014), *Data warehouse* y diseño de repositorios institucionales (Durán et al., 2014), sistemas pregunta-respuesta (Talitha et al., 2014; Amorim et al., 2012; Amorim et al., 2011a; Amorim et al., 2011b; Akerkar et al., 2010; Balduccini et al., 2008; Sasaki, 2005), información geográfica (Janowicz et al., 2011; Goodchild, 2010; Kauppinen et al., 2006), integración de esquemas (Rivero et al., 2014), entornos *e-learning* (Patel et al., 2015) o librerías digitales (Abascal-Mena, 2009), entre otras.

De todo lo anterior se deduce que las ontologías son tecnologías que permiten representar conocimiento de manera estructurada y formal, para que éste sea procesable por las máquinas de forma automática. Una muestra de la importancia actual de las ontologías es la cantidad de metodologías que han surgido para su creación (Ochoa et al., 2013), contando también con la diversidad de idiomas en los que se han desarrollado. En la metodología propuesta en esta tesis doctoral las ontologías juegan un papel fundamental, ya que de su correcto diseño, estructura y completitud dependen directamente los resultados obtenidos en los procesos de identificación de características y cálculo del análisis de sentimientos en las opiniones de los usuarios.

1.2.3.2 Elementos de las ontologías

Las ontologías proporcionan un vocabulario común de un área y definen, a diferentes niveles de formalismo, el significado de los términos y las relaciones entre ellos. El

conocimiento en ontologías se formaliza principalmente usando cinco tipos de componentes (Gruber, 1993): clases, atributos, relaciones, axiomas e instancias.

- **Clases.** Se suele usar tanto el término “clases” como “conceptos”. Un concepto representa cualquier entidad que se puede describir, tiene asociado un identificador único, puede poseer diferentes atributos y establecer relaciones con otros conceptos. Las clases en la ontología se suelen organizar en taxonomías. Algunas veces, la noción de ontología se diluye, ya que las taxonomías se consideran ontologías completas (Studer et al., 1998).
- **Atributos.** Los atributos representan la estructura interna de los conceptos. Atendiendo a su origen, los atributos se clasifican en específicos y heredados. Los específicos son los propios del concepto al que pertenecen, mientras que los heredados vienen dados por las relaciones taxonómicas en las que el concepto desempeña el rol de hijo y, por tanto, hereda los atributos del padre. Los atributos se caracterizan por el rango en el cual pueden tomar valor.
- **Relaciones.** Las relaciones representan un tipo de interacción entre los conceptos del dominio. Se definen formalmente como cualquier subconjunto del producto cartesiano de n conjuntos, esto es: “ $R: C_1 \times C_2 \times \dots \times C_n$ ”. Las relaciones normalmente suelen ser binarias, esto es, entre dos conceptos. No todas las relaciones tienen el mismo significado. Existen relaciones binarias de especialización como “*is-a*” o de composición como “*part-whole*”, que se pueden modelar con distintas propiedades como la simetría, reflexividad, transitividad, asimetría, etc.
- **Axiomas.** Los axiomas son expresiones que son siempre ciertas, es decir, modelan las “verdades” que siempre se cumplen en el modelo y en el dominio. Pueden ser incluidas en una ontología con muchos propósitos, tales como definir el significado de los componentes ontológicos, definir restricciones complejas sobre los valores de los atributos, argumentos de relaciones, etc. Estos axiomas verifican la corrección de la información especificada en la ontología y pueden ayudar en la generación de nuevo conocimiento.
- **Instancias.** Las instancias son las ocurrencias en el mundo real de los conceptos. En una instancia todos los atributos del concepto tienen asignado un valor concreto.

1.2.3.3 Lenguajes para la representación de Ontologías

A lo largo de la historia han ido apareciendo distintos lenguajes ontológicos como forma de representación del conocimiento. En 1997 apareció SHOE (*Simple HTML Ontology Extensions*) (Luke et al., 1997), el primer lenguaje para la definición de ontologías para la Web. Posteriormente surgieron otros lenguajes con funciones similares, entre los que se pueden destacar, como más representativos, *RDF* (*Resource Description Framework*), DAML+OIL (James Hendler & Deborah L. McGuinness, 2000) y, por último, *OWL* (*Web Ontology Language*).

En esta sección se mostrará una breve introducción a los lenguajes para la representación de ontologías *RDF* y *OWL*, ya que son los lenguajes utilizados para la implementación de la metodología propuesta en esta tesis doctoral.

1.2.3.3.1 Resource Description Framework (RDF)

Mientras que *XML* es un lenguaje para modelar datos, *RDF* es un lenguaje para especificar metadatos (Studer et al., 1998). *XML*⁹ falla en la escalabilidad de los datos puesto que el orden de los elementos es antinatural y su mantenimiento es muy difícil y costoso. Por el contrario, *RDF* (*Resource Description Framework*) permite la interoperabilidad entre aplicaciones que intercambian información comprensible por la página Web, para proporcionar una infraestructura que soporte actividades de metadatos (Modoni et al., 2014).

El lenguaje *RDF* está recogido en las recomendaciones del W3C: *Primer*¹⁰, *Concepts*¹¹, *Syntax*¹², *Semantics*¹³, *Vocabulary (Schema)*¹⁴ y *Test Cases*¹⁵.

RDF está basado en la idea de identificar los recursos en la Web usando los *Uniform Resource Identifiers* o URIs, y describiendo los recursos en términos de propiedades simples y valores. Una descripción *RDF* es un conjunto de proposiciones simples

⁹ W3C. XML. <http://www.w3.org/XML/>, última consulta realizada en Abril de 2015.

¹⁰ W3C. RDF Primer. <http://www.w3.org/TR/rdf-primer/>, última consulta realizada en Abril de 2015.

¹¹ W3C. RDF: Concepts and Abstract Syntax. <http://www.w3.org/TR/rdf-concepts/>, última consulta realizada en Abril de 2015.

¹² W3C. RDF/XML Syntax Specification (Revised). <http://www.w3.org/TR/rdf-syntax-grammar/>

¹³ W3C. RDF Semantics. <http://www.w3.org/TR/rdf-mt/>, última consulta realizada en Abril de 2015.

¹⁴ W3C. RDF Vocabulary Description Language 1.0: RDF Schema. <http://www.w3.org/TR/rdf-schema/>, última consulta realizada en Abril de 2015.

¹⁵ W3C. RDF Test Cases. <http://www.w3.org/TR/rdf-testcases/>, última consulta realizada en Abril de 2015.

(también llamadas sentencias o declaraciones), y una proposición se conoce también como una “tripleta”, porque está compuesta de 3 elementos (Klyne & Carroll, 2004):

- **Sujeto:** El sujeto identifica el recurso (persona, lugar o cosa) que la sentencia describe. Un recurso *RDF* puede ser cualquier cosa en un modelo de datos (documento, usuario, producto, etc.). Cada recurso está identificado de forma única a través de una *URI*.
- **Predicado:** representa una propiedad del sujeto. Al igual que ocurre con los sujetos, cada propiedad se identifica con una *URI* única.
- **Objeto:** especifica el valor del predicado para un sujeto. En *RDF* un objeto puede ser otro recurso o un literal. En el caso de que sea un recurso, el objeto definirá otra *URI* que identifique al recurso. Por su parte, los literales son una cadena simple de caracteres u otro tipo de datos primitivo definido por *XML*. En términos de *RDF*, un literal puede contener marcado *XML* pero no es interpretado por *RDF*. El modelo *RDF* distingue a los literales de los recursos restringiendo a los literales ser sujeto o predicado en una declaración.

Una tripleta *RDF* se puede representar también como un grafo dirigido de nodos y arcos que representan los recursos, y sus propiedades y valores. Los sujetos y objetos son nodos, mientras que los predicados son arcos (ver figura 1.3).



Figura 1.3. Ejemplo de una sentencia en *RDF* (tripleta).

A continuación se especifica un ejemplo concreto, donde se muestran una serie de declaraciones o sentencias: hay una persona identificada por la *URI* “<http://www.um.es/people/facinf/klt/contact#me>”, cuyo nombre es “Isidro Peñalver”, cuya dirección de correo electrónico es “isidro@um.es”, y cuyo título es “Ing.”, que podría representarse en un grafo *RDF* como el que aparece en la figura 1.4.

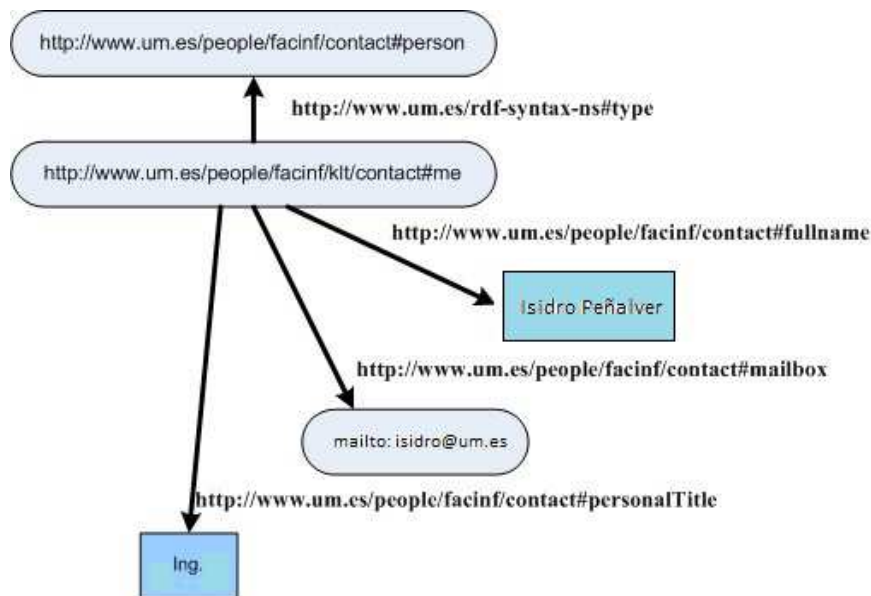


Figura 1.4. Ejemplo de un grafo en *RDF*.

Por último, *RDF* también provee una sintaxis basada en *XML* (llamada *RDF/XML*) para guardar e intercambiar estos grafos. La figura 1.5 muestra la definición en *RDF/XML* del grafo presentado en la figura 1.4.

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf=http://www.um.es/rdf-syntax-ns#type
  xmlns:contact="http://www.um.es/people/facinf/contact#">

  <contact:Person
    rdf:about="http://www.um.es/people/facinf/klt/contact#me">
    <contact:fullName>Isidro Peñalver</contact:fullName>
    <contact:mailbox rdf:resource="mailto:isidro@um.es"/>
    <contact:personalTitle>Ing.</contact:personalTitle>
  </contact:Person>

</rdf:RDF>
```

Figura 1.5. Ejemplo de definición en *RDF/XML*.

1.2.3.3.2 Resource Description Framework Schema (RDFS)

RDF Schema extiende a *RDF* para incluir un amplio vocabulario con un significado adicional (Chang-Su et al., 2015). *RDF* no proporciona lo necesario para indicar tipos o clases específicas de recursos. Para solucionar esto, RDFS proporciona un vocabulario de modelado de datos para *RDF*, constituyendo una extensión del vocabulario básico de

*RDF*¹⁶. Este vocabulario permite expresar clases y sus relaciones jerárquicas, al tiempo que se definen propiedades que se asocian con clases (Cyganiak et al., 2004). De este modo, RDFS facilita mejoras en los procesos de búsqueda y permite hacer inferencias.

El sistema de clases y propiedades de RDFS es similar al sistema de tipos de algunos lenguajes de programación orientados a objetos, con la diferencia de que, en vez de definir clases en términos de las propiedades que las instancias de estas clases deben de tener, RDFS describe propiedades en términos de las clases de recursos a los que éstas se pueden aplicar. Para este propósito, aparecen los roles de ‘dominio’ y ‘rango’. Así, por ejemplo, se puede definir la propiedad *eg: Autor* con un dominio *eg: Documento* y un rango *eg: Persona*. De este modo, es posible definir más propiedades que tengan como dominio *eg: Documento* o como rango *eg: Persona* sin necesidad de modificar la descripción original de estas clases.

Por tanto, se puede definir RDFS como un lenguaje ontológico primitivo que ofrece ciertas cláusulas para el modelado de clases, relaciones de subclases, propiedades, relaciones de sub-propiedades y restricciones de dominio y rango, con un significado fijo (Antoniou & van Harmelen, 2004). Sin embargo, se puede afirmar que RDFS es demasiado primitivo como lenguaje de modelado para la Web y carece de ciertas opciones de modelado interesantes en relación a otros lenguajes más recientes. Entre sus limitaciones, cabe destacar las siguientes (Antoniou & van Harmelen, 2004):

- **Ámbito local de las propiedades.** No permite que las restricciones se apliquen sólo a algunas clases (por ejemplo, ‘*Los leones sólo comen carne*’).
- **Clases disjuntas.** Por ejemplo, la clase *Hombre* versus la clase *Mujer*.
- **Combinaciones booleanas de clases.** Definición de clases mediante unión, intersección, complementario, etc. (por ejemplo, *Persona = Hombre U Mujer*).
- **Restricciones de cardinalidad.** Por ejemplo, una persona sólo tiene 2 progenitores.
- **Características de propiedades.** Transitividad, unicidad, propiedad inversa, etc.

De lo anterior se deduce que el problema de RDFS es que es un lenguaje demasiado simplista, que carece de muchas características de modelado interesantes que otros lenguajes más actuales aportan. No obstante, pese a las limitaciones presentadas, es

¹⁶ W3C. RDF Schema. <http://www.w3.org/TR/rdf-schema/>, última consulta realizada en Abril de 2015.

posible, mediante *RDF* y *RDFS*, realizar diversas tareas de razonamiento automáticas para inferir nuevas relaciones a partir de una base de conocimiento dada.

1.2.3.3 *Web Ontology Language (OWL)*

OWL, desarrollado por el grupo de trabajo de ontologías Web del W3C, es un lenguaje ontológico para la Web basado en *RDF* que facilita su publicación (McGuinness & Van Harmelen, 2004). Ha sido diseñado para representar la información sobre las categorías de los objetos y cómo los objetos están relacionados entre sí. El lenguaje *OWL* proporciona un mecanismo para interpretar el contenido de la Web que supera los propuestos previamente, *XML*, *RDF*, *RDFS*, suministrando vocabulario adicional junto con una semántica formal.

En el diseño del lenguaje *OWL* hubieron multitud de recomendaciones, tal y como aparece en (Horrocks et al., 2003): “*peticiones que incluyen influencias de formalismos establecidos y los paradigmas de representación del conocimiento*”. Fue construido pensando en que este lenguaje sería la base para la Web Semántica, ya que desde el principio fue promovido por el W3C. Por tal razón, tenía que encajar con el conjunto de lenguajes existentes hasta ese momento, incluidos *XML* y *RDF*. *OWL*, al definirse como un lenguaje de ontologías, tenía que ser capaz de representar a un grupo de características ontológicas útiles. Otra petición que se realizó al diseñarse este lenguaje es que fuera lo más compatible posible con los lenguajes existentes como SHOE (Heflin et al., 1999), OIL (Fensel et al., 2001) y DAML + OIL (Connolly et al., 2001).

En este contexto, no fue sencillo encontrar una solución única para *OWL*. Las diferentes versiones de *OWL* han venido acompañadas de diferentes sub-lenguajes expresivos con los que construir las ontologías. La primera versión de *OWL* definió los lenguajes *OWL-Lite*, *OWL DL* y *OWL FULL* con una expresividad creciente. Cada sub-lenguaje presenta características diferentes que se analizan a continuación:

- **OWL Lite:** es el sub-lenguaje más simple. Permite una jerarquía de clasificación y restricciones simples (restricciones de cardinalidad, pero sólo valores de cardinalidad de 0 ó 1). Entre las ventajas, se encuentran la facilidad de entender (por parte de los usuarios) y la facilidad de implementar (para constructores de herramientas). También proporciona una ruta de migración rápida de tesauros y otras taxonomías. La mayor desventaja es, por supuesto, la restringida

expresividad. En (Pulido et al., 2006), se dice que *OWL Lite* es un lenguaje expresivo de ontologías con la inferencia deseable.

- **OWL DL** (*OWL Description Logic*): proporciona la máxima expresividad posible garantizando que se mantiene la integridad de cómputo y la decidibilidad en un tiempo finito para tareas de inferencia. Incluye todos los constructores del lenguaje *OWL*, pero solamente se pueden utilizar bajo ciertas restricciones. *OWL DL* se llama así debido a su correspondencia con la lógica de descripción, un campo de investigación que ha estudiado las lógicas que constituyen la base formal de *OWL*. *OWL DL* es poco complejo y permite un razonamiento e inferencia eficaces, y ahorra tiempo en su uso, pero pierde la compatibilidad con *OWL Full* (Pulido et al., 2006).
- **OWL Full**: con este sub-lenguaje se pueden utilizar todos los constructores y primitivos definidos en *OWL* y, además, aporta la libertad sintáctica ofrecida por *RDF*. Sin embargo, carece de garantías en cuanto a la propiedad de ser computable. Permite asimismo a una ontología aumentar el significado del vocabulario pre-definido, y se puede ver como una extensión de *RDF*, tanto a nivel de sintaxis como de semántica.

En la figura 1.6 se muestra una representación gráfica del alcance, a nivel de expresividad, de cada uno de los lenguajes vistos: *OWL-Lite*, *OWL DL* y *OWL Full*. Como se puede apreciar en la figura, *OWL Full* engloba, en términos de expresividad, a los lenguajes *OWL DL* y *OWL Lite*.

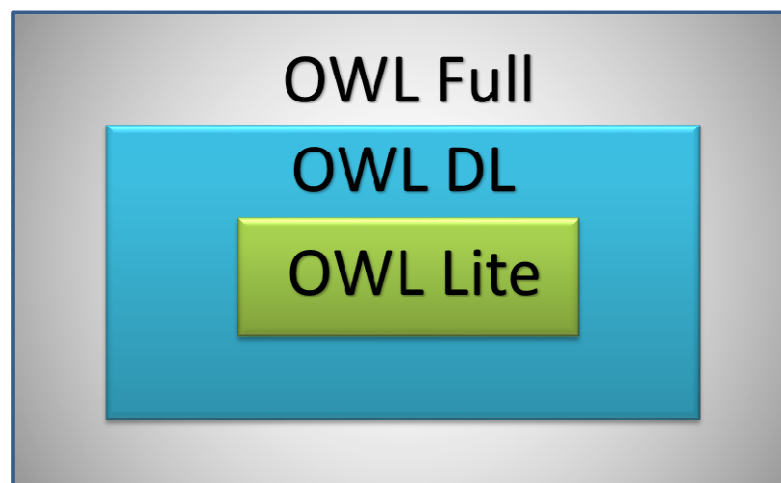


Figura 1.6. Versiones del lenguaje *OWL*.

La primera versión del lenguaje *OWL* apareció en julio de 2002 y fue presentado de manera formal por la *W3C* en 2004. En el año 2009 apareció la segunda y última versión hasta el momento de este lenguaje, la versión 2.0, que tiene una estructura muy similar a la versión 1, pero con una sintaxis diferente. La sintaxis funcional en *OWL* 2 se diferencia en la forma de la sintaxis abstracta de *OWL* 1, pero el rol que desempeña sobre la estructura *OWL* es exactamente el mismo, de ahí que la compatibilidad de las versiones sea casi total. La sintaxis de *OWL* 2 es muy similar a la representación con grafos *RDF* mostrada anteriormente.

OWL 2.0 añade nuevas funcionalidades con respecto a *OWL* 1.0 para mejorar la expresividad del lenguaje. Entre las nuevas características que aporta este lenguaje se pueden destacar las siguientes: (i) definición de claves en las clases, (ii) cadenas de propiedades, (iii) tipos de datos y rangos de datos más complejos, (iv) restricciones de cardinalidad cualificadas, (v) propiedades simétricas, reflexivas y disjuntas, y (vi) mejora de las características de las anotaciones.

Por otro lado, *OWL* 2 tiene una estructura de sub-lenguajes distinta a *OWL* 1. En esta versión se definen tres perfiles, siendo cada uno de ellos más restrictivo que *OWL* DL. Los diferentes perfiles existentes en *OWL* 2 son los siguientes (Cardoso et al., 2015):

- ***OWL* 2 EL:** es un subconjunto de *OWL* 2 que está diseñado para aplicaciones que trabajen con ontologías que contienen un gran número de propiedades y/o clases. El acrónimo “EL” en el nombre del perfil refleja la incorporación a la familia de lógicas descriptivas EL, que permite realizar tareas básicas de razonamiento en tiempos polinómicos aunque su expresividad sea limitada restringiendo el uso de cuantificadores universales.
- ***OWL* 2 QL:** es un subconjunto de *OWL* 2 destinado para aplicaciones que requieren de grandes volúmenes de instancias de datos y donde la consulta de información es la tarea más importante de razonamiento. En *OWL* 2 QL se permite la unión de consultas utilizando los sistemas de base de datos relacionales. Como en el perfil *OWL* 2 EL, incorpora algoritmos de tiempo polinomial para comprobar la consistencia de la ontología y las expresiones lógicas. El acrónimo QL relaciona el perfil con las consultas de información y con el hecho de que permite la utilización de un lenguaje de consultas estándar relacional.
- ***OWL* 2 RL:** es un subconjunto de *OWL* 2 destinado para aplicaciones que requieren un razonamiento escalable sin sacrificar el poder de la expresividad. Este perfil ha sido diseñado para permitir que las aplicaciones *OWL* 2 pierdan

poca expresividad y ganen mucha eficiencia. También ha sido diseñado para aplicaciones RDFS que necesitan expresividad añadida. Los sistemas de razonamiento *OWL 2 RL* pueden utilizar sistemas de razonamiento basados en reglas. Problemas como la consistencia de la ontología, satisfacibilidad de las expresiones, subsunción de las expresiones, chequeo de instancias y unión de consultas pueden ser solucionados en tiempo polinómico respecto al tamaño de la ontología. El acrónimo *RL* refleja el hecho de que el razonamiento en este perfil puede ser implementado mediante un lenguaje de reglas estándar.

La figura 1.7 representa gráficamente la estructura de sub-lenguajes de *OWL 2.0*. A diferencia de *OWL 1.0*, la organización de los sub-lenguajes no es estrictamente jerárquica, sino que en esta versión los diferentes sub-lenguajes comparten características entre ellos. La lógica descriptiva se encuentra presente en cada uno de ellos y la perspectiva del diseño cambia radicalmente con respecto a la versión 1.0. En este caso, cada uno de los sub-lenguajes ha sido diseñado para ofrecer importantes ventajas en escenarios concretos de aplicación, en lugar de estar jerárquicamente diseñados como ocurre en el caso de la versión *OWL 1.0*.

OWL 2 ha incorporado algunas nuevas funcionalidades con respecto a la versión 1 (Zhang et al., 2015). Una de ellas es la función de “*punning*”¹⁷, que permite la definición de recursos como instancias y clases al mismo tiempo. Esta función permite definir una clase y una instancia asociada a la misma clase y que, además, compartan la misma *URI*, a diferencia de *OWL* versión 1, donde a cada recurso se le asigna una *URI* diferente. Por lo tanto, *OWL 2* permite agrupar los recursos a través de una *URI* que representa un recurso que, en función del contexto, puede ser tratado como una clase o como una instancia. A nivel de clase, estos recursos pueden tener a su vez sub-clases, y a nivel de instancia pueden contener otras instancias del modelo ontológico. El objetivo de esta nueva funcionalidad es facilitar la reutilización a nivel de instancia de la semántica de una conceptualización de un dominio a nivel de clase.

¹⁷ W3C. OWL2 New Features. <http://www.w3.org/TR/owl2-new-features/>, última consulta realizada en Abril de 2015.



Figura 1.7. Perfiles del lenguaje OWL2.

Por otro lado, en *OWL 2* se definen diferentes sintaxis que añaden diferentes funcionalidades al lenguaje. Todas las sintaxis deben estar obligatoriamente basadas en *RDF/XML*. De hecho, según el estándar, todas las herramientas que implementen *OWL 2* deben de poder leer y escribir usando esta sintaxis. La sintaxis *RDF/XML* (W3C, 2014) permite representar la información de las ontologías para que se puedan almacenar en documentos *OWL* y, además, tiene por objeto servir como medio de intercambio de documentos *OWL 2*. Por su parte, el objetivo de la sintaxis *OWL/XML* (W3C, 2012b) es cumplir el estándar *XML* para que puedan reutilizarse herramientas *XML* para procesar este tipo de ontologías. En *OWL 2* también se ha definido una sintaxis funcional (W3C, 2012a) orientada a facilitar la lectura de la estructura de las ontologías. Finalmente, la sintaxis Turtle (Beckett & Berners-Lee, 2008) facilita la lectura y escritura de tripletas *RDF*.

Como conclusión, se puede afirmar que *OWL 2* se originó como consecuencia de las peticiones realizadas por numerosos usuarios de *OWL* ante las carencias de *OWL 1*. *OWL 2* añade características indispensables para satisfacer escenarios particulares, con el único objetivo de facilitar el desarrollo de las ontologías y su distribución a través de la Web, haciendo el contenido Web más accesible para las máquinas. La elección entre los diferentes perfiles (a saber, *OWL 2 EL*, *OWL 2 QL* y *OWL 2 RL*) depende, principalmente, de la expresividad necesaria para la aplicación, la prioridad dada al razonamiento sobre las clases o los datos, el tamaño de los conjuntos de datos y la importancia de la escalabilidad.

1.2.4 Web Semántica y ontologías aplicadas en minería de opiniones

La aplicación de la Web Semántica y, más concretamente, las ontologías en el proceso de minería de opiniones basada en características no ha sido muy generalizado a lo largo de la historia. Sin embargo, se pueden destacar algunos trabajos importantes sobre ontologías y minería de opiniones basadas en características (Zhou & Chaovalit, 2008; Zhao & Li, 2009). El objetivo de estos trabajos es calcular la polaridad teniendo en cuenta las características de un concepto en una ontología. Se utilizan ontologías del dominio que modelan el contenido de los textos en el corpus y, a partir de dichas ontologías, extraen los adjetivos que describen dichas características.

En el trabajo de Zhou y Chaovalit (2008), la polaridad de un texto se calcula basándose en los pesos de las características, y el valor de la polaridad de la característica se calcula mediante la “estimación de máxima verosimilitud”. En el trabajo de Zhao y Li (2009), por su parte, se obtiene la orientación semántica del texto desde la jerarquía de la ontología. Además, el método que proponen es capaz de obtener la positividad, negatividad y neutralidad de un texto. Estos dos trabajos demuestran que el uso de ontologías mejora los resultados en el problema del análisis de sentimientos.

En (Baldoni et al., 2012) se combinan léxicos computacionales de sentimientos disponibles en la Web con una ontología de categorías emocionales. Este trabajo está ambientado en el marco de la llamada Web Semántica Social, ya que implica a los usuarios finales en el proceso de anotación de recursos a través de una plataforma online. Se desarrolla un software llamado ArsEmotica para asociar emociones a un conjunto de obras de arte, siendo los resultados de la validación del mismo prometedores.

En (Narock et al., 2012) se propone un algoritmo para mejorar el cálculo de la similitud semántica entre características de opiniones e instancias de ontologías. Para ello se utilizan técnicas de minería de opiniones y cálculo de polaridades.

Más recientemente, en el trabajo de (Farman et al., 2015) se propone un nuevo sistema de extracción de información y minería de opiniones basada en características empleando una ontología llamada T2FOBOMIE. El sistema propuesto recupera opiniones de usuarios sobre hoteles y extrae las características de las mismas en base a una ontología del dominio. La ontología del dominio, la información del usuario y la información del hotel se integran para formar una ontología fusionada de tipo 2 que se empleará para calcular la polaridad de las características y la polaridad individual del

hotel. Los resultados de los experimentos demostraron que el uso de la ontología tipo 2, T2FOBOMIE, es prometedor para llevar a cabo el análisis de opiniones de usuarios en el dominio hotelero.

Finalmente, en (Shaban-Nejad et al., 2015) se lleva cabo un trabajo sobre el análisis de sentimientos de las opiniones de los usuarios sobre ciertas vacunas para prevenir determinadas enfermedades. Esta aproximación presenta un enfoque para el diseño y desarrollo de una plataforma semántica integrada que hace uso de las tecnologías de la Web Semántica, el análisis de sentimientos y las ontologías en el ámbito de la medicina. Se construye una ontología llamada VASON que se utiliza para llevar a cabo el análisis de sentimientos de opiniones. Esta plataforma está en vías de implantación y tiene como objetivo proporcionar conocimiento sobre los factores que impulsan el rechazo de una vacuna determinada.

En esta tesis doctoral se propone una metodología alternativa para la minería de opiniones basada en características donde se aprovechen las bondades de las ontologías, expuestas en los trabajos citados, junto con el empleo de métodos matemáticos avanzados en el proceso de análisis de sentimientos de las opiniones de los individuos.

1.3 Procesamiento del lenguaje natural

1.3.1 Introducción

La disciplina que se ocupa del tratamiento automático del lenguaje natural en textos se llama “Procesamiento del Lenguaje Natural” (PLN). El PLN es parte del área de investigación de la Inteligencia Artificial (IA), que se define como “*la ciencia y la ingeniería de hacer máquinas inteligentes*” (McCarthy, 1959) mediante la simulación de los mecanismos de la inteligencia humana.

“El concepto de PLN hace referencia a las técnicas de tratamiento del lenguaje y su aplicación en diversas áreas por medio de métodos computacionales. Paralelamente se hace uso de otros términos próximos, como lingüística computacional o ingeniería lingüística”, esta definición fue aportada en (Sosa, 1997).

Según (Darriba, 2007), “*el PLN se define como el desarrollo de modelos computacionales de determinados aspectos de lenguaje humano para que, partiendo de*

esos modelos se puedan realizar programas capaces de comprender o producir enunciados en lenguaje natural”.

“El procesamiento del lenguaje natural es un conjunto de técnicas computacionales teóricamente motivadas para analizar y representar naturalmente textos de origen natural en uno o más niveles de análisis lingüísticos para lograr el propósito de procesar el lenguaje humano por una serie de tareas o aplicaciones”. Esta definición fue aportada en (Liddy, 2001), y hace hincapié en la capacidad de estos sistemas para llevar a cabo diferentes niveles de análisis lingüísticos.

En (Chowdhury, 2003) se define *PLN* como *“un área de investigación que explora cómo las computadoras pueden utilizarse para entender y manipular texto escrito en lenguaje natural o del habla para hacer operaciones útiles”.*

Todas estas definiciones sobre *PLN* tienen algo en común: establecer el objetivo del procesamiento del lenguaje natural en el desarrollo de técnicas y construcción de herramientas, que permitan a los sistemas informáticos entender los lenguajes naturales para facilitar la comunicación entre seres humanos y máquinas.

Hasta la década de 1980, la mayoría de los sistemas de *PNL* se basaban en un complejo conjunto de reglas diseñadas a mano. A partir de finales de 1980 hubo una revolución en *PNL* con la introducción de algoritmos de aprendizaje automático para el procesamiento del lenguaje. Durante los últimos años esta área de investigación ha seguido en constante movimiento. En primer lugar, los modelos probabilísticos y los modelos conducidos por datos se encuentran bastante estandarizados en todo el procesamiento del lenguaje natural. Los algoritmos de análisis, etiquetado gramatical, resolución de referencia y discurso han comenzado a incorporar modelos probabilísticos y a emplear metodologías de evaluación en el reconocimiento de voz y recuperación de la información. Estos años se han caracterizado por el incremento del nivel de procesamiento y la rapidez de los ordenadores, lo que ha permitido la explotación de diferentes áreas del *PLN*, como el reconocimiento de la voz y la detección de errores ortográficos y gramaticales. Por último, el auge de la Web ha resaltado la necesidad de innovación en los procesos de minería de opinión y extracción de información de los sistemas de almacenamiento.

En conclusión, el *PLN* es una técnica esencial de la inteligencia artificial que tiene como propósito el modelado y procesamiento computacional del lenguaje humano. Todo sistema de *PLN* intenta simular el comportamiento lingüístico humano. Para ello debe

tomar conciencia tanto de las estructuras propias del lenguaje, como de su dominio de aplicación. Es decir, un sistema de *PLN* deberá conocer los significados de las palabras y cómo éstas afectan al significado global del discurso.

1.3.2 Información expresada en lenguaje natural

Tradicionalmente los esfuerzos para la gestión de la información en las empresas se han centrado principalmente en los datos estructurados y semiestructurados, con el único objetivo de sacar el máximo partido posible a las costosas inversiones en bases de datos y sistemas transaccionales. Sin embargo, en muchos casos, la información pertinente no se encuentra en fuentes estructuradas (es decir, tablas o bases de datos), sino en documentos no estructurados, escritos en lenguaje humano. La elevada cantidad en la que estos datos se encuentran en la Web requiere el uso de técnicas de procesamiento automático para poder tratarlos (Balahur & Montoyo, 2010).

La información no estructurada la encontramos en fuentes tales como documentos, la Web y cualquier otro medio que contenga información expresada en lenguaje natural y en formatos muy diversos como texto, videos, audio o imágenes. Este tipo de información ocupa más del 94 % del universo digital (Gantz et al., 2008). Con la llegada de Internet la información no estructurada (como el correo electrónico, los informes, contratos, facturas, formularios, hojas de cálculo, imágenes, presentaciones, etc.) ha llegado a ser muy importante. De hecho el 80% de la información de una empresa reside hoy en estos contenidos no estructurados que son vitales para las operaciones diarias de los departamentos más estratégicos de la empresa, desde marketing a recursos humanos y finanzas (Mazón, 2008).

Los buscadores actuales de información en la Web (*Google, Yahoo, MSN Search, Altavista, Lycos*, etc.) hacen que las necesidades de los usuarios se satisfagan únicamente de forma parcial. El gran número de documentos irrelevantes que usualmente recuperan estos buscadores, y el gran número de documentos relevantes que estos sistemas no recuperan, son un obstáculo substancial para una mejor utilización de los recursos de información disponibles en la Web (Avancini, 2000).

1.3.3 Niveles del lenguaje natural

El lenguaje humano puede ser dividido en distintos niveles de cara a su análisis y entendimiento. En (Liddy, 2001; Feldman, 1999) se muestran los siete niveles del lenguaje que utilizan los humanos para extraer el significado del lenguaje natural. La tabla 1.1 muestra un cuadro resumen con los distintos niveles del lenguaje natural, su ámbito de actuación, las herramientas de procesamiento que podríamos utilizar para analizarlos de forma automática y las estructuras lingüísticas obtenidas en cada nivel.

Los niveles del lenguaje humano no son niveles aislados sino que están interrelacionados. Además, existen niveles que, para desarrollar su proceso de análisis, requieren conocimiento de niveles anteriores, o incluso posteriores. Por ejemplo, para los procesos de desambiguación de las categorías morfológicas es necesario recurrir a la sintaxis para determinar qué función desempeña un término en una frase.

A continuación se explican en detalle las características más importantes de cada uno de los niveles del lenguaje humano según Liddy (2001) y Feldman (1999). Los siete niveles son los siguientes:

- **Nivel fonológico.** La fonética es la parte de la lingüística que estudia los sonidos del lenguaje humano. El funcionamiento de un sistema de procesamiento fonológico acepta como entrada la voz humana, y analiza y codifica las ondas producidas en una señal digitalizada para interpretar varias reglas o para establecer comparaciones con otro modelo de lenguaje.
- **Nivel morfológico.** La morfología es la parte de la lingüística que estudia las reglas que rigen la flexión, la composición y la derivación de las palabras. El funcionamiento de un sistema de procesamiento morfológico transforma cada secuencia de caracteres en una secuencia de morfemas, mediante la utilización de diferentes técnicas como lematizadores, o etiquetadores *POS*. Con estas técnicas es posible determinar aspectos como tiempo, género, número, grado, etc. y buscar sufijos, prefijos, sinónimos, generalizaciones, o especializaciones. También permiten clasificar las unidades lingüísticas en categorías gramaticales (sustantivo, verbo, adjetivo, adverbio, etc.). Por último, los sistemas de procesamiento morfológico pueden reconocer el significado de cada morfema para obtener el significado total de la palabra.

Tabla 1.1. Niveles del lenguaje natural y herramientas de *PLN*.

Niveles del Procesamiento del Lenguaje Natural	Ámbito de actuación	Herramientas Utilizadas	Resultados del Procesamiento
Fonético	Sonidos	Corpus de Aprendizaje Modelos Acústicos Diccionario de Unidades de Síntesis	Fonemas
Léxico	Formas	Lematizador Etiquetado POS Lexicón Computacional	Palabras
Morfológico	Formas	Lematizador Etiquetado POS Lexicón Computacional	Palabras
Sintáctico	Estructuras	Bases de Datos Sintácticas Treebank	Frases
Semántico	Significados	Bases de Datos Semánticas Lexicón Computacional Ontologías	Relaciones
Discurso	Comunicación	Bases de Datos Semánticas Ontologías	Textos
Pragmático	Comunicación	Bases de Datos Semánticas Ontologías	Textos

- **Nivel léxico.** La lexicología es la parte de la lingüística que estudia el léxico, sus unidades y las relaciones entre ellas dentro del sistema de la lengua. El funcionamiento de un sistema de procesamiento léxico se centra en la interpretación del significado de las palabras de manera individual. En general, los sistemas de *PLN* requieren de un léxico que les permita conocer las categorías léxicas existentes en el lenguaje. En este nivel existen diferentes técnicas que permiten el análisis léxico de la palabra, pero la técnica más extendida se basa en la asignación de etiquetas individuales a cada palabra del texto. Estas etiquetas determinan la categoría léxica de la palabra y, en el caso de que una palabra pueda desempeñar diferentes funciones dentro de la sentencia, será etiquetada con la categoría léxica más probable en función del contexto. El proceso de análisis léxico finaliza cuando todas las palabras del texto han sido etiquetadas.
- **Nivel sintáctico.** La sintaxis es la parte de la lingüística que estudia el orden y la relación de las palabras o sintagmas en la oración, así como las funciones que cumplen. Un sistema de procesamiento sintáctico analiza la estructura gramatical de la secuencia de palabras y produce una representación de su estructura. Este proceso de análisis requiere contar con los patrones sintácticos más frecuentes del lenguaje que se esté analizando y un analizador sintáctico que se encargue de obtener la estructura gramatical de la secuencia de unidades léxicas para representarla en forma de árbol o red. Por lo tanto, el resultado de este nivel de procesamiento es una representación de la estructura de la oración que pone de manifiesto las relaciones de dependencia estructural de las palabras.
- **Nivel semántico.** La semántica es la parte de la lingüística que estudia el significado de las expresiones lingüísticas. En principio cualquier medio de expresión (lenguaje formal o natural) admite una correspondencia entre palabras y situaciones o conjuntos de cosas que se encuentran en el mundo físico, y que pueden ser descritas por dicho medio de expresión. A partir de la estructura generada por el análisis sintáctico, se genera otra estructura o forma lógica asociada, que representa el significado o sentido de la sentencia, independientemente del contexto. El objetivo del procesamiento semántico es determinar los posibles significados de una frase, centrándose en las relaciones de significados a nivel de palabra. Este nivel de procesamiento puede requerir de técnicas de desambiguación semántica para aquellas palabras que tengan varios significados. La desambiguación semántica permite seleccionar y representar un solo significado de palabras polisémicas. Actualmente, existen varios métodos

que pueden ser implementados en los sistemas de procesamiento semántico para eliminar las ambigüedades. Algunos de estos métodos precisan de información relacionada con la frecuencia en que aparece cada tópico de interés en el corpus y otros utilizan el conocimiento pragmático del dominio del documento.

- **Nivel del discurso o contextual.** En este nivel el sistema no interpreta los textos como conjuntos independientes de sentencias concatenadas, sino que se centra en analizar las propiedades del texto en su conjunto que transmiten significado al conectar entre las oraciones que lo componen. El objetivo de este nivel de procesamiento es utilizar la estructura semántica del nivel semántico para desarrollar una interpretación final de la oración en función de circunstancias del contexto. El contexto de la palabra es considerado como un conjunto de palabras que la acompañan, junto con las relaciones sintácticas y categorías semánticas (Vázquez et al., 2003). Es precisamente este contexto, el que parece haberse demostrado como el medio más eficaz para identificar el sentido de una palabra polisémica. Para lograr este objetivo, existen varios tipos de procesamiento del discurso que pueden darse en este nivel, siendo los más conocidos la resolución anafórica y el reconocimiento de la estructura del texto/discurso (Carrión, 2014). La resolución anafórica se basa en reemplazar palabras vacías semánticamente, como pronombres, por la entidad a la que referencian. El reconocimiento de la estructura del texto/discurso determina las funciones de las sentencias en el texto y añade esta representación significativa al texto.
- **Nivel pragmático:** La pragmática es la parte de la lingüística que estudia las estrategias comunicativas enmarcándolas en un contexto socio-cultural. En este nivel se analiza el texto más allá de los límites de la frase. El análisis que se realiza en este nivel se encuentra relacionado con los factores extralingüísticos que condicionan el uso del lenguaje en situaciones comunicativas concretas. Precisamente todos aquellos factores extralingüísticos que no pueden ser analizados por los niveles anteriores, como la intención comunicativa, el contexto verbal, y la situación o conocimiento del mundo, resultan de gran importancia en este nivel. En este análisis se utiliza el contexto, por encima de los contenidos, para comprender el significado del texto. Para lograr este análisis, los sistemas de procesamiento pragmático disponen de bases de conocimiento y módulos de inferencia que permiten interpretar las intenciones, los planes y los objetivos de un texto.

1.3.4 Recursos lingüísticos en sistemas de PLN

Actualmente, los sistemas de *PLN* hacen uso de distintos recursos lingüísticos, entre los que podemos destacar:

- **Bases de datos léxicas:** se trata de bases de datos con listas de términos, junto a su significado léxico, dentro de un contexto específico. Como ejemplo de base de datos léxica más conocida podemos destacar *WordNet* (Miller, 1995), que es una base de datos léxica del idioma inglés empleada en la implementación de la metodología propuesta en esta tesis doctoral. También es destacable *EuroWordNet*¹⁸, que es una base de datos léxica multilingüe con relaciones semánticas entre las palabras de varios idiomas de la comunidad europea: inglés, holandés, español, italiano, alemán, francés, checo y estonio (Hirst, 1999).
- **Diccionarios electrónicos:** se trata de diccionarios tradicionales en formato electrónico. Los diccionarios electrónicos son bases de datos léxicas que agrupan grandes cantidades de palabras, proporcionando definiciones y almacenando las relaciones semánticas entre los conjuntos de palabras. Como ejemplo de diccionario electrónico cabe destacar *LDOCE*¹⁹, que es uno de los diccionarios más ampliamente utilizados. Tiene aproximadamente 50.000 definiciones de palabras. Una de las características de *LDOCE* es la inclusión de etiquetas por área, las cuales pueden ser equivalentes a las etiquetas de dominio utilizadas para cada sentido de la palabra. Cuenta con una versión especialmente desarrollada para *PLN*, llamada *LDOCE3 NLP database*, muy utilizada en experimentos de *Word Sense Disambiguation (WSD)*.
- **Tesauros:** los tesauros son vocabularios controlados que representan las relaciones semánticas con otras palabras y sus significados. Se trata de agrupaciones de términos en clases para un determinado dominio. Los tesauros documentales son herramientas de almacenamiento y recuperación de información utilizados por un profesional de la documentación para registrar ordenadamente datos e informaciones en un índice y por el usuario para buscar. Este instrumento está compuesto por listados de palabras o términos analizados y normalizados, que guardan entre sí relaciones semánticas y funcionales, y que se utilizan para representar conceptos, temas o contenidos de documentos. Su misión es que los usuarios accedan fácilmente a la información que contiene una biblioteca, un archivo o los centros de documentación. Los tesauros están configurados en torno a términos que se ordenan de forma alfabética, temática y

¹⁸ <http://www.illc.uva.nl/EuroWordNet>

¹⁹ <http://www.ldoceonline.com>

jerárquica; sinónimos (también llamados descriptores) y definiciones de los términos. Están organizados de forma jerárquica, interrelacionados por equivalencia o asociación. Normalmente los tesauros son muy específicos y se dedican a un área temática determinada. Por ejemplo, el *tesauroArtLex*²⁰ es un diccionario de arte que incluye más de tres mil términos con imágenes, y *The Getty Thesaurus of Geographic Names*²¹ contiene nombres geográficos del Getty Information Institute.

- **Redes semánticas:** se trata de representaciones de los términos con una estructura de red para mostrar las jerarquías conceptuales existentes entre los conceptos. Es una forma de representación de conocimiento lingüístico en la que los conceptos y sus interrelaciones se representan mediante un grafo. En caso de que no existan ciclos, estas redes pueden ser visualizadas como árboles. Las redes semánticas son usadas, entre otras cosas, para representar mapas conceptuales y mentales. La figura 1.8 muestra un ejemplo de red semántica de animales.
- **Matrices léxicas:** se trata de representaciones en forma matricial de la relación entre cada palabra y su significado. Mediante esta representación, se pueden obtener las relaciones de polisemia y sinonimia. La información léxica que se requiere para llenar tales matrices se puede obtener no sólo haciendo introspecciones hipotéticas, sino que se puede extraer de corpus textuales. Esto permite reflejar un comportamiento más cercano a la realidad de una lengua natural.

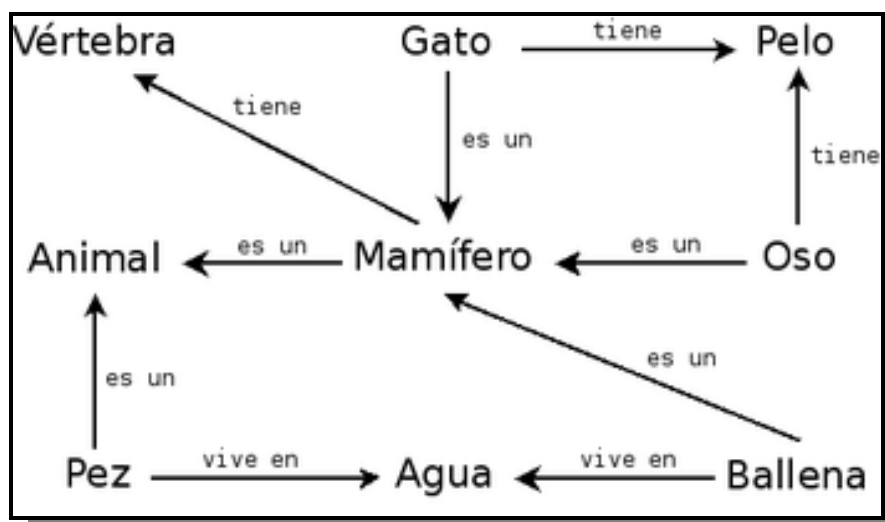


Figura 1.8. Ejemplo de red semántica.

²⁰ <http://www.artlex.com/>

²¹ <http://www.getty.edu/research/tools/vocabularies/index.html>

En la investigación realizada por (Valencia-García, 2005), podemos encontrar una colección de procesadores asociados a cada nivel de lenguaje, de entre los indicados arriba. Por desgracia, cabe destacar que la mayoría de los recursos léxicos solo están disponibles para el idioma inglés.

Como se ha comentado anteriormente, en la implementación de la metodología propuesta en esta tesis doctoral, se ha utilizado la base de datos léxica *WordNet*, ya que se ha empleado un corpus de opiniones de usuarios escritas en inglés. Durante el estudio y comprensión de esta herramienta se observaron diversos solapamientos terminológicos entre *WordNet* y otros recursos lingüísticos. Un primer caso es la confusión con un lexicón computacional (Martí Antonín et al., 2000), recurso con el que *WordNet* tiene algunos puntos en común (que contiene información léxica y se encuentra en soporte informático), pero con el que difiere en una de las características esenciales que define a dicho recurso, la presencia de reglas que permitan realizar inferencias. Un segundo caso es la confusión de *WordNet* con un tesoro (Hirst, 1999; Kilgarriff, 2000), que coincide con *WordNet* en que incluye palabras que en determinados contextos pueden reemplazarse unas a otras (sinónimos). Esta confusión viene dada por la aplicación de *WordNet* a la expansión de búsquedas (*query expansion*) en el campo de la recuperación de información (Voorhees, 1994; Mandala et al., 1999), donde se asume la función de dicha base de datos léxica como fuente de sinónimos. Un tercer caso es la confusión de *WordNet* con una ontología (Berland, 1999; Jones, 1998; Kietz et al., 2000; Burgun, 2001; Fensel et al., 2001). En *WordNet* los nombres no están organizados jerárquicamente entorno a un único concepto que englobe a todos los demás, sino que se agrupan en torno a un conjunto de conceptos denominados *synsets* que originan jerarquías léxicas separadas. Esta estructuración plantea el problema del rol de los *synsets*: mientras unas veces son tratados como elementos léxicos, en otras son tomados como conceptos. Por lo tanto, se plantea la duda de si *WordNet* trabaja con unidades léxicas o con conceptos. La estructuración por jerarquías léxicas separadas es distinta a la utilizada en las ontologías que inician su jerarquía conceptual a partir un único concepto superordinado (*all*).

Finalmente, debemos comentar que muchos de los recursos lingüísticos que vistos en esta sección han constituido la base para el proceso de construcción de ontologías. Dicho de otra manera, la construcción de sistemas de cómputo capaces de generar la arquitectura de una red organizada de conceptos. En la figura 1.9 se muestra esquemáticamente este proceso.

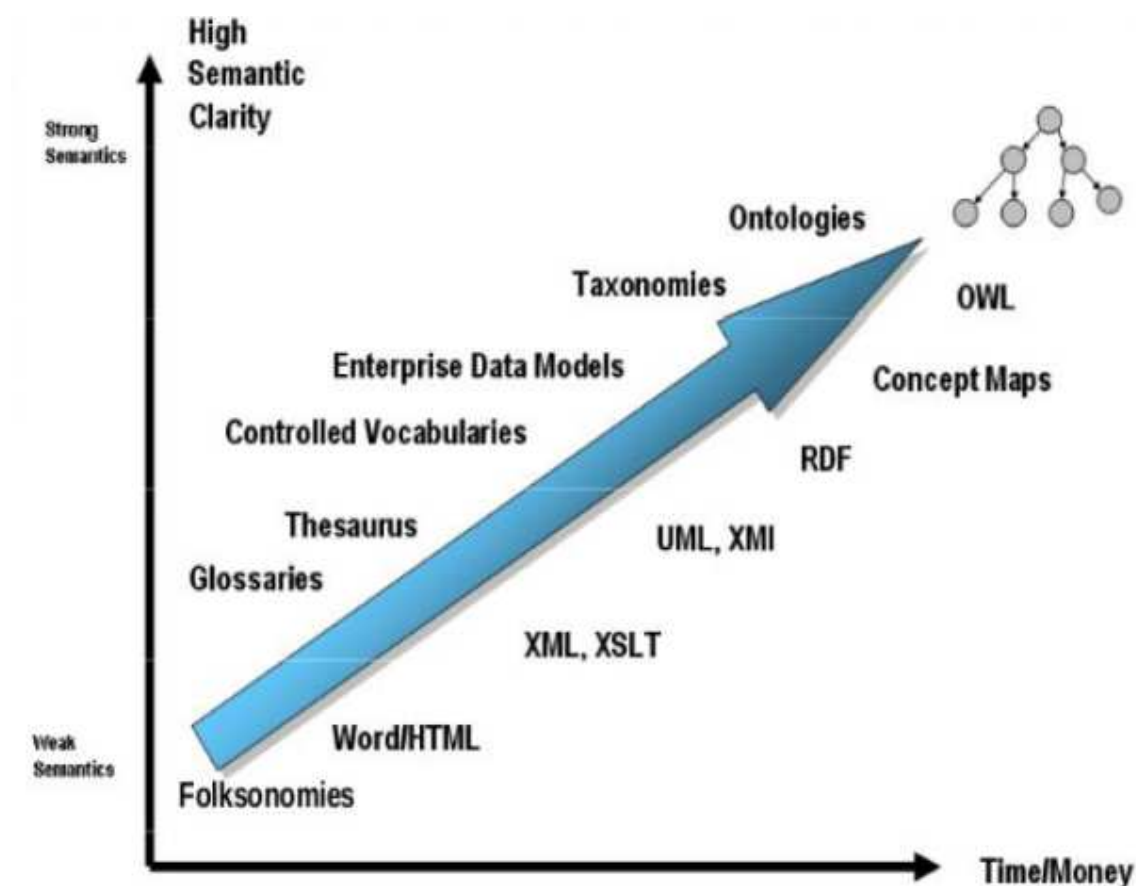


Figura 1.9. Recursos lingüísticos en PLN frente a ontologías.

1.3.5 PLN aplicado en minería de opiniones

Dentro del campo de la inteligencia artificial (IA) existen dos áreas que, desde hace tiempo, han empezado a adquirir una gran importancia: el procesamiento del lenguaje natural (PLN) y la minería de opiniones o análisis de sentimientos. Estas tecnologías permiten ayudar a resolver el complejo problema de la gestión de información, originado por el exceso de información que existe hoy en día.

Así, mientras el PLN se dedica a entender el lenguaje humano para poder explotar el conocimiento lingüístico de los textos (Valencia-García et al., 2004), la minería de opiniones centra en la extracción de información, identificación de patrones en textos de opinión, y la clasificación de la polaridad de las opiniones realizadas por los usuarios (Chen & Zimbra, 2010; Pang et al., 2008). Teniendo en cuenta lo anterior, es lógico que ambas disciplinas se complementen para trabajar conjuntamente en un esquema en el que las computadoras entiendan a los seres humanos.

De todo lo anterior se deduce que cualquier proyecto que se emprenda con el ánimo de llevar a cabo el proceso de minería de opiniones de usuarios en un dominio cualquiera, entre los que se incluye el presente trabajo de tesis doctoral, necesitará llevar a cabo de forma implícita, en mayor o menor medida, procesos relacionados con el procesamiento del lenguaje natural. De hecho, en (Liu, 2006) se define la minería de opiniones o análisis de sentimientos como “*una tarea del procesamiento del lenguaje natural (PLN) que identifica opiniones relacionadas con un objeto*”.

1.4 Espacios Vectoriales

1.4.1 Introducción

Los espacios vectoriales han sido objeto de estudio intensivo a lo largo de la historia en las ciencias matemáticas y físicas (Balakrishnan, 1981; Berezansky, 1996; Birkhoff, 1984; Hutson, 1980; Kreyszig, 1989; Reed, 1972; Yoshida, 1971; Zeidler, 1989; Zeidler, 1995). Según su definición en el álgebra abstracta, un espacio vectorial es una estructura algebraica creada a partir de un conjunto no vacío, una operación interna (llamada suma, definida para los elementos del conjunto) y una operación externa (llamada producto por un escalar, definida entre dicho conjunto y otro conjunto, con estructura de cuerpo) con 8 propiedades fundamentales que se analizarán a continuación. A los elementos de un espacio vectorial se les llama vectores y a los del cuerpo escalares (Strang, 2009). La figura 1.10 muestra una representación gráfica de un espacio vectorial.

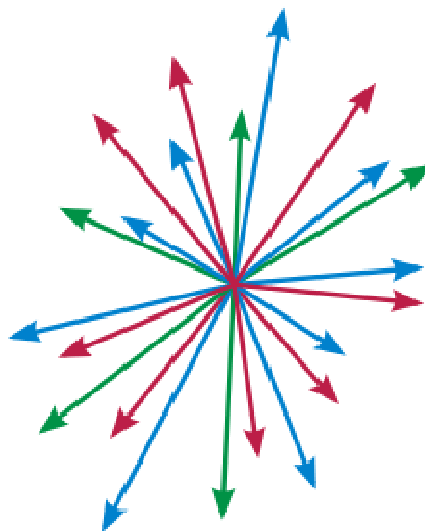


Figura 1.10. Representación artística de un espacio vectorial.

Un espacio vectorial sobre un cuerpo K (como el cuerpo de los números reales o los números complejos) es un conjunto V no vacío dotado de dos operaciones para las cuales será cerrado (la suma y el producto por un escalar). Los elementos de V (que denotaremos como u, v, w) se llaman vectores (Apostol, 2006).

En física un vector (también llamado vector euclidiano o vector geométrico) es “una magnitud física definida por un punto del espacio donde se mide dicha magnitud, además de un módulo (o longitud), su dirección (u orientación) y su sentido (que distingue el origen del extremo)” (Arfken, 1985).

En matemáticas se define un vector como un “elemento de un espacio vectorial” (Apostol, 2006). Esta noción es más abstracta y para muchos espacios vectoriales no es posible representar sus vectores mediante el módulo, la longitud y la orientación. En particular los espacios de dimensión infinita sin producto escalar no son representables de ese modo (Köthe, 1979). Los vectores en un espacio euclídeo se pueden representar geométricamente como segmentos de recta dirigidos (representados como “flechas”) en el plano R^2 o en el espacio R^3 (Zimmer, 1990).

Como se ha comentado anteriormente, en un espacio vectorial hay definidas dos operaciones: la suma (interna) y el producto a escalar (externa). Las representaciones en lenguaje matemático de ambas operaciones se muestran en la figura 1.11. La figura muestra, mediante notación matemática, que podemos sumar dos vectores y se obtiene otro vector, al igual que también podemos multiplicar un vector por un número (escalar) y se obtiene otro vector.

$\begin{array}{l} \text{Suma} \quad + : V \times V \longrightarrow V \\ \quad \quad \quad (u, v) \quad \mapsto \quad u + v \\ \\ \text{Producto} \quad \cdot : K \times V \longrightarrow V \\ \quad \quad \quad (a, u) \quad \mapsto \quad a \cdot u \end{array}$

Figura 1.11. Representación matemática de las operaciones definidas en un espacio vectorial.

Dentro de un espacio vectorial V , la *suma* es una operación interna que cumple las siguientes propiedades: conmutativa, asociativa, elemento neutro y elemento opuesto. La expresión matemática de todas estas propiedades es la siguiente:

- **Propiedad Conmutativa:** $u + v = v + u, \forall u, v \in V$
- **Propiedad Asociativa:** $u + (v + w) = (u + v) + w, \forall u, v, w \in V$
- **Elemento Neutro:** $\exists 0 \in V : u + 0 = u, \forall u \in V$
- **Elemento Opuesto:** $\forall u \in V, \exists -u \in V : u + (-u) = 0$

Además, dentro de un espacio vectorial V , el *producto por un escalar* es una operación externa que cumple las siguientes propiedades: asociativa, elemento neutro, distributiva para suma de vectores y distributiva para suma de escalares. La expresión matemática de todas estas propiedades es la siguiente:

- **Propiedad Asociativa:** $a.(b.u) = (a.b).u, \forall a, b \in K, \forall u \in V$
- **Elemento Neutro:** $\exists 1 \in K : 1.u = u, \forall u \in V$
- **Distributiva para suma de vectores:** $a.(u+v) = a.u+a.v, \forall a \in K, \forall u, v \in V$
- **Distributiva para producto de vectores:** $(a+b).u = a.u+b.u, \forall a, b \in K, \forall u \in V$

Cualquier conjunto que posea unas operaciones *suma* y *producto por escalares*, cumpliendo todas las propiedades anteriores, diremos que es un espacio vectorial, y los elementos de tal conjunto se llamarán vectores. Además, diremos que el espacio vectorial es real o complejo, según sean los escalares.

1.4.2 Sistema de referencia en el espacio

Una vez definido el concepto de espacio vectorial y sus elementos, se puede construir, a partir de los vectores, un sistema de referencia que permita expresar los puntos del espacio ordinario (ver figura 1.12).

Un sistema de referencia en el espacio, denotado por R , consiste en un conjunto de tres vectores y un punto (Apostol, 2006). El conjunto de tres vectores formará una base y el punto fijo será el origen común de los vectores. A este punto, llamado origen, se le nombra con la letra O . A los vectores de la base se les representa como $B = \left\{ \vec{i}, \vec{j}, \vec{k} \right\}$.

Normalmente la base utilizada es siempre ortonormal. Teniendo en cuenta lo anterior, se puede nombrar al sistema de referencia en el espacio R^3 como $R = \left\{ O, \left(\vec{i}, \vec{j}, \vec{k} \right) \right\}$.

A cada punto P del espacio ordinario le corresponde un vector de origen O y extremo P , que se denotará por $\left(\vec{OP}\right)$ y que tiene unas coordenadas (a,b,c) en la base $B = \left\{\vec{i}, \vec{j}, \vec{k}\right\}$ del sistema de referencia dado. Se dice que (a,b,c) son las coordenadas del punto P en la referencia R . A cada una de las ternas le corresponde un único punto.

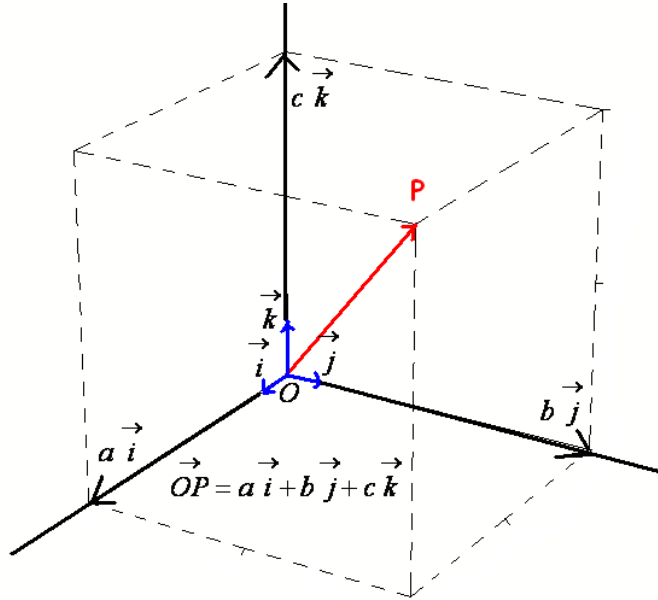


Figura 1.12. Sistema de referencia en el espacio

1.4.3 Sub-espacios vectoriales

Dado un espacio vectorial V , se puede considerar una parte S de él que funcione como un espacio vectorial menor incluido en V . Como V es un espacio vectorial, posee unas operaciones (suma, producto por un escalar) que en particular se pueden efectuar en S . Sólo se requiere que, al efectuarlas, su resultado quede dentro de S .

Si V es un espacio vectorial sobre K y $S \subset V$, siendo V no vacío, entonces S es un sub-espacio vectorial de V si se cumplen las siguientes dos propiedades:

- **Propiedad 1:** $\forall u, v \in S, u + v \in S$
- **Propiedad 2:** $\forall u \in S, \forall k \in K, k \cdot u \in S$

Además, S heredaría las operaciones de V como aplicaciones bien definidas. Es decir, que no escaparían de S . Otra consecuencia añadida de la definición anterior es que S es un espacio vectorial sobre K (Strang, 2009).

Por otro lado, con cualquier subconjunto de elementos seleccionados sobre un espacio vectorial no vacío se pueden generar sub-espacios vectoriales (Apostol, 2006). Las principales formas en que se puede presentar un sub-espacio vectorial son: sistema de generadores, base, ecuaciones paramétricas y ecuaciones cartesianas o implícitas. La figura 1.13 resume las formas de expresión de un sub-espacio vectorial y muestra un esquema que indica las posibilidades de transformación de unas a otras, según pueda interesar.

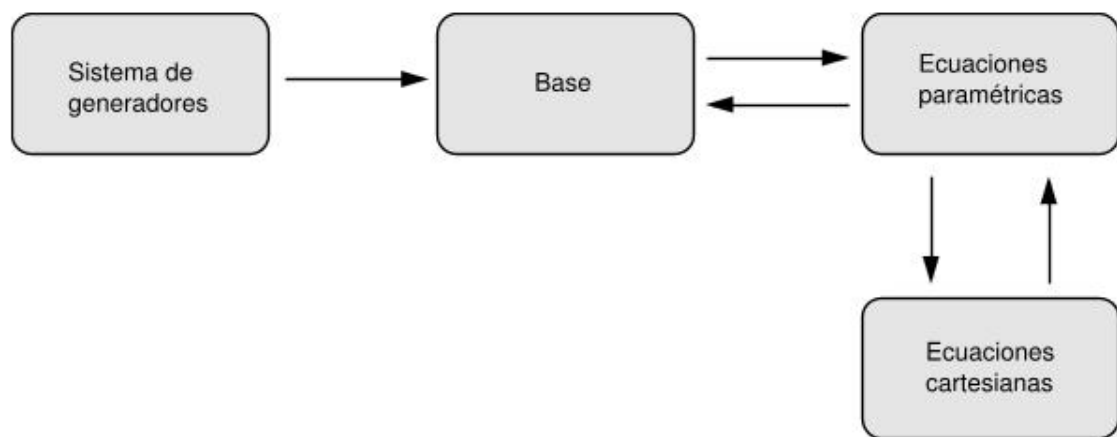


Figura 1.13. Expresiones de un sub-espacio vectorial

En la metodología propuesta en la presente tesis doctoral se empleará la expresión de un sub-espacio vectorial como sistema de generadores, debido a que es la opción que mejor se ajusta a las necesidades de implementación del sistema. A continuación se explica en detalle este concepto.

1.4.4 Sistemas de generadores

En álgebra lineal, dado un espacio vectorial V , se llama sistema de generadores al conjunto de vectores, pertenecientes a V , a partir del cual se puede generar el espacio vectorial V completo (Apostol, 2006).

Es importante no confundir este concepto con el de base, ya que si bien toda base es un sistema generador, la implicación inversa no es cierta. Mientras que una base ha de ser obligatoriamente un sistema libre (todos sus elementos han de ser linealmente independientes) un sistema generador puede ser ligado. Es decir, linealmente dependiente.

Hay que destacar que para cualquier sistema generador W formado por n elementos, siempre se puede hallar una base B comprendida en W con un número de elementos estrictamente menor que n (de ser igual se obtendría la base en sí y no se hablaría de sistema generador).

En general, a la hora de representar un sistema de generadores se emplea la siguiente notación: $V = W(\vec{V}_1, \vec{V}_2, \dots, \vec{V}_n)$. Donde V es el espacio vectorial generado por el sistema W compuesto por n vectores. Siendo n mayor o igual a la dimensión del espacio V . La figura 1.14 muestra una visión espacial abstracta de un sistema de generadores.

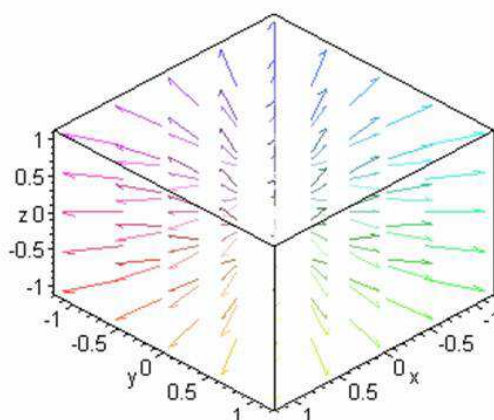


Figura 1.14. Representación espacial de un sistema de generadores

1.4.5 Espacios vectoriales en minería de opiniones

A lo largo de la historia han aparecido multitud de trabajos que se han llevado a cabo en minería de opiniones utilizando técnicas basadas en vectores (Cambria et al., 2013). En concreto, este tipo de soluciones se basan en el uso de máquinas de soporte vectorial (*Support Vector Machines*, abreviado *SVM*) y son un conjunto de algoritmos de aprendizaje supervisado desarrollados por Vladimir Vapnik y su equipo en los laboratorios AT&T²².

Los métodos basados en *SVM* están propiamente relacionados con problemas de clasificación y regresión (Vapnik, 1992). Dado un conjunto de ejemplos de entrenamiento, podemos etiquetar las clases y entrenar una *SVM* para construir un modelo que prediga la clase de una nueva muestra. Una *SVM* es un modelo que representa a los puntos de muestra en el espacio separando las clases por un espacio lo más amplio posible. Cuando las nuevas muestras se ponen en correspondencia con

²² <http://www.corp.att.com/attlabs/>.

dicho modelo en función de su proximidad pueden ser clasificadas en una u otra clase. Una *SVM* construye un hiperplano o conjunto de hiperplanos en un espacio de dimensionalidad muy alta (o incluso infinita) que puede ser utilizado en problemas de clasificación o regresión. Una buena separación entre las clases permitirá una clasificación correcta.

Dado un conjunto de puntos, subconjunto de un conjunto mayor (espacio), en el que cada uno de ellos pertenece a una de dos posibles categorías, un algoritmo basado en *SVM* construye un modelo capaz de predecir si un punto nuevo (cuya categoría desconocemos) pertenece a una categoría o a la otra. Como en la mayoría de los métodos de clasificación supervisada los datos de entrada (los puntos) son vistos como un vector p -dimensional. La *SVM* busca un hiperplano que separe de forma óptima a los puntos de una clase de los de la de otra que, eventualmente, han podido ser previamente proyectados a un espacio de dimensionalidad superior. En ese concepto de “separación óptima” es donde reside la característica fundamental de las *SVM*: este tipo de algoritmos busca el hiperplano que tenga la máxima distancia (margen) con los puntos que estén más cerca de él mismo. Por eso también, a veces, se les conoce a las *SVM* como clasificadores de margen máximo. De esta manera los puntos del vector que son etiquetados con una categoría estarán a un lado del hiperplano y los casos que se encuentren en la otra categoría estarán al otro lado (Bautista, 2004).

Los algoritmos *SVM* pertenecen a la familia de los clasificadores lineales. En la literatura de las *SVM*, se llama atributo a la variable predictora y característica a un atributo transformado que es usado para definir el hiperplano. La elección de la representación más adecuada del universo estudiado se realiza mediante un proceso denominado “selección de características”. Al vector formado por los puntos más cercanos al hiperplano se le llama *vector de soporte* (Vapnik, 1992).

Uno de los trabajos más relevantes aparecidos en los últimos años en minería de opiniones utilizando mecanismos de *SVN* es el trabajo de M. Rushdi-Saleh (2011). En este trabajo, se lleva a cabo la minería de opiniones utilizando máquinas de soporte vectorial (*SVM*). El objetivo es probar diferentes conjuntos de datos de distintos dominios y usando varios esquemas de ponderación. Los resultados de este trabajo demostraron la viabilidad de la *SVM* para llevar a cabo la minería de opiniones en diferentes dominios.

Recientemente ha aparecido otro trabajo en minería de opiniones que utiliza *SVM* (Smailovic et al., 2014). En este trabajo se estudian las relaciones existentes entre las

opiniones vertidas por los usuarios en *Twitter* sobre compañías y sus productos, y los cambios de los precios en bolsa de estos productos. Para ello se realiza un análisis estático de las opiniones escritas por los usuarios en *Twitter* con el objetivo de establecer el mejor conjunto de datos de entrenamiento para el clasificador de sentimientos basado en *SVM*. Los resultados de este trabajo determinaron que los cambios de opinión de los usuarios con respecto a una compañía específica podrían ser utilizados como indicadores de los cambios en los precios de cierre en bolsa para esa compañía.

Por otro lado, los modelos basados en *SVM* están estrechamente relacionados con las redes neuronales. Usando una función kernel, resultan un método de entrenamiento alternativo para clasificadores polinomiales, funciones de base radial y perceptrón multicapa. Sin embargo, las *SVM* han sido ampliamente utilizadas con éxito como método de clasificación de sentimientos, mientras que las redes neuronales artificiales (*Artificial Neural Networks*, abreviado *ANN*) escasa vez se han considerado en los estudios comparativos en la literatura del análisis de sentimientos. En (Moraes, 2013) se presenta una comparación empírica entre *SVM* y *ANN* con respecto al análisis de sentimientos a nivel de documento. Los experimentos llevados a cabo en este trabajo han indicado que *ANN* producía mejores resultados que *SVN* en el proceso de clasificación de sentimientos a nivel de documento. Además, los resultados de los experimentos también confirmaron algunas limitaciones potenciales de ambos modelos (el coste computacional de *SVN* en tiempo de ejecución y de *ANN* en tiempo de entrenamiento) que han sido escasamente discutidos en la literatura sobre la clasificación de sentimientos.

La metodología que se propone en esta tesis doctoral no es utilizar algoritmos de *SVM* para llevar a cabo la minería de opiniones de usuarios. De hecho, la solución que se aporta no está basada en aprendizaje supervisado, sino que es totalmente automática. La idea es asociar un vector euclídeo a cada una de las características identificadas en la opinión de cada usuario. Este vector tendrá tres coordenadas asociadas en un sistema de referencia en el espacio R^3 , que se corresponderán con los valores de polaridad positiva, negativa y neutra asociados a dicha característica en la opinión del usuario. Posteriormente, gracias a las operaciones con vectores disponibles en el sub-espacio vectorial formado por los vectores de polaridad asociados a cada una de las características identificadas en la opinión del usuario, y a otras operaciones de geometría espacial avanzada, se calculará la polaridad de la característica dentro de la opinión del usuario. Por último, la polaridad individual de cada característica

identificada en la opinión del usuario determinará la polaridad global de dicha opinión haciendo uso un vez más de cálculos con vectores en el espacio.

1.5 Minería de opiniones

1.5.1 Introducción

La minería de opiniones permite valorar cuantitativamente expresiones subjetivas como sentimientos o sensaciones (Chen & Zimbra, 2010; Pang et al., 2008). Permitiendo conocer si se está hablando positivamente, negativamente o de forma neutra sobre productos o servicios. Así como medir la intensidad de dicha opinión. Debido a la importancia actual de foros y redes sociales para algunos sectores de actividad, la posibilidad de disponer de herramientas que permitan monitorizar y valorar las opiniones vertidas en estos foros y redes sociales es fundamental (Rushdi-Saleh et al., 2011).

Cuando se habla de minería de opiniones, también llamada análisis del sentimiento, se refiere a la aplicación de una serie de técnicas del procesamiento del lenguaje natural, lingüística computacional y minería de textos, que tienen como objetivo la extracción de información subjetiva a partir de contenidos generados por los usuarios, como comentarios en blogs u opiniones en revistas de productos (Cruz et al., 2013; Moreo et al., 2012; Rushdi-Saleh et al., 2011). Con este tipo de tecnologías es posible extraer un valor tangible y directo a partir de un comentario en lenguaje natural.

La minería de opinión tiene distintos campos de aplicación (Qiong Wu et al., 2011). Por ejemplo, utilizando esta técnica se podrían monitorizar y gestionar las opiniones de los usuarios acerca de productos y servicios en redes sociales y foros concretos. Se podría intentar identificar qué se opina sobre un producto o servicio, o cuál es la opinión sobre determinadas características del producto o servicio concreto (Min & Park, 2012; Hye-Jin et al., 2012). También se podrían monitorizar y gestionar las opiniones sobre personas o entidades en foros y redes sociales concretas. Así como abordar temas como las tendencias y *hot topics*. Esto se traduce en unos datos subjetivos con gran potencial. De hecho pueden ser explotados por administraciones públicas, empresas y particulares para, por ejemplo, conocer la opinión sobre personalidades públicas, elegir la propaganda idónea según las preferencias u opiniones de la gente o para encontrar el producto mejor valorado por los usuarios (Liu, 2007).

Además la minería de opiniones puede cubrir diversas necesidades, por lo que, dependiendo de cuál sea la necesidad concreta, es posible adaptar la tecnología para cubrir dicha necesidad. Como se apunta en (Balahur & Montoyo, 2010), un proyecto de minería de opinión requiere, en primer lugar, que se determinen las páginas Web de las que se van a extraer las opiniones de los usuarios. A continuación se establece la periodicidad con la que extraer dichas opiniones o comentarios. Por último se procesan las opiniones o comentarios obtenidos para aplicar las métricas que permiten visualizar y gestionar de forma adecuada los resultados obtenidos.

Teniendo en cuenta las tendencias actuales del *eCommerce*, y en particular la socialización del comercio electrónico junto con la ingente cantidad de datos generados por los usuarios disponibles en la red (y creciendo con la penetración de Internet en los dispositivos móviles) la minería de opiniones resulta cada vez más útil para exprimir al máximo toda esta información. Teniendo como objetivo que las empresas entiendan a sus clientes y les ofrezcan lo que necesitan. Por ejemplo, una empresa en cuya tienda online se permita que los clientes realicen comentarios u opiniones de los productos que oferta, podrá analizar de forma activa lo que sus clientes tienen que decir de los productos que les venden. Sin embargo, si el volumen de comentarios u opiniones sobre los productos es muy elevado es posible que no se tenga la capacidad de procesar de forma manual todos esos contenidos. Poder disponer de herramientas que ayuden a procesar de forma automática todas esas opiniones para centrarse en los grupos de opinión que interesen (posiblemente las opiniones negativas) permitirá ahorrar mucho tiempo y dinero. Dando como resultado una mejor calidad de servicio para los clientes (Chen, 2012).

Gracias a esto la minería de opiniones está ganando terreno dentro de la actual Web Social. Además, un número muy elevado de investigadores de los campos de los sistemas de recuperación de información y la lingüística computacional han focalizado sus investigaciones en este campo. A continuación se revisaran los trabajos más relevantes en esta área a lo largo de la historia para comprender la evolución que ha experimentado la minería de opiniones. Posteriormente, en la sección 1.5.5 se seleccionaran los trabajos más relevantes que se han hecho en minería de opiniones en los últimos tiempos comparándolos con el trabajo propuesto en esta tesis doctoral.

1.5.2 Evolución histórica de la minería de opiniones

La minería de opiniones ha ido evolucionando a lo largo del tiempo. En esta sección se presenta un recorrido histórico a través de los años citando aquellos trabajos que en su momento supusieron un significativo avance en este campo.

En 1997, el trabajo de (Hatzivassiloglou & McKeown, 1997) se basó en la investigación de la orientación semántica, también llamada polaridad, de los adjetivos identificados en las opiniones de los usuarios.

En el inicio del siglo XXI aparecieron diversas soluciones en el campo de la minería de opiniones que utilizaron distintas técnicas para llevar a cabo su cometido. Estas técnicas se pueden clasificar en tres grandes grupos: medidas probabilísticas de asociaciones de palabras (Turney et al., 2003), técnicas que usan información sobre relaciones léxicas (Kamps et al., 2002; Kim et al., 2004), y técnicas que usan bases de datos léxicas (Esuli & Sebastiani, 2005; Andreevskaia et al., 2006). Todos estos trabajos clasifican las palabras en dos categorías (positiva o negativa) y proporcionan una puntuación global positiva o negativa para un texto. Si un documento contiene más términos positivos que negativos, se asume que es un documento positivo y viceversa. Estos mecanismos de clasificación fueron útiles como mecanismos de partida y mejoraron la eficiencia de la clasificación de sentimientos.

Poco después han aparecido estudios que no determinan solamente la polaridad de un texto (positivo o negativo) sino que también definen el nivel de polaridad del mismo como alto/medio/bajo positivo/negativo (Dave et al., 2003; Goldberg et al., 2006; Shimada et al., 2008).

En estos años, trabajos como los de (Pang et al., 2002; Pang & Lee, 2004; Mukras & Carroll, 2004; Gamon, 2005; Baccianella et al., 2009) explotaron los datos provenientes de opiniones sobre películas, servicios y productos. Haciendo uso de un completo conjunto de métodos estadísticos de selección de características y aplicando técnicas de aprendizaje computacional (*machine learning*). Las investigaciones llevadas a cabo por este conjunto de investigadores concluyeron que las técnicas de aprendizaje computacional, a nivel general, no alcanzan un rendimiento satisfactorio en la tarea de análisis de sentimientos.

En este contexto aparecen otros trabajos que abordan el análisis de sentimientos desde un punto de vista distinto al aprendizaje computacional. En (Zhou & Chaovalit, 2008; Zhao & Li, 2009) se utilizan técnicas de la Web Semántica, más concretamente las

ontologías, en el proceso de minería de opiniones basada en características. El principal objetivo de esos trabajos es calcular la polaridad teniendo en cuenta las características de un concepto. Los autores proponen primero la construcción de las ontologías del dominio para modelar el contenido de los textos del corpus para posteriormente extraer los adjetivos de las características. En (Zhou & Chaovalit, 2008) la polaridad de un texto se calcula sobre la base de los pesos de las características, y el valor de la polaridad de una característica se estima por el enfoque de “*estimación de máxima probabilidad*”. En (Zhao & Li, 2009), los autores obtienen la orientación semántica de un texto en base a la jerarquía de una ontología del dominio. El método propuesto también es capaz de obtener la orientación positiva, negativa o neutra del texto.

Sin embargo, hasta este momento los sistemas de clasificación de sentimientos se habían enfocado típicamente en un dominio específico. Su rendimiento disminuía considerablemente cuando se aplicaban en un dominio diferente. Además, la construcción de este tipo de sistemas requiere anotar una gran cantidad de datos para cada dominio, y este proceso necesita mucha mano de obra humana. Por lo tanto, una forma razonable para dar solución a este asunto sería utilizar los datos anotados en un dominio existente para la clasificación de sentimientos en un dominio de destino. Para abordar este problema en (Qiong Wu et al., 2011) se propone una plataforma basada en dos etapas para clasificación de sentimientos a través de varios dominios. La idea es construir un puente entre el dominio de origen y el dominio de destino para obtener documentos etiquetados en el dominio de destino. En la siguiente etapa se explota la estructura intrínseca, dada por estos documentos etiquetados, para etiquetar los datos en el dominio de destino. Los resultados experimentales indican que este enfoque podría mejorar el rendimiento de la clasificación de sentimientos entre dominios en un porcentaje alto.

Un interesante estudio descrito en (Hye-Jin et al., 2012) trabaja sobre las numerosas opiniones que realizan los usuarios en la compraventa de productos online. La idea de este trabajo es prestar más atención a la evaluación de la calidad de dichas opiniones. En este artículo se propone un nuevo indicador para clasificar opiniones por “*menciones sobre las experiencias*”, lo cual representa las experiencias personales de los clientes como una forma de identificar opiniones de alta calidad. La métrica propuesta en este trabajo está basada en la captura de dos parámetros, a saber: (i) las expresiones de tiempo relacionadas con el uso de los productos, y (ii) las entidades de los productos en diferentes períodos de compra. Los resultados empíricos muestran que esta medida no sólo es tan útil como los mejores indicadores existentes, ‘*voto útil*’ o ‘*rango crítico*’, sino que también está libre de sesgos indeseables que, o bien penalizan lo reciente o están

basados en la popularidad. En este estudio también se muestra que ordenar las opiniones con la métrica propuesta se considera provechoso desde el punto de vista de la satisfacción de los clientes.

En (Moreo et al., 2012) se propone un sistema de clasificación de sentimientos donde se abordan los siguientes aspectos: (i) la tendencia de muchos usuarios a expresar sus puntos de vista en un lenguaje no estándar, (ii) la detección del objetivo de las opiniones de los usuarios en un escenario de múltiples dominios, y (iii) el diseño de un modelo de conocimiento lingüístico modular con capacidad de adaptación de manera sencilla. El sistema propuesto consiste en un módulo automático de detección del enfoque de la opinión y un módulo de análisis de sentimientos capaz de evaluar las opiniones de los usuarios en las noticias sobre dicho enfoque o temas. Estos módulos utilizan un léxico diseñado específicamente para el análisis de noticias. Los experimentos muestran que los resultados obtenidos en este trabajo son muy prometedores.

En los últimos años se han impuesto dos métodos principales para clasificar las opiniones de los usuarios a nivel de características: los métodos basados en modelos y los métodos estadísticos. Sin embargo, existen pocos estudios comparativos entre estos métodos en lo que respecta a sus habilidades prácticas en la identificación de los distintos elementos dentro de una opinión, tales como las características, intensificadores, frases de entidad y entidades poco frecuentes. En (Chen & Qi, 2012), se lleva a cabo una comparación experimental de estos métodos basada en datos del mundo real y, además, se adopta un modelo condicional de campos aleatorios y se evalúa su rendimiento en comparación con algoritmos relacionados. En definitiva, en este trabajo se lleva a cabo una comparación exhaustiva de los métodos de aprendizaje a nivel de características para la minería de opiniones de usuarios en la red.

Muchos investigadores han propuesto diversos sistemas de extracción de opinión, la mayoría independientes del dominio, para extraer automáticamente las representaciones estructuradas de las opiniones contenidas en dichos textos. Sin embargo, en (Cruz et al., 2013) se aborda esta tarea mediante un enfoque orientado al dominio, es decir, se definen un conjunto de recursos específicos de dominio que capturan conocimiento valioso acerca de cómo las personas expresan opiniones sobre un determinado dominio. Estos recursos se crean de forma automática a partir de un conjunto de documentos con anotaciones. Los experimentos se llevaron a cabo en tres dominios diferentes (comentarios hechos por los usuarios sobre teléfonos, hoteles y coches), comparando este enfoque con otras técnicas independientes del dominio. Los resultados confirmaron

la importancia del dominio con el fin de construir sistemas precisos de extracción de opinión.

Las metodologías de clasificación de opiniones pueden ser supervisadas o no supervisadas. Una metodología supervisada utiliza algoritmos de aprendizaje automático cuando existen datos de entrenamiento, mientras que una metodología no supervisada se aplica cuando existen recursos lingüísticos disponibles. En (Martin-Valdivia et al., 2013) se combinan *SVM* con métodos de aprendizaje no supervisado. Es decir, en este trabajo se combinan los dos enfoques. Se propone el uso de meta-clasificadores que combinan aprendizaje supervisado y no supervisado con el fin de desarrollar un sistema de clasificación de sentimientos.

Recientemente, en (Kim et al., 2014) se desarrolla un sistema experto de ayuda a la decisión basado en minería de opiniones para predecir las subidas y bajadas de los índices bursátiles. Para ello se trabaja sobre un corpus formado por noticias en el dominio de las finanzas y la bolsa. En un primer experimento los investigadores construyeron un diccionario de sentimientos específico para el dominio de la bolsa mediante el procesamiento de artículos online sobre finanzas. En un segundo experimento los investigadores probaron la capacidad predictiva del sistema de ayuda a la decisión construido tomando como entrada noticias del dominio financiero provenientes de distintos medios. Los resultados de estos experimentos demostraron que aplicar minería de opiniones sobre noticias del dominio de las finanzas puede ser un mecanismo efectivo para predecir cambios en el mercado bursátil.

1.5.3 Minería de opiniones basada en características

La detección de la polaridad o, en definitiva, ser capaces de determinar si una opinión es positiva o negativa es la tarea que tradicionalmente se ha asociado a la minería de opiniones clásica. La minería de opiniones basada en características va un paso más allá. Esta parte de la minería de opiniones trata de determinar las distintas características relevantes tratadas en la opinión escrita por el usuario y, para cada una de esas características mencionadas en la opinión, tratar de extraer una polaridad (Hu & Liu, 2004; Jindal & Liu, 2006; Miao et al., 2008). Este tipo de procesos son mucho más complejos y de un grano mucho más fino que la detección de la polaridad global mencionada en el párrafo anterior, tal y como se apunta en (Ding et al., 2008).

El trabajo presentado en (Hu & Liu, 2004) describe una aproximación para descubrir las características presentes en opiniones sobre productos. Estas características se obtuvieron mediante el uso de minería de asociaciones entre palabras. En la actualidad se tiene claro que las características desempeñan un papel fundamental en la clasificación de sentimientos (Zhai et al., 2011), pero ¿cómo se pueden seleccionar eficazmente diferentes tipos de características para mejorar el rendimiento de la clasificación de sentimientos?. Este es el tema principal del trabajo de Zhai (2011) del que se extraen como características las siguientes unidades lingüísticas: palabras sobre sentimientos, sub-series, grupos de sub-cadenas y grupos de sub-cadenas clave. Esta forma de extracción de características nunca antes había sido considerada en el área de la clasificación de sentimientos. Una vez extraídas las características, éstas se comparan y analizan. Para poder realizar sus experimentos, en este trabajo se utilizan dos conjuntos de datos en diferentes dominios y en el idioma chino. Los resultados experimentales que se obtuvieron indicaron lo siguiente: (i) diferentes tipos de características poseen diferentes capacidades discriminativas en la clasificación de sentimientos para el lenguaje chino, (ii) las características *bigram character*²³ obtienen el mayor rendimiento entre las características *N_Gram*²⁴, (iii) las características de grupos de sub-cadenas mejoran el rendimiento de la clasificación de sentimientos mediante la combinación de sub-cadenas de diferentes longitudes, (iv) las palabras de sentimiento o frases extraídas de los léxicos de sentimiento no son eficaces para la clasificación de sentimientos en el idioma chino, y (v) las características eficaces para el proceso de análisis de sentimiento están presentes, por lo general, en diferentes longitudes en lugar de longitudes fijas.

Por lo tanto, en el trabajo presentado en (Zhai et al., 2011) queda demostrado que las características juegan un papel fundamental en la clasificación de sentimientos. Además se resalta la importancia utilizar métodos de selección de características efectivos de las distintas tipologías de características.

1.5.4 Recursos lingüísticos actuales en minería de opiniones

Debido al auge que ha ido tomando la minería de opiniones, en los últimos años se han ido construyendo multitud de recursos lingüísticos listos para poder utilizarse en las

²³ El término *bigram character* hace referencia en este trabajo a las características identificadas tomando cadenas de caracteres de longitud dos dentro de las opiniones.

²⁴ El término *N_Gram* hace referencia en este trabajo a los distintos tamaños de cadenas empleados para el proceso de identificación de características dentro de las opiniones.

soluciones ideadas en esta rama de la ciencia (Esuli, 2008). Estos recursos lingüísticos dotan de valor de sentimiento a las palabras identificadas en las opiniones de los usuarios (Cerini et al., 2007). A continuación se citan algunos ejemplos de los recursos lingüísticos más significativos aparecidos en los últimos años.

En el trabajo de (Miller, 1995) se presenta *WordNet*. Se trata de una base de datos léxica (también llamada lexicón) para el idioma inglés. Donde se agrupan las palabras en inglés en conjuntos de sinónimos llamados *synsets*, proporcionando definiciones cortas y generales, y almacenando las relaciones semánticas entre los conjuntos de sinónimos. El significado de los *synsets* queda más claro con la definición de *glosses*, que son definiciones y/o sentencias de ejemplo. Un ejemplo de *synset* con *gloss* es el que aparece en la tabla 1.2.

Tabla 1.2. Ejemplo de *synset* con *gloss* dentro de *WordNet*.

<i>good, right, ripe – (most suitable or right for a particular purpose; "a good time to plant tomatoes"; "the right time to act"; "the time is ripe for great sociological changes")</i>

En el año 2004 aparece *WordNet Affect* (Strapparava y Valitutti, 2004). Se trata de un recurso lingüístico para la representación léxica de conocimiento afectivo. Este recurso se construyó partiendo de *WordNet* a través de la selección y etiquetado de subconjuntos de *synsets* que representaban el conocimiento afectivo o emocional. Se partió de un conjunto inicial de palabras afectivas semilla, clasificadas de acuerdo con las seis categorías básicas de emociones (alegría, tristeza, miedo, sorpresa, ira y disgusto) y se expandió el léxico utilizando las relaciones de *WordNet*.

Otro método similar fue utilizado en la creación de *SentiWordNet* (Esuli y Sebastiani, 2005). En el trabajo de Esuli y Sebastiani (2005) se pone de manifiesto que la idea subyacente en la construcción de este recurso es que “*los términos con glosses similares en WordNet tienden a tener polaridad similar*”. Por lo tanto, *SentiWordNet* (SWN) fue construido usando un conjunto de palabras semilla cuya polaridad se conocía y se expandió usando similitud de *glosses*. SWN es una base de datos sentimental que proporciona valores de positividad y negatividad asociados a nombres, adjetivos, verbos y adverbios en el idioma inglés (Ohana & Tierney, 2009). En esta tesis doctoral, a la hora de evaluar el comportamiento de la metodología en un idioma específico, en este caso el idioma inglés, se recuperarán estos valores para todas las palabras que están

situadas cerca de la expresión lingüística que representa cada característica en la opinión de los usuarios.

En la tabla 1.3 se muestra la estructura de la base de datos sentimental *SWN 3.0*, que es la versión de esta base de datos utilizada en este trabajo. En esta tabla se muestran puntuaciones de ejemplo para algunos *synsets* (positivos y negativos) en *SWN 3.0*. La estructura de campos de *SWN* es la siguiente:

- **Etiqueta:** esta columna indica la categoría gramatical de la palabra. De esta manera, el contenido de esta columna será “a” para adjetivos, “n” para nombres, “v” para verbos y “r” para adverbios.
- **Posición:** esta columna indica la posición de la palabra dentro de la base de datos.
- **Positividad:** esta columna indica el grado de positividad de la palabra. Es decir, esta columna indica el valor de polaridad positiva asociado a la palabra. El 0 es el valor mínimo y el 1 es el valor máximo.
- **Negatividad:** esta columna indica el grado de negatividad de la palabra. Es decir, esta columna indica el valor de polaridad negativa asociado a la palabra. El 0 es el valor mínimo y el 1 es el valor máximo.
- **Synset:** esta columna indica el conjunto de palabras sujetas a los valores de positividad y negatividad concretos.
- **Descripción:** esta columna aporta una pequeña descripción que ayuda a definir el significado concreto de las palabras que aparecen en el *synset*.
- **Ejemplos:** esta columna aporta, opcionalmente, un conjunto de ejemplos que ayudan a clarificar el significado concreto de los *synsets*.

Cabe destacar que todas las puntuaciones tendrán uno de los siguientes valores: 0, 0.125, 0.25, 0.375, 0.5, 0.625, 0.75, 0.875 y 1. Esto es debido al efecto de la ponderación de todos los clasificadores de *SWN* por igual y el cálculo promedio de sus clasificaciones. En el Capítulo III de esta tesis doctoral explicaremos más en detalle el funcionamiento de *SWN* y la manera en la que ha sido construida esta base de datos de sentimientos para palabras en inglés.

Además, tal y como se puede apreciar en la tabla 1.3 puede haber un número indeterminado de entradas para una misma palabra en la base de datos o diccionario de sentimientos *SWN*. El número que aparece detrás de la palabra y el símbolo ‘#’ indica el número de la acepción concreta para esa palabra. Este es el caso, por ejemplo, de la palabra ‘good’, que aparece con tres acepciones diferentes en la tabla identificadas

como ‘good#1’, ‘good#2’ y ‘good#3’. Esto indica que la palabra ‘good’ puede tener diferentes valores de polaridad en función de la manera en la que se la invoque dentro de la opinión de los usuarios, ya que puede aparecer con diferentes significados e incluso con diferentes categorías gramaticales.

SWN es un recurso lingüístico para aplicaciones de minería de opiniones obtenido a partir de *WordNet*. El objetivo de *SWN* es proporcionar una extensión para *WordNet*, de tal manera que todos los *synsets* puedan estar asociados con un valor relativo a la connotación negativa, positiva o neutra. *SWN* etiqueta cada *synset* con un valor para cada categoría de entre 0.0 y 1.0. La suma de los tres valores es siempre 1.0, de modo que cada *synset* puede tener un valor distinto de cero para cada sentimiento, porque algunos *synsets* pueden ser positivos, negativos o neutros en función del contexto en el que se utilizan.

En (Kreutzer & Witte, 2014) se lleva a cabo un estudio exhaustivo sobre *WordNet* y *SentiWordNet*. Además, en este trabajo se muestra un ejemplo de aplicación software que utiliza *SWN* para llevar a cabo el proceso de minería de opiniones.

Otro lexicón que contiene palabras con su polaridad asociada es *MicroWNOp* (Cerini et al., 2007). Este recurso lingüístico se construyó sobre la base de un conjunto de términos extraídos del lexicón general *Inquirer* (Stone et al., 1966) y posteriormente añadiendo todos los *synsets* en *WordNet* donde aparecen estas palabras. La crítica hacia estos recursos es que no toman en cuenta el contexto en el que aparecen las palabras o expresiones. Otros métodos intentaron superar esta deficiencia y construyeron lexicones de sentimiento con el contexto local de las palabras (Lu et al., 2011).

Tabla 1.3. Puntuaciones de algunos synsets en SentiWordNet.

Etiqueta	Posición	Positividad	Negatividad	Synset	Descripción	Ejemplos
a	01123148	0.875	0	good#1	having desirable or positive qualities especially those suitable for a thing specified	"good news from the hospital"; "a good report card"; "when she was good she was very very good"; "a good knife is one good for cutting"; "this stump will make a good picnic table"; "a good check"; "a good joke"; "a good exterior paint"; "a good secretary"; "a good dress for the office"
a	01179547	0.125	0	godly#2 divine#1	emanating from God	"divine judgment"; "divine guidance"; "everything is black or white...satanic or godly"-Saturday Review
n	15046900	0	0	solid#1	matter that is solid at room temperature and pressure	
a	01125154	0.875	0	superb#2	surpassingly good	"a superb meal"
a	01049941	0	1	abject#2	most unfortunate or miserable	"the most abject slaves joined in the revolt"; "abject poverty"
a	01126841	0	1	sorry#2 sad#3 pitiful#2 lamentable#1 distressing#2 deplorable#1	bad; unfortunate	"my finances were in a deplorable state"; "a lamentable decision"; "her clothes were in sad shape"; "a sorry state of affairs"
n	05144079	0	0.875	badness#1 bad#1	that which is below standard or expectations as of ethics or decency	"take the bad with the good"
n	09630641	0	0.25	unfortunate_person#1 unfortunate#1	a person who suffers misfortune	
n	09637684	0	0.125	darky#1 darkie#1 darkey#1	(ethnic slur) offensive term for Black people	
n	09637013	0	0	black_american#1 afro-american#1 african_american#1 african-american#1	an American whose ancestors were born in Africa	
n	04849241	0.875	0	goodness#2 good#2	moral excellence or admirableness	"there is much good to be found in people"
a	01129977	0.875	0	good#3	morally admirable	

En la actualidad, los recursos léxicos más importantes disponibles para el idioma inglés (de manera libre) para poder ser usados en tareas de minería de opiniones son los siguientes:

- *WordNet Affect* (Strapparava y Valitutti, 2004).
- *SentiWordNet* (Esuli y Sebastiani, 2006).
- *Emotion triggers* (Balahur y Montoyo, 2008).
- *MicroWNOp* (Cerini et al., 2007).
- *General Inquirer* (Stone et al., 1966).
- *Appraisal terms* (Whitelaw et al., 2005).

En el trabajo que se presenta, la polaridad de cada una de las características identificadas dentro de la opinión de un usuario se calcula teniendo en cuenta el valor sentimental de las palabras cercanas a la expresión lingüística que identifica a la característica. Para ello se utiliza la base de datos sentimental SWN a la hora de proyectar la metodología sobre el idioma inglés, ya que actualmente no existen recursos lingüísticos con contenido sentimental de palabras en otro idioma que no sea el inglés. Este es uno de los motivos por el que se ha decidido validar la metodología propuesta en esta tesis doctoral (que es genérica, independiente del dominio e independiente del idioma) en el idioma inglés.

Además, en la implementación de la metodología propuesta en esta tesis doctoral también se utiliza *WordNet* para agrupar aquellas características identificadas que tienen expresiones lingüísticas diferentes pero que hacen referencia al mismo concepto. Esto es, se utiliza *WordNet* en el proceso de agrupación de características sinónimas. Es importante resaltar que dentro de la metodología propuesta el uso de una base de datos de sinónimos es opcional. Si la ontología del dominio que se utilice está bien construida cada concepto de la misma llevará asociado un conjunto de *labels* que contendrán todas las posibles expresiones lingüísticas asociadas a dicho concepto. Por tanto, la introducción de esta base de datos de sinónimos tiene como fin suplir la posible falta de completitud de términos lingüísticos asociados a cada concepto de la misma.

1.5.5 Comparativa de soluciones actuales en minería de opiniones

En los últimos años hemos asistido a la aparición de varios métodos para llevar a cabo el proceso de la minería de opiniones. La mayoría de las técnicas actuales de minería de

opiniones están diseñadas para trabajar con los lenguajes chino e inglés. En la tabla 1.4 se muestran algunos de los trabajos más recientes con respecto al análisis de sentimientos que han sido aplicados en diferentes dominios. Esta clasificación incluye los trabajos más relevantes en minería de opiniones desde el año 2011 hasta el año 2014. La heterogeneidad de este campo queda patente también en la tabla. De hecho, dado que esta disciplina está abierta a varias sub-disciplinas de investigación, el objetivo de estos trabajos es muy diferente y muestran resultados diferentes. Muchas veces estos resultados no son comparables entre sí, ya que son el resultado de objetivos distintos.

Teniendo en cuenta lo anterior, en este trabajo se han definido una serie de parámetros en torno a los cuales se pueden comparar los trabajos más recientes en el campo de la minería de opiniones de una manera justa. Los parámetros que se han identificado, y que son ampliamente analizados en las siguientes subsecciones, son los siguientes:

- **Recursos iniciales:** son los recursos que necesita inicialmente un enfoque para poder llevar a cabo su cometido.
- **Técnicas de clasificación de sentimientos:** son el conjunto de técnicas empleadas por un enfoque para llevar a cabo su objetivo de clasificar los sentimientos presentes en las opiniones de los usuarios.
- **Grado de automatización:** se trata de determinar qué enfoques requieren la intervención de un experto humano con el fin de obtener resultados, y cuáles no.
- **Identificación de características:** se trata de determinar qué enfoques procesan las opiniones de los usuarios en su conjunto, y cuáles tratan de identificar un conjunto de características en las opiniones de los usuarios y proporcionan un informe detallado sobre la polaridad de cada característica.
- **Dominios:** son los dominios en los que han sido validadas las metodologías y aproximaciones propuestas en cada uno de los enfoques estudiados.
- **Idiomas:** son los idiomas en los que han sido aplicadas las metodologías y aproximaciones propuestas en cada uno de los enfoques estudiados.
- **Nivel de clasificación de sentimientos:** se trata de determinar qué enfoques llevan a cabo la clasificación de sentimientos a nivel de palabra, a nivel de frase o a nivel de documento.

1.5.5.1 Recursos iniciales

Con respecto a los recursos iniciales necesarios para que un enfoque determinado lleve a cabo su objetivo, la mayoría de los enfoques hacen uso de léxicos de palabras que modelan sentimiento. Por ejemplo, los trabajos presentados en (Cruz et al., 2013; Martín-Valdivia et al., 2013; Singh et al., 2013) utilizan *SentiWordNet* (Esuli y Sebastiani, 2006), mientras que los presentados en (Min. & Park, 2012; Baccianella et al., 2010; Kontopoulos et al., 2013) utilizan otros lexicones de sentimiento menos populares.

Otras propuestas usan lexicones de sentimiento desarrollados por sus autores, como el trabajo presentado en (Eirinaki et al., 2012), que crea un nuevo lexicón sentimental formado por adjetivos. Por otro lado, en (Zhai et al., 2011), se utiliza un lexicón sentimental para el idioma chino. Los enfoques presentados en (Eirinaki et al., 2012) y en (Ghiassi et al., 2013) incorporan lexicones de sentimientos para *micro-blogging*, que es un servicio que permite a sus usuarios enviar y publicar mensajes breves, generalmente solo de texto. Por último, en (Chen et al., 2012) se presenta un estudio comparativo de distintos métodos de minería de opiniones basada en características aplicados de sobre un corpus de opiniones online, por tanto, en este trabajo los recursos externos necesarios para llevar a cabo su objetivo son los propios algoritmos utilizados para su comparativa.

Por el contrario, hay otros enfoques que no usan lexicones de sentimiento para llevar a cabo sus objetivos. En los trabajos de (Zhai et al., 2011; Rushdi-Saleh et al., 2011; Deng et al., 2014) se utilizan técnicas de *SVM* para el proceso de minería de opiniones, por lo tanto, requieren de un conjunto de documentos de entrenamiento inicial para alcanzar su objetivo.

En la definición de la metodología formal propuesta en esta tesis doctoral también se hace uso de un lexicón de palabras sentimentales. En concreto, en la implementación de la metodología dentro de una aplicación de software sobre un dominio e idioma concretos, utilizaremos *SentiWordNet* en el proceso de cálculo de polaridades sentimentales de aquellas palabras cercanas a las expresiones lingüísticas que representan a las características identificadas en las opiniones de los usuarios. Se ha elegido *SentiWordNet* como lexicón de palabras sentimentales porque la validación de la metodología se ha hecho utilizando un corpus de opiniones escritas en inglés.

1.5.5.2 Técnicas de clasificación de sentimientos

Las técnicas de clasificación de sentimientos se pueden dividir en técnicas de *machine-learning* y técnicas basadas en diccionarios. A pesar del hecho de que los enfoques de *machine-learning* han hecho avances significativos en la clasificación de sentimientos, para aplicarlos a opiniones de usuarios, en ocasiones, se requiere del uso de conjuntos de datos de entrenamiento etiquetados (aunque existen soluciones que no requieren este entrenamiento). Por otro lado, los enfoques basados en diccionarios pueden proporcionar ventajas significativas, tales como el hecho de que una vez que se han construido, no son necesarios datos de entrenamiento.

Actualmente, varios sistemas de minería de opiniones utilizan técnicas de *machine-learning* o *aprendizaje automático* para el proceso de clasificación de sentimientos. Más concretamente, la técnica de *Support Vector Machines* (SVM) se aplica ampliamente en una serie de obras (Zhai et al., 2011; Rushdi-Saleh et al., 2011; Ghiassi et al., 2013; Deng et al., 2014). En otros enfoques se combinan dos o más algoritmos de aprendizaje automático. Por ejemplo, en (Min & Park, 2012) se combinan SVM y reglas de árboles de decisión.

Algunas otras propuestas, como la de (Cruz et al, 2013), han sido diseñadas para ser totalmente flexibles en cuanto a la aplicación de técnicas de clasificación de sentimientos basadas en aprendizaje automático. Por tanto, en este tipo de trabajos se puede aplicar cualquier método de clasificación sentimental. En (Chen et al, 2012) se ha utilizado otra técnica de aprendizaje automático llamada *Conditional Random Fields* (CRF) para la clasificación de sentimientos.

Otros enfoques, tales como (Eirinaki et al., 2012; Singh et al., 2013), utilizan algoritmos basados en diccionario para llevar a cabo la clasificación de sentimientos. Por último, el sistema presentado en (Kontopoulos et al., 2013) utiliza un servicio *third-party* (o servicio de terceros) de análisis de sentimientos llamado *OpenDover*²⁵. Este servicio no es libre y su código fuente no está disponible, por lo que no están publicadas las técnicas de análisis de sentimientos empleadas por el mismo.

²⁵ Disponible en <http://www.opendover.nl/> . Última consulta realizada en Abril de 2015.

Tabla 1.4. Sistemas de minería de opiniones actuales.

Autores	Recursos externos	Técnicas de Clasificación de sentimientos	Grado de Automatización	Identificación de Características	Dominio	Lenguaje	Nivel de Clasificación de sentimientos
(Chen et al., 2012)	Algoritmos base de la comparativa	Conditional Random Fields (CRFs)	Semiautomático	Si	Productos	Inglés	Documento
(Cruz et al., 2013)	Palabras de sentimiento	Diferentes clasificadores de sentimientos	Semiautomático	Si	Productos (Headphones) Hoteles Coches	Inglés	Frase
(Deng et al., 2014)	Documentos de entrenamiento	SVM	Automático	No	Películas	Inglés	Documento
(Eirinaki et al., 2012)	Adjetivos de sentimiento	Diccionarios	Semiautomático	Si	Productos	Inglés	Documento
(Ghiassi et al., 2013)	Lexicon Micro-blogging, Doc. de entrenamiento	Redes neuronales y SVM	Automático	Si	Música	Inglés	Frase
(Kontopoulos et al., 2013)	Palabras de sentimiento	Servicio Third-party de análisis de sentimientos	Semiautomático	Si	Productos (Smartphones)	Inglés	Frase
(Martín-Valdivia et al., 2013)	Documentos de entrenamiento y Palabras de sentimiento	SVM y aproximaciones no supervisadas	Semiautomático	Si	Películas	Inglés Español	Documento
(Min & Park, 2012)	Palabras de sentimiento, lexicon Micro-blogging	Reglas, Árboles de decisión y SVM	Semiautomático	Si	Productos	Inglés	Documento
(Rushdi-Saleh et al., 2011)	Documentos de entrenamiento	SVM	Automático	Si	Productos (Cámaras digitales, Computadores) Películas	Inglés	Documento
(Singh et al., 2013)	Palabras de sentimiento	Diccionarios	Automático	Si	Películas	Inglés	Documento
(Zhai et al., 2011)	Palabras de sentimiento, Doc. De entrenamiento	SVM	Semiautomático	Si	Hoteles Productos	Chino	Documento
Nuestra solución	Palabras de sentimiento	Ontologías y Espacios Vectoriales	Automático	Si	Películas Finanzas	Inglés	Documento

Finalmente, la metodología descrita en esta tesis doctoral combina el uso de ontologías de dominio con una nueva técnica basada en cálculos con vectores en el espacio para mejorar el proceso de análisis de sentimientos. Los resultados de este trabajo son el fruto de un análisis detallado de las opiniones de los usuarios teniendo en cuenta dos factores: la opinión total del usuario y todas las características señaladas en la opinión. El método propuesto no está limitado al empleo de un tema específico dentro de una opinión, ya que es capaz de analizar y clasificar los sentimientos de cada tema discutido por los usuarios en su conjunto de opiniones.

1.5.5.3 Grado de automatización

Otro parámetro a tener en cuenta para clasificar los diversos enfoques de minería de opiniones es el grado de automatización. Como se muestra en la tabla 1.4, la mayoría de los trabajos de minería de opiniones, como los presentados en (Zhai et al., 2011; Min & Park, 2012; Chen et al., 2012; Cruz et al., 2013; Martín-Valdivia et al., 2013; Eirinaki et al., 2012; Kontopoulos et al., 2013), tienen un grado de automatización que se puede categorizar como semiautomático. Esto se debe principalmente al hecho de que estos enfoques requieren, por lo general, la intervención de un experto con el fin de obtener resultados. Por ejemplo, el sistema descrito en (Kontopoulos et al., 2013) requiere la creación de una ontología de dominio basada en los tweets de los usuarios. Para hacer frente a este reto, los autores recurrieron a un editor de ontologías semiautomático basado en datos. Otros métodos son automáticos, tales como los trabajos de (Ruchdi Saleh et al., 2011; Ghiassi et al., 2013; Singh et al., 2013; Deng et al., 2014).

La metodología propuesta en esta tesis doctoral se basa en el uso de una ontología del dominio y comprende la aplicación de un método automático para la clasificación de los sentimientos de las opiniones de los usuarios. Este enfoque no requiere la intervención humana pudiéndose considerar como totalmente automático.

1.5.5.4 Identificación de características

Muchas de las aportaciones en minería de opiniones se han centrado tradicionalmente en la construcción de aplicaciones de análisis de emociones que procesan las opiniones de los usuarios en su conjunto. Sin embargo, la última tendencia en este campo es

identificar un conjunto de características en las opiniones de los usuarios y proporcionar un informe detallado sobre la polaridad de cada característica.

Como se muestra en la tabla 1.4, casi todas las obras seleccionadas en minería de opiniones incorporan una etapa de identificación de características como parte del proceso de análisis de sentimientos (Zhai et al., 2011; Rushdi-Saleh et al., 2011; Min & Park, 2012; Chen et al., 2012; Cruz et al., 2013; Martín-Valdivia et al., 2013; Eirinaki et al., 2012; Kontopoulos et al., 2013; Ghiassi et al., 2013; Singh et al., 2013). En los últimos años, este enfoque ha demostrado con éxito su eficacia en el campo de la minería de opiniones.

El método propuesto en esta tesis doctoral lleva a cabo la identificación de características mediante el uso de ontologías del dominio. Se identifica y asocia cada característica con una de las clases de la ontología del dominio. Además, el proceso de selección de características se puede aplicar a cualquier dominio cambiando la ontología que modela dicho dominio.

1.5.5.5 Dominios

Tradicionalmente, las aportaciones en el campo de la minería de opiniones han prestado poca atención a los dominios en los que se aplican estas técnicas. Por tanto, las particularidades de cada dominio de aplicación no han sido investigadas hasta hace poco. En este sentido, en (Rushdi-Saleh et al., 2011) se demuestra empíricamente que la precisión de los métodos de análisis de sentimientos que proponen depende en gran medida del dominio de estudio. El dominio de las películas cinematográficas, por ejemplo, es difícil de trabajar ya que las opiniones vertidas sobre ellas a menudo contienen partes que comentan la trama de la misma. Esto desencadena en una modificación del valor medio de la orientación semántica de las opiniones. Por otra parte, el dominio de los productos comerciales, por lo general, da lugar a mejores resultados de análisis de sentimientos en términos de precisión.

Por otro lado, en (Qiong Wu et al., 2011) se demuestra la implicación directa entre el dominio de aplicación elegido y la efectividad de los métodos de clasificación sentimental de opiniones supervisados. Ya que muchas de las soluciones aportadas históricamente en minería de opiniones bajan su efectividad cuando son aplicadas en un dominio distinto del que fueron ideadas.

Con respecto a los trabajos relacionados que hemos seleccionado, la mayoría han sido validados en el dominio de los productos (Zhai et al., 2011; Rushdi-Saleh et al., 2011; Min & Park, 2012; Chen et al., 2012; Cruz et al., 2013; Eirinaki et al., 2012; Kontopoulos et al., 2013), el turismo (Zhai et al., 2011; Cruz et al., 2013) o las películas (Rushdi-Saleh et al., 2011; Martín-Valdivia et al., 2013; Singh et al., 2013; Deng et al., 2014). Por último, el trabajo presentado en (Ghiassi et al., 2013) se validó en el dominio de la música.

La solución de análisis de sentimientos propuesta en este trabajo es independiente del dominio. Puede trabajar en cualquier dominio mediante el uso de una ontología y un corpus con las opiniones de usuarios en ese mismo dominio. El enfoque propuesto en esta tesis se ha validado en el dominio de las películas cinematográficas y en el dominio de las finanzas, quedando como línea futura de investigación la aplicación de la metodología en nuevos dominios. Se ha terminado la fase de validación de la metodología aportando un estudio detallado de los resultados obtenidos en cada dominio y justificando las diferencias obtenidas en cada uno de ellos. Toda esta información se encuentra en el Capítulo IV de esta tesis doctoral.

1.5.5.6 Idiomas

La mayoría de soluciones de análisis de sentimientos han sido diseñadas para funcionar correctamente con opiniones de usuarios en inglés (Rushdi-Saleh et al., 2011; Min & Park, 2012; Chen et al., 2012; Cruz et al., 2013; Martín-Valdivia et al., 2013; Eirinaki et al., 2012; Kontopoulos et al., 2013; Ghiassi et al., 2013; Singh et al., 2013; Deng et al., 2014). Sin embargo, hay actualmente muchas investigaciones en curso cuyo objetivo es aplicar las técnicas de minería de opiniones para el lenguaje chino (Zhai et al., 2011) y el español (Martín-Valdivia et al., 2013).

En (Martín-Valdivia et al., 2013) se ha utilizado un corpus de opiniones en español sobre críticas de cine junto con su corpus paralelo traducido al inglés. En primer lugar, se generan dos modelos individuales que utilizan estos dos corpus y aplican algoritmos de aprendizaje automático. En segundo lugar, se integra *SentiWordNet* (SWN) en el corpus en inglés generando un nuevo modelo de supervisión. Por último, los tres sistemas se combinan usando un meta-clasificador que permite aplicar varios algoritmos de combinación como sistema de votación. Los resultados obtenidos con este método superan a los obtenidos mediante los sistemas que trabajan de manera individual y

demuestran que este enfoque podría ser considerado como una solución favorable para la clasificación de sentimientos cuando se trabaja con corpus paralelos.

La metodología de minería de opiniones propuesta en esta tesis doctoral es general e independiente del lenguaje empleado en las opiniones de los usuarios. Se puede adaptar a cada lenguaje específico utilizando una ontología de dominio y lexicón de sentimientos en la lengua de que se trate. El método ha sido validado con opiniones de usuarios escritas en inglés, por lo que se han tenido que elegir los recursos lingüísticos específicos de este idioma, esto es, *SentiWordNet* (Esuli y Sebastiani, 2006) como léxico de sentimientos y *WordNet* (Miller, 1995) como base de datos de palabras sinónimas. Esta elección ha venido determinada por la ausencia de recursos lingüísticos de calidad en otros idiomas distintos del inglés.

1.5.5.7 Nivel de clasificación de sentimientos

La clasificación de sentimientos se puede realizar a nivel de palabra, de frase o de documento. Pocos de los enfoques estudiados se basan en análisis de sentimientos a nivel de frase (Cruz et al., 2013; Kontopoulos et al., 2013; Ghiassi et al., 2013). Por ejemplo, en (Kontopoulos et al., 2013), los autores presentan un método de análisis de opiniones realizado a nivel de frases cortas idóneo para analizar las opiniones publicadas por los usuarios en *Twitter*. En esta plataforma de *micro-blogging* (o mensajería corta), los usuarios pueden enviar mensajes de texto, llamados tweets, limitados a 140 caracteres. Dado que estos mensajes están basados en un tema muy específico y los usuarios expresan sus opiniones sobre ese tema con frases cortas, se facilita el proceso de análisis de sentimientos.

La mayoría de los enfoques existentes están basados en la clasificación de sentimientos a nivel de documento (Zhai et al., 2011; Rushdi-Saleh et al., 2011; Min & Park, 2012; Chen et al., 2012; Martín-Valdivia et al., 2013; Eirinaki et al., 2012; Singh et al., 2013; Deng et al., 2014).

En el trabajo propuesto en esta tesis doctoral el análisis de sentimientos se lleva a cabo a nivel de documento. Cada opinión de usuario puede extenderse a un promedio de unas 100 líneas de texto, por lo que es común que los usuarios realicen comentarios relativos a varios temas diferentes dentro de una misma opinión. Este hecho hace que sea más necesario el análisis de sentimientos basado en características, ya que, por las características de los corpus utilizados en los experimentos, al no hablar sobre un único

asunto en cada opinión, se requiere analizar la polaridad de cada uno de los temas que aparecen en las opiniones.

1.5.5.8 Conclusiones de la comparativa

Los resultados de la investigación en el campo de la minería de opiniones han demostrado que el análisis de sentimientos es una ardua tarea que puede ser realizada de distintas maneras.

Para las aplicaciones finales, el análisis de sentimientos no es la primera ni la última tarea a realizar. En primer lugar, es necesario recuperar un conjunto de documentos relevantes en el dominio en que estemos trabajando. Posteriormente, tras realizar el proceso de análisis de sentimientos se deberán interpretar los resultados convenientemente, ya que el resultado del análisis de sentimiento puede contener mucha información redundante que puede distorsionar el resultado final.

Los sistemas implementados para la tarea de análisis de sentimientos se basan en reglas, léxicos de palabras que tienen asociadas una polaridad (positiva o negativa), métodos estadísticos o aprendizaje automático.

Llevar a cabo una comparativa exhaustiva entre los distintos trabajos de investigación en el campo de la minería de opiniones no resulta sencillo, ya que la tarea de análisis de sentimientos no está definida de una forma única en las distintas aproximaciones llevadas a cabo en la actualidad. En particular, no es posible en todos los casos comparar la eficacia y rendimiento de los distintos sistemas, ya que, por ejemplo, los textos sobre los que actúan pueden tener diferentes elementos anotados.

Uno de los grandes problemas detectados en el campo de la minería de opiniones es la dificultad para encontrar lexicones de palabras que expresen sentimientos para otros idiomas distintos al inglés. Este hecho, ha propiciado que nos hayamos visto “forzados” a validar la metodología genérica propuesta en esta tesis, que es automática e independiente del dominio y del idioma, utilizando documentos escritos en inglés. El camino alternativo, es decir, validar la metodología en un idioma distinto del inglés, pasaba por construirnos nosotros mismos tales recursos o usar algún lexicón disponible de menor calidad que *SWN*.

Por último, tras los resultados de nuestras investigaciones en el campo de la minería de opiniones, se ha podido constatar que la mayoría de sistemas implementados en este campo trabajan a nivel léxico, utilizando reglas, léxicos, métodos estadísticos o

aprendizaje automático. Sin embargo, son escasas las aproximaciones que toman en cuenta otros niveles de análisis, como el sintáctico o semántico. Por tanto, el objetivo del método propuesto en esta tesis doctoral es proponer una metodología alternativa para llevar a cabo el proceso de minería de opiniones basada en características que tome en cuenta el nivel semántico. Para ello, se utiliza una ontología del dominio que aporte un vocabulario común del dominio específico en el que se trabaje y, además, permita aprovechar toda su estructura y potencia semántica para realizar el proceso de clasificación de sentimientos de opiniones.

1.6 Motivación y objetivos

Analizando las metodologías disponibles de minería de opiniones, nos damos cuenta de que existen diversos problemas que impiden la implantación de éstas como soluciones plausibles. Algunas de ellas utilizan medidas probabilísticas de asociación de palabras, otras, técnicas que usan información sobre relaciones léxicas, y otras, técnicas que usan bases de datos léxicas. Todas ellas usan datos provenientes de las opiniones sobre automóviles, bancos, películas, hoteles, etc. Además, clasifican las palabras en dos categorías (positiva o negativa) y calculan la puntuación positiva y negativa total para el texto. Si un documento contiene más términos positivos que negativos, se asume que el documento es positivo y, en caso contrario, negativo. Estos mecanismos de clasificación son útiles y mejoran la eficiencia de la clasificación de sentimientos, pero no pueden determinar la polaridad de cada característica particular dentro de una opinión determinada. Otras aproximaciones usan un conjunto completo de métodos de selección de características estadísticos y aplican técnicas de aprendizaje computacional. Estos experimentos demuestran que las técnicas de aprendizaje computacional no obtienen resultados satisfactorios en la clasificación de sentimientos.

Por otra parte, tal y como se ha apuntado anteriormente, gran parte de los enfoques actuales en el campo de la minería de opiniones poseen una serie de inconvenientes derivados de la ausencia de relaciones semánticas en el proceso de búsqueda de características, o la falta del empleo de métodos matemáticos avanzados en el proceso de análisis de sentimientos en las opiniones de los individuos.

En nuestra investigación se ha considerado que para que una metodología de minería de opiniones posea aplicabilidad real debe, primeramente, identificar todos los conceptos que existan en el corpus y sus características asociadas para, después, poder calcular la polaridad de las mismas y llevar a cabo un proceso de análisis de sentimientos

completo, con el objetivo de adquirir todo el conocimiento que existe en el corpus de opiniones.

En base a las carencias detectadas en las metodologías y sistemas de minería de opiniones, el trabajo presentado en esta tesis se planteó con el fin de proporcionar una solución a estos problemas. El trabajo realizado en esta tesis doctoral persigue los siguientes objetivos:

- **Definición y formalización de una metodología de minería de opiniones basada en características para clasificar los sentimientos de las opiniones.**

Esta metodología incluirá:

- **Procesamiento del lenguaje natural** del corpus de opiniones de los usuarios, empleando para ello: tokenizadores, reconocimiento de frases, etiquetado *POS* y lematizadores.
 - **Identificación de características** usando una ontología del dominio. Este proceso incluirá: (i) búsqueda de características en el corpus de opiniones de usuarios, y (ii) ponderación de cada una de ellas según su posición relativa dentro de la opinión de cada usuario.
 - **Cálculo de la polaridad de las características** identificadas, teniendo en cuenta su ponderación y utilizando una base de datos con valores de sentimiento asociados a palabras.
 - **Análisis de sentimientos** de cada opinión de cada usuario, así como de las características presentes en ella, usando cálculos vectoriales en el espacio R^3 .
- **Diseño e implementación de una aplicación software para llevar a cabo el proceso de análisis de sentimientos de las opiniones de los usuarios.**

Este sistema se diseñará para cumplir con los requisitos que implica el entorno de minería de opiniones descrito anteriormente. Se emplearán para ello las tecnologías más novedosas y los recursos mejor valorados disponibles actualmente.

- **Validación de la aplicación de software en varios dominios.**

La metodología de validación se compone de los siguientes puntos:

- Definición de las métricas de validación empleadas: precisión, exhaustividad y medida-F.
- Construcción de una ontología en el dominio de validación.
- Evaluación del proceso de identificación de características.
- Evaluación del proceso de clasificación de sentimientos de características.
- Evaluación del proceso de clasificación de sentimientos de opiniones.
- Validación de la metodología de análisis de sentimientos de opiniones en el dominio de las películas cinematográficas.
- Validación de la metodología de análisis de sentimientos de opiniones en el dominio de las finanzas.

1.7 Resumen

En este capítulo se presenta un exhaustivo estudio del estado del arte de las principales áreas relacionadas con la investigación que se relata en esta tesis doctoral, a saber: Web Semántica y ontologías, procesamiento del lenguaje natural, espacios vectoriales y minería de opiniones.

En la primera sección se lleva a cabo un estudio del estado del arte en Web Semántica y ontologías. En primer lugar, se define el concepto de Web Semántica y se enumeran los fundamentos tecnológicos sobre los que se sustenta. Seguidamente se comentan los problemas actuales con los que se está encontrando la Web Semántica hasta poder llegar a su implantación definitiva, y acabamos esta sección enumerando y explicando los componentes de los que está formada la Web Semántica. Uno de estos componentes, el más importante en el trabajo presentado en esta tesis, es la ontología. De esta manera continuamos esta sección definiendo las ontologías y enumerando los elementos que las conforman. Seguidamente se aporta un estudio sobre los distintos lenguajes de los que disponemos actualmente para representar a las ontologías. Finaliza esta sección con un apartado en donde se muestran algunos de los trabajos más relevantes en minería de opiniones que utilizan las tecnologías de la Web Semántica y las ontologías. Además, se comenta la importancia de estas tecnologías en la solución aportada en esta tesis doctoral.

En la segunda sección se lleva a cabo un estudio del estado del arte en tareas de procesamiento del lenguaje natural (*PLN*). En primer lugar, se define el concepto de *PLN* desde sus distintos puntos de vista. A continuación, hablamos sobre las

características que posee la información expresada en lenguaje natural y, seguidamente, se enumeran los siete niveles en los que se puede dividir el lenguaje natural, a saber: fonético, léxico, morfológico, sintáctico, semántico, discurso y pragmático. Tras esto, aportamos un estudio sobre los tipos de recursos lingüísticos que se pueden utilizar actualmente en tareas de *PLN*. Finalizamos esta sección hablando sobre las tareas de *PLN* que se aplican usualmente en problemas de minería de opiniones, destacando el hecho de que se trata de dos sub-disciplinas íntimamente relacionadas.

En la tercera sección se hace un amplio repaso de los conceptos fundamentales de análisis vectorial y geometría espacial que se necesitan para implementar la metodología de análisis de sentimientos de opiniones. En concreto, se define el concepto de espacio vectorial y el conjunto de operaciones vectoriales que pueden llevarse a cabo dentro de él. A continuación, se define el concepto de sistema de referencia en el espacio, que será un concepto clave en la solución propuesta en esta tesis. Tras esto, se definen los conceptos de sub-espacio vectorial y sistema de generadores, ya que ambos conceptos jugarán un papel decisivo a la hora de calcular la polaridad de las características identificadas en las opiniones de los usuarios. Finalmente, la sección acaba enumerando los trabajos más significativos en minería de opiniones que utilizan vectores para llevar a cabo sus tareas, y se comenta la manera en la que hemos utilizado todos los conceptos descritos en esta sección dentro de nuestra solución.

En la cuarta sección empezamos definiendo el concepto de minería de opiniones. Tras esto se hace un recorrido histórico sobre los trabajos más relevantes a lo largo de los tiempos en esta sub-disciplina. Seguidamente introducimos el concepto de minería de opiniones basada en características, y, por último, realizamos una comparativa exhaustiva sobre los trabajos más significativos aparecidos en los últimos años en el campo de la minería de opiniones atendiendo a características como los recursos iniciales, las técnicas de clasificación de sentimientos empleadas, el grado de automatización, los procesos de identificación de características, los dominios estudiados, los idiomas utilizados y el nivel de clasificación de sentimientos utilizado.

Finalizamos el capítulo enumerando los objetivos a cumplir durante el desarrollo de esta tesis doctoral. El objetivo principal de esta tesis es proporcionar una metodología innovadora para llevar a cabo la minería de opiniones basada en características a partir de textos en lenguaje natural. Esta metodología estará basada en la identificación de características utilizando una ontología del dominio. De esta manera, será posible llevar a cabo la clasificación de sentimientos de las opiniones de los usuarios utilizando para

ello cálculos vectoriales en el espacio R^3 . Para ello se desarrollará una aplicación de software que realice el trabajo de forma automática cumpliendo con los requisitos de la metodología definida y que permita, además, validarla de manera fácil y rápida en distintos dominios.

Capítulo II

Sistema basado en tecnologías de la
Web Semántica para la minería de
opiniones basada en características

2.1 Introducción

El uso de la minería de opiniones aporta diversos beneficios. Uno muy destacable es la comprensión de las opiniones vertidas en grandes conjuntos de información escrita sin necesidad de intervención humana (Ruchdi Saleh et al., 2011; Ghiassi et al., 2013; Singh et al., 2013; Deng et al., 2014). Esto supone un ahorro de tiempo en el tedioso trabajo de analizar textos uno a uno. Además, el analista de textos puede concentrarse en los documentos relevantes o en fragmentos que expresan opinión y utilizarlos en sus informes. La minería de opiniones nos permite hacer una distinción entre opiniones basadas en datos objetivos y aquellas que están cargadas de subjetividad (Balahur & Montoyo, 2010). Por otro lado, en (Zhai et al., 2011) queda demostrado que las características juegan un papel fundamental en el proceso de minería de opiniones, además de resaltar la importancia de utilizar métodos de selección de características efectivos en las opiniones.

También se ha visto que la aplicación de la Web Semántica, y más concretamente las ontologías, en el proceso de minería de opiniones basada en características produce buenos resultados, tal y como se refleja en los trabajos de (Zhou & Chaovalit, 2008) y (Zhao & Li, 2009).

Por todo lo anterior, este capítulo se centra en describir la metodología para llevar a cabo el análisis de sentimientos de opiniones que combine los beneficios de la minería de opiniones basada en características junto con los beneficios derivados del uso de ontologías. Para cumplir con este reto, en este capítulo se presenta la arquitectura general del sistema. Esta arquitectura pretende cumplir el principal objetivo de esta tesis doctoral: desarrollar una nueva metodología para la minería de opiniones basada en características guiada por ontologías. Esta metodología es independiente del dominio e idioma en los que se aplique. Sin embargo, la ausencia de recursos lingüísticos de alta calidad para idiomas distintos del inglés propiciará que, para la implementación y validación de la misma, utilicemos un corpus de opiniones y una base de datos léxica de sentimientos en inglés.

La arquitectura se compone de cuatro módulos principales: (i) módulo de procesamiento del lenguaje natural, (ii) módulo de identificación de características, (iii) módulo de cálculo de polaridades de características y (iv) módulo de análisis de sentimientos.

También se muestra un diagrama general de procesos donde se explica en detalle el funcionamiento de cada tarea llevada a cabo dentro de cada módulo de la arquitectura de minería de opiniones propuesta. Por último se muestra un ejemplo práctico de cómo funciona la aplicación de nuestra metodología en el dominio de las películas cinematográficas con opiniones escritas en inglés.

2.2 *Objetivos del sistema*

Como se ha comentado en la sección anterior, el objetivo principal de esta tesis doctoral es el desarrollo de una nueva metodología para llevar a cabo la minería de opiniones basada en características guiada por ontologías. Este objetivo general se puede descomponer en los siguientes sub-objetivos:

- **Desarrollo de métodos para la identificación de características a partir de textos en lenguaje natural.** Estos métodos de identificación de características utilizan la semántica de las ontologías para llevar a cabo su cometido. Cualquier expresión lingüística presente en la opinión de un usuario que tenga relación con una clase o instancia de la ontología es identificada como una nueva característica dentro de esa opinión.
- **Desarrollo de métodos para el cálculo de la polaridad de las características identificadas en las opiniones.** Estos métodos de cálculo de la polaridad asociada a una característica utilizan una base de datos léxica de sentimientos para llevar a cabo su cometido.
- **Desarrollo de métodos para el análisis de sentimientos de opiniones basados en cálculos vectoriales en el espacio R^3 .** Estos métodos de análisis de sentimientos de opiniones utilizan la teoría de geometría espacial y espacios vectoriales para llevar a cabo los cálculos necesarios para clasificar los sentimientos de las opiniones.

Se han tenido en cuenta, a su vez, una serie de objetivos secundarios con los que se pretende conseguir que el sistema que implementa la metodología propuesta en esta tesis doctoral satisfaga los estándares de calidad industriales, de forma que sea posible su aplicación en entornos reales. Los objetivos de diseño secundarios son los siguientes:

- **Independencia del dominio.** Se persigue el diseño de un entorno de trabajo que permita realizar el proceso de análisis de sentimientos de opiniones expresadas en lenguaje natural que sea independiente del dominio (por ejemplo: productos, geografía, medicina, finanzas, turismo, etc.).

- **Independencia del idioma.** Se persigue el diseño de un entorno de trabajo que permita llevar a cabo el proceso de análisis de sentimientos de opiniones que sea independiente del idioma. En este punto se requiere la existencia de recursos lingüísticos de calidad en el idioma con el que se quiera trabajar o la creación de los mismos.
- **Adaptabilidad y parametrización.** Se requiere el diseño de un sistema capaz de adaptarse a los cambios en el entorno de forma dinámica. El sistema deberá ofrecer la posibilidad de configurar parámetros que permitan ejecutar el proceso de minería de opiniones en distintas condiciones.
- **Eficiencia y flexibilidad.** El sistema deberá incorporar tecnologías de eficiencia máxima a la hora de llevar a cabo su cometido. Los recursos externos del sistema (librerías, utilidades, etc.) podrán intercambiarse de forma flexible, de modo que sea posible disponer en todo momento de las últimas versiones de estos.

2.3 *Arquitectura del Sistema*

En este apartado se presenta la arquitectura general del sistema que sirve como base de la metodología de análisis de sentimientos propuesta en esta tesis doctoral. En primer lugar se expondrá la estructura de capas de aplicación de esta arquitectura y a continuación se mostrará el esquema general de la misma.

2.3.1 Capas de la arquitectura

En la figura 2.1 se muestra la división en capas de software de la arquitectura general del sistema. Se pueden distinguir tres capas: la capa configurable, la capa de sistema y la capa de aplicaciones.

En la capa configurable aparecen aquellas partes del sistema que pueden ser configuradas o intercambiadas, es decir, aquellas partes que se configuran según las necesidades del usuario sin realizar ningún cambio adicional en el sistema. Estas necesidades pueden surgir, por ejemplo, por la necesidad de actualizar a una versión superior alguna de las piezas que conforman esta capa. Por tanto, es factible, por ejemplo, cambiar el dominio en el que va a trabajar el sistema sin que haya que realizar ningún tipo de cambio en la implementación del mismo. De manera análoga, podemos

elegir aquellas fuentes de datos que van a suministrar la información que necesita el sistema para trabajar. Por último, el sistema se ayuda de aplicaciones externas para conseguir llevar a cabo determinados trabajos. Por ejemplo, se utiliza el conjunto de librerías del *Stanford Log-linear Part-Of-Speech Tagger*²⁶ (Toutanova & Manning, 2000; Toutanova et al., 2003) para que el sistema lleve a cabo las tareas de procesamiento del lenguaje natural. Sin embargo, sustituir esta herramienta por otra equivalente, tipo *FreeLing*²⁷ (Padró, 2011; Padró & Stanilovsky, 2012) o *GATE*²⁸ (Cunningham et al., 2011; Cunningham et al., 2013) sería inmediato y el sistema no vería alterado su funcionamiento.

La capa de sistema es aquella donde se encuentra el núcleo central del sistema, es decir, aquella parte del sistema que no cambia y donde está implementada toda la funcionalidad del mismo. Esta capa intermedia hará uso de las herramientas, aplicaciones y paquetes que fueron configuradas en la capa de configuración. En esta capa se encuentran los distintos módulos que llevan a cabo los procesos definidos para el sistema de minería de opiniones propuesto en este trabajo. Además, esta parte sirve como base para la creación de aplicaciones del lado del cliente que consuman las funcionalidades expuestas en este núcleo central del sistema.

La tercera capa, la capa de aplicaciones, es la capa en la que aparecen aquellas aplicaciones que utilizan el núcleo central del sistema de minería de opiniones para dar una serie de servicios a los clientes o usuarios. Es una capa abierta donde los arquitectos de aplicaciones de software pueden aportar diversas soluciones que den servicios de calidad y cubran necesidades reales de la sociedad. Por ejemplo, utilizando el *API* de funcionalidades que ofrece el sistema propuesto se podría crear fácilmente una aplicación que valorase la satisfacción de los clientes cuando realizan compras en un centro comercial, o cuando están valorando la facilidad de uso de un determinado producto. Otro ejemplo de aplicación basada en el conjunto de utilidades que ofrece el sistema podría ser un sistema de ayuda a la decisión (*SAD*) que ayudase a psicólogos a decidir sobre el tratamiento a aplicar a determinados pacientes en función del grado de positividad o negatividad detectado en las opiniones dadas por estos pacientes al ser preguntados sobre determinados temas.

²⁶ <http://nlp.stanford.edu/software/tagger.shtml>, última consulta realizada en Enero de 2015.

²⁷ <http://nlp.lsi.upc.edu/freeling/>, última consulta realizada en Enero de 2015.

²⁸ <https://gate.ac.uk/>, última consulta realizada en Enero de 2015.

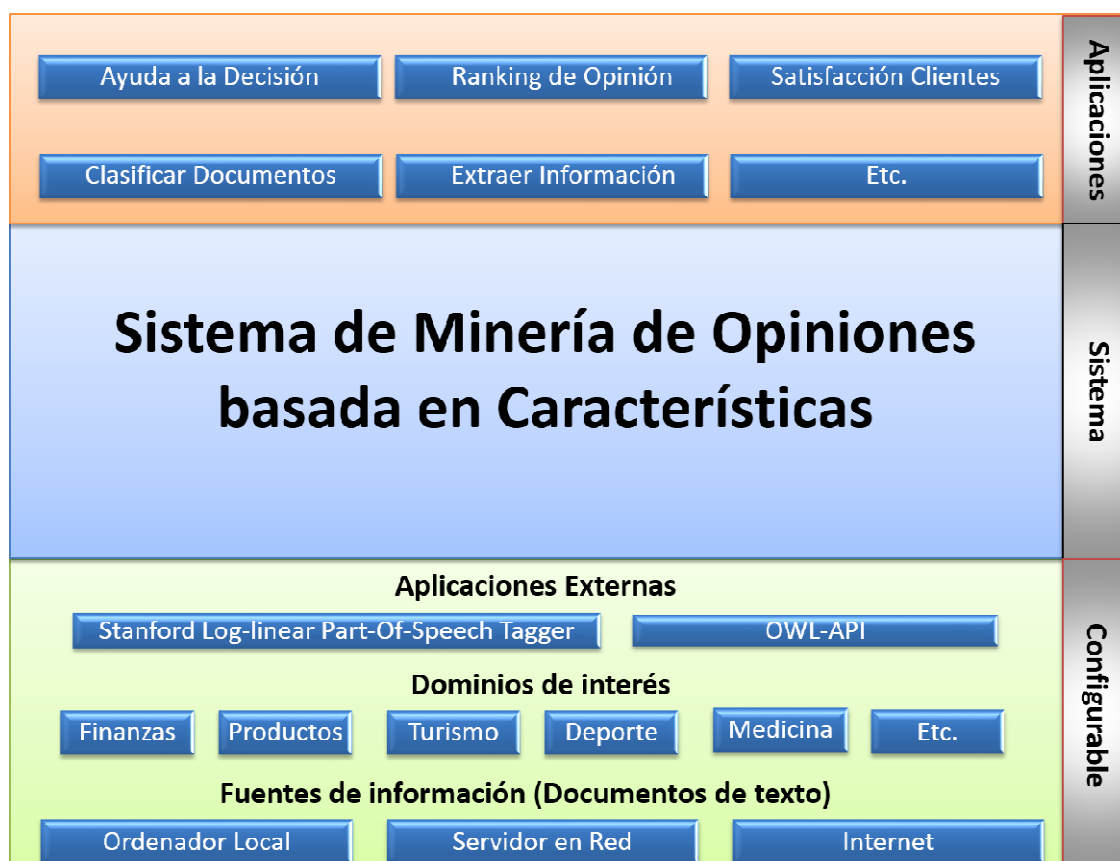


Figura 2.1. Capas de software de la arquitectura del sistema.

2.3.2 Esquema general de la arquitectura

La arquitectura general del sistema que sirve como base de la metodología de minería de opiniones basada en características presentada en esta tesis doctoral se muestra en la figura 2.2.

Esta arquitectura se compone de cuatro módulos principales: el módulo de procesamiento del lenguaje natural (*PLN*), el módulo de identificación de características, el módulo de cálculo de la polaridad de características y el módulo de minería de opiniones o análisis de sentimientos. Cada uno de estos módulos se describe a continuación:

- **Módulo de procesamiento del lenguaje natural.** Este módulo tiene como principal objetivo obtener la estructura morfológica y léxica de cada frase contenida en las opiniones del usuario. Para ello se aplican técnicas de *PLN* sobre el corpus de opiniones de los usuarios.

- **Módulo de identificación de características.** Este módulo tiene como principal objetivo identificar las características presentes en las opiniones de los usuarios. Para ello se utiliza una ontología del dominio que sirva para identificar los términos lingüísticos que representan características.
- **Módulo de cálculo de la polaridad de características.** Este módulo lleva a cabo el cálculo de la polaridad de las características identificadas en las opiniones de los usuarios. Para llevar a cabo este cometido utiliza una base de datos léxica de sentimientos. También se utiliza una base de datos de sinónimos para la agrupación de polaridades de características.
- **Módulo de análisis de sentimientos de opiniones.** Este módulo lleva a cabo el proceso de análisis de sentimientos de las opiniones efectuadas por los usuarios así como de las características contenidas en ellas.

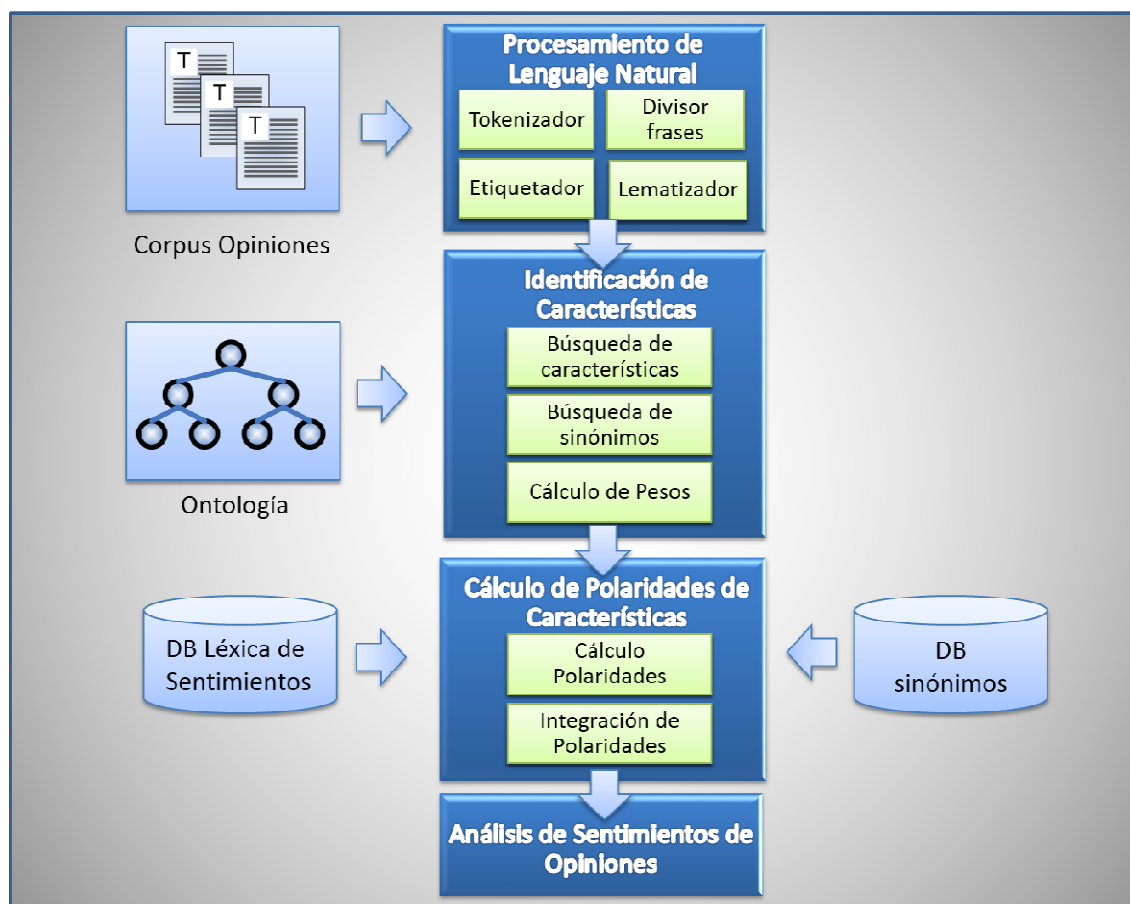


Figura 2.2. Arquitectura general del sistema.

2.4 Módulo de procesamiento del lenguaje natural

El objetivo principal de este módulo es el análisis del texto de las opiniones de forma lingüística. En la implementación de este módulo se ha empleado una herramienta de análisis del lenguaje natural llamada *Stanford Log-linear Part-Of-Speech Tagger* (Toutanova & Manning, 2000; Toutanova et al., 2003). Las funcionalidades básicas que se llevan a cabo dentro de este módulo son las siguientes:

- **Tokenizador:** proceso que se aplica a un texto en lenguaje natural y devuelve una lista de *tokens* separados en palabras individuales.
- **Divisor de oraciones:** proceso que obtiene las oraciones de un texto en lenguaje natural.
- **Etiquetado gramatical (*PosTagger*):** proceso que asigna a cada una de las palabras del texto su categoría gramatical. Algunas categorías gramaticales son: determinante, sustantivo o nombre, pronombre, verbo, adjetivo, adverbio, preposición, conjunción e interjección.
- **Lematizador:** proceso que permite el reconocimiento, la generación y la manipulación de las relaciones morfo-léxicas a partir de cualquier palabra, incluyendo la recuperación de toda su información léxico-genética hasta llegar a una palabra primitiva. En lexicografía, un lema es “una entrada léxica del diccionario en que se suministra diversa información y es a menudo representativa de distintas formas flexionadas”²⁹. Por ejemplo, *ir* es el lema de *voy*, *vas*, *íbamos*, *fueron* y el resto de sus formas conjugadas.

La aplicación de cada uno de los procesos mencionados se realiza por cada una de las frases identificadas en el corpus. Para cada opinión, primero se ejecutan secuencialmente los módulos tokenizador, divisor de oraciones, etiquetado gramatical y lematizador. Después, como resultado, se genera un documento de texto etiquetado morfológicamente.

La figura 2.3 muestra esquemáticamente el funcionamiento del módulo de *PLN*. Como se puede apreciar en el figura, el módulo recibe como entrada un corpus con las opiniones de los usuarios escritas en lenguaje natural (se asume una sola opinión escrita en cada fichero de texto). Una vez llevados a cabo los procesos internos descritos

²⁹ <http://www.hispanoteca.eu/Lexikon%20der%20Linguistik/I/LEMMA%20%20%20Lema.htm>. Última consulta realizada en Enero de 2015. Definición de lema en lexicografía aportada en este sitio Web utilizando la siguiente fuente: *Diccionario de lingüística*, Anaya, p. 174.

en el párrafo anterior, se devuelve un corpus de opiniones etiquetado léxica y morfológicamente.

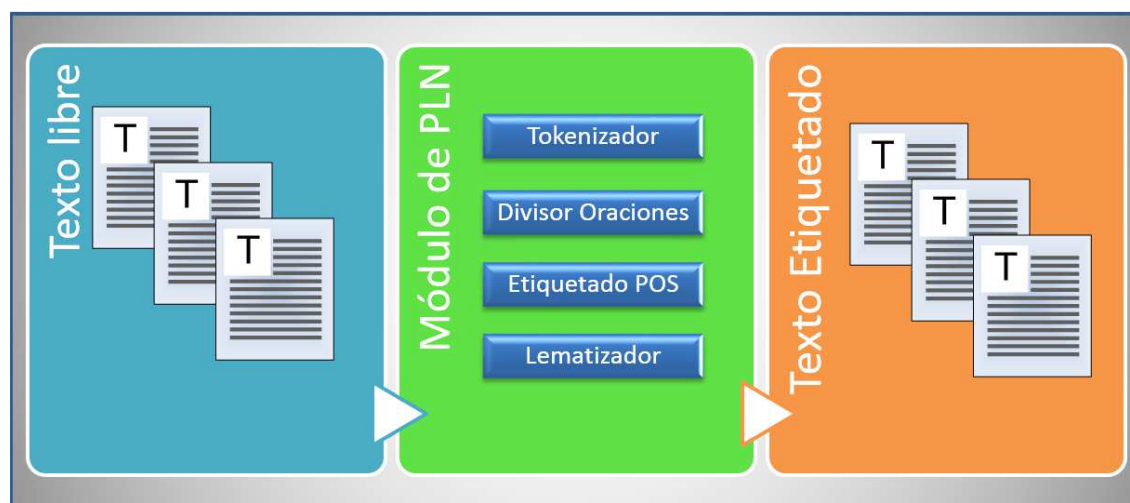


Figura 2.3. Módulo de procesamiento del lenguaje natural.

Como se ha comentado antes, en la implementación de este módulo de *PLN* se ha empleado una herramienta de análisis del lenguaje natural llamada *Stanford Log-linear Part-Of-Speech Tagger*. Esta herramienta contiene tres modelos de etiquetado que están representados en la tabla 2.1. En esta tabla se enumeran las características más relevantes que describen a cada modelo.

Tabla 2.1. Modelos de etiquetado de la herramienta *Stanford Log-linear Part-Of-Speech*.

Etiquetador: *bidirectional-distsim-wsj-0-18.tagger*

Entrenado en el *Wall Street Journal* (WSJ, www.wsj.com) secciones 0-18 usando una arquitectura bidireccional.

Conjunto de etiquetas Penn Treebank.

Rendimiento:

97.28% de precisión con WSJ 19-21

(90.46% de precisión con palabras desconocidas)

Etiquetador: *left3words-wsj-0-18.tagger*

Entrenado en el WSJ secciones 0-18 usando una arquitectura left3words.

Conjunto de etiquetas Penn Treebank.

Rendimiento:

96.97% de precisión con WSJ 19-21

(88.85% de precisión con palabras desconocidas)

Etiquetador: *left3words-distsim-wsj-0-18.tagger*

Entrenado en el WSJ secciones 0-18 usando una arquitectura left3words.

Conjunto de etiquetas Penn Treebank.

Rendimiento:

97.01% de precisión con WSJ 19-21

(89.81% de precisión con palabras desconocidas)

Los etiquetadores bidireccionales son un poco más precisos cuando se trabaja con el idioma inglés, pero el proceso de etiquetado es mucho más lento, y por lo tanto es necesario elegir el etiquetador adecuado en función de las necesidades de velocidad y rendimiento que se tengan. En el desarrollo del sistema presentado en esta tesis la opción elegida ha sido el etiquetador *left3words-WSJ-0-18.tagger*. Se ha elegido esta opción después de considerar las opciones de rendimiento y precisión ofrecidas por cada uno de los etiquetadores disponibles en la herramienta utilizada.

La herramienta *Stanford Log-linear Part-Of-Speech Tagger* está basada en el conjunto de etiquetas *Penn Treebank* (Santorini, 1990) que aparece en la tabla 2.2. Se usan estas etiquetas para la anotación morfosintáctica de lexicones y corpus en distintas lenguas. Dependiendo de la lengua hay atributos que pueden o no especificarse. Una de las ventajas que tiene la forma de definición de estas etiquetas es que se pueden omitir atributos sin ningún problema. Un atributo puede ser omitido cuando expresa un tipo de información que no existe en la lengua o cuando la información no se considera relevante.

El conjunto de etiquetas *Penn Treebank* ha sido utilizado con éxito en multitud de proyectos de índole lingüística para etiquetar textos escritos en alemán, inglés, francés, italiano, holandés, español, búlgaro, ruso, portugués, gallego, chino, swahili, eslovaco, latín, estonio, polaco y textos franceses antiguos. Además, este conjunto de etiquetas es adaptable a otros idiomas si se dispone de un lexicón y un corpus de entrenamiento etiquetado manualmente.

Tabla 2.2. Etiquetas *Penn Treebank*.

Número	Etiqueta	Descripción
1.	CC	Conjunción coordinativa
2.	CD	Número cardinal
3.	DT	Determinante
4.	EX	Existencial
5.	FW	Palabra foránea
6.	IN	Preposición o conjunción subordinada
7.	JJ	Adjetivo
8.	JJR	Adjetivo, comparativo
9.	JJS	Adjetivo, superlativo
10.	LS	Marcador de lista de ítems
11.	MD	Modal
12.	NN	Nombre, singular
13.	NNS	Nombre, plural

14.	NNP	Nombre propio, singular
15.	NNPS	Nombre propio, plural
16.	PDT	Predeterminante
17.	POS	Posesivo
18.	PRP	Pronombre personal
19.	PRP\$	Pronombre posesivo
20.	RB	Adverbio
21.	RBR	Adverbio, comparativo
22.	RBS	Adverbio, superlativo
23.	RP	Partícula
24.	SYM	Símbolo
25.	TO	to
26.	UH	Interjección
27.	VB	Verbo, forma base
28.	VBD	Verbo, pasado
29.	VBG	Verbo, gerundio o participio presente
30.	VBN	Verbo, participio pasado
31.	VBP	Verbo, no-3ª persona del singular presente
32.	VBZ	Verbo, 3ª persona del singular presente
33.	WDT	Wh-determinante
34.	WP	Wh-pronombre
35.	WP\$	Posesivo wh-pronombre
36.	WRB	Wh-adverbio

Para finalizar este apartado, en la figura 2.4 se muestra un extracto de la opinión real de un usuario sobre una película. La opinión está escrita en inglés. Esta será una de las entradas para el módulo de procesamiento de lenguaje natural (*PLN*). Como se puede apreciar en la figura, las opiniones de los usuarios puede que no lleguen escritas en un lenguaje formal y bien estructurado. En ocasiones, dependiendo del dominio en el que estemos trabajando, es usual que los usuarios empleen expresiones coloquiales y utilicen términos de alguna jerga cuando opinan. Además, es frecuente el uso de la ironía en sus expresiones y el uso de frases largas con oraciones subordinadas de difícil seguimiento incluso para un lector humano. Todos estos aspectos dificultan el procesamiento del lenguaje natural. En la sección de validación de la metodología expuesta en esta tesis doctoral se utilizan dos dominios totalmente diferentes: uno con un lenguaje formal y otro con un lenguaje informal. De esta manera se puede evaluar la solución propuesta en distintas situaciones y se puede llegar a conclusiones que sean relevantes.

*“The actors are pretty good for the most part, although Wes Bentley just seemed to be playing the exact same character that he did in American Beauty, only in a new neighborhood. But my biggest kudos go out to Melissa Sagemiller, who holds her own throughout the entire film, and actually has you feeling her character's unraveling. Overall, the film doesn't stick because it doesn't entertain, it's confusing, it rarely excites and it feels pretty redundant for most of its runtime, despite a pretty cool ending and explanation to all of the craziness that came before it.
Oh, and by the way, this is not a horror or teen slasher flick . . . it's just packaged to look that way because someone is apparently assuming that the genre is still hot with the kids. It also wrapped production two years ago and has been sitting on the shelves ever since. Whatever... skip it!”*

Figura 2.4. Extracto de la opinión de un usuario en lenguaje natural e idioma inglés.

Los resultados del análisis morfológico llevado a cabo por el módulo de *PLN* se muestran en la figura 2.5. Esta es la salida del módulo de *PLN* explicado en este apartado. Si miramos la figura vemos que un fragmento de la primera frase etiquetada es el siguiente: *“The/DT actors/NNS are/VBP pretty/RB good/JJ”*. Las etiquetas que aparecen en este fragmento indican que la palabra *The* ha sido etiquetada como determinante, ya que si miramos la tabla 2.2 vemos que la etiqueta *DT* en la posición tres así lo indica. La palabra *actors* ha sido etiquetada como un sustantivo, ya que la etiqueta *NNS* se emplea para etiquetar sustantivos, tal y como se puede apreciar en la entrada trece de la tabla de etiquetas *Penn*. Además, en este caso, la *S* final de la etiqueta *NNS* indica que se trata de un sustantivo en plural. La palabra *are* ha sido etiquetada como verbo en presente según aparece en la entrada treinta y uno de la tabla de etiquetas. A continuación, la palabra *pretty* ha sido etiquetada como adverbio, tal y como podemos ver en la entrada veinte de la tabla de etiquetas. Por último, la palabra *good* ha sido etiquetada como adjetivo, tal y como se puede apreciar en la entrada siete de la tabla de etiquetas *Penn*.

The/DT actors/NNS are/VBP pretty/RB good/JJ for/IN the/DT most/JJS part/NN ./, although/IN Wes/NNP Bentley/NNP just/RB seemed/VBD to/TO be/VB playing/VBG the/DT exact/JJ same/JJ character/NN that/IN he/PRP did/VBD in/IN American/JJ Beauty/NN ./, only/RB in/IN a/DT new/JJ neighborhood/NN ./ But/CC my/PRP\$ biggest/JJS kudos/NNS go/VBP out/RP to/TO Melissa/NNP Sagemiller/NNP ./, who/WP holds/VBZ her/PRP\$ own/JJ throughout/IN the/DT entire/JJ film/NN ./, and/CC actually/RB has/VBZ you/PRP feeling/VBG her/PRP\$ character/NN 's/POS unraveling/NN ./ Overall/RB ./, the/DT film/NN does/VBZ n't/RB stick/VB because/IN it/PRP does/VBZ n't/RB entertain/VB ./, it/PRP 's/VBZ confusing/JJ ./, it/PRP rarely/RB excites/VBZ and/CC it/PRP feels/VBZ pretty/RB redundant/JJ for/IN most/JJS of/IN its/PRP\$ runtime/NN ./, despite/IN a/DT pretty/RB cool/JJ ending/VBG and/CC explanation/NN to/TO all/DT of/IN the/DT craziness/NN that/WDT came/VBD before/IN it/PRP ./ oh/UH ./, and/CC by/IN the/DT way/NN ./, this/DT is/VBZ not/RB a/DT horror/NN or/CC teen/JJ slasher/NN flick/NN .../: it/PRP 's/VBZ just/RB packaged/VBN to/TO look/VB that/DT way/NN because/IN someone/NN is/VBZ apparently/RB assuming/VBG that/IN the/DT genre/NN is/VBZ still/RB hot/JJ with/IN the/DT kids/NNS ./ It/PRP also/RB wrapped/VBD production/NN two/CD years/NNS ago/RB and/CC has/VBZ been/VBN sitting/VBG on/IN the/DT shelves/NNS ever/RB since/IN ./ Whatever/WDT .../: skip/VB it/PRP !/.

Figura 2.5. Análisis morfológico llevado a cabo por el módulo de *PLN*.

2.5 Módulo de identificación de características

El principal objetivo de este módulo es identificar las características presentes en las opiniones de los usuarios. Para ello, este módulo utiliza una ontología del dominio que le ayuda en este proceso de identificación. La ontología del dominio que utilizaremos en nuestra metodología puede ser una ontología disponible en la Web o puede ser una ontología creada por expertos.

También es posible partir de una ontología disponible en la Web y modificarla y/o instanciarla convenientemente por parte de los expertos del dominio.

El correcto funcionamiento de este módulo determinará, en buena medida, el éxito del sistema de clasificación de opiniones basado en características propuesto en este trabajo. De hecho, otros trabajos como (Zhai et al., 2011) ponen de manifiesto el papel fundamental que juega la identificación de características en el proceso de clasificación de sentimientos.

En la figura 2.6 se muestra un esquema gráfico de las entradas y salidas del módulo de identificación de características, así como de los procesos llevados a cabo dentro de este módulo: proceso de identificación o búsqueda de características, proceso de búsqueda de sinónimos y proceso de cálculo del peso de las características.

Este módulo recibe como entrada el texto etiquetado de la opinión de los usuarios (tras el proceso anterior de *PLN*) y una ontología del dominio. En primer lugar se lleva a cabo el proceso de identificación de características. Para ello se utiliza una ontología del dominio para identificar los fragmentos del texto que contienen las clases, instancias, atributos o relaciones de la ontología del dominio. En paralelo con este proceso, también se identifican aquellos fragmentos del texto que contienen términos que son sinónimos de los objetos de la ontología nombrados anteriormente. Una vez que las características se han identificado, a continuación tiene lugar el proceso de cálculo del peso de cada una de las características identificadas. Este proceso básicamente consiste en calcular una puntuación que representa la importancia de una determinada característica dentro de la opinión de un usuario.

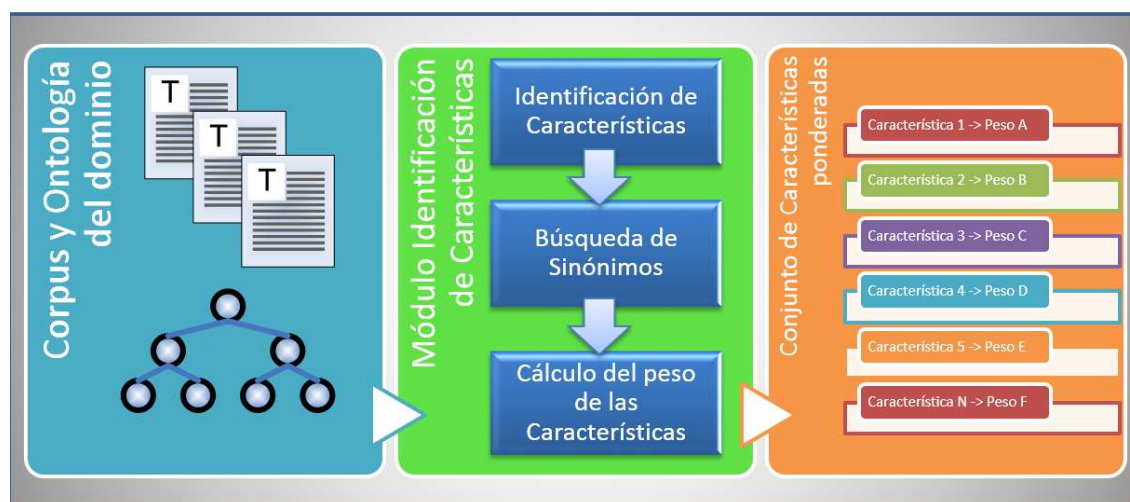


Figura 2.6. Módulo de identificación de características.

Para analizar el funcionamiento de este módulo en detalle, en los siguientes apartados se explica el funcionamiento de cada uno de los procesos que intervienen en él.

2.5.1 Búsqueda de características

El objetivo del proceso de búsqueda de características es identificar aquellas características presentes en las opiniones de los usuarios. Para llevar a cabo esta tarea, tal y como se ha comentado anteriormente, el proceso utiliza una ontología del dominio.

Para ilustrar el conjunto de tareas que tienen lugar dentro de este proceso se va a utilizar un ejemplo sencillo como hilo conductor. Se partirá de una ontología del dominio que representa el dominio de las películas cinematográficas. En la figura 2.7 se muestra un extracto de la ontología del dominio *The Movie Ontology (MO)*³⁰. Esta ontología ha sido instanciada de manera manual utilizando la herramienta *Protégé*³¹ y la base de datos *Internet Movie Database (IMDB)*³².

En la figura 2.7 se puede apreciar como la clase *Movie* se presenta como el concepto central de la ontología. Este concepto es el que tiene mayor número de relaciones con el resto de conceptos o clases de la ontología. Así, por ejemplo, el concepto *Movie* está relacionado con el concepto *Actor*, ya que en una película intervienen un número concreto de actores. Además, el concepto *Movie* está relacionado con los conceptos *Producer* y *Film_Director*, ya que una película tiene uno o varios productores y directores. También se observa que una película tiene una banda sonora asociada y que pertenece a un género determinado, por tanto, la clase *Movie* está relacionada con las clases *SoundTrack* y *Genre* respectivamente. Por último, cabe destacar que una clase puede tener subclases en la ontología. Este es el caso de la clase *Genre*, que tiene como subclases a las clases *Information* y *Entertainment*, ya que existen distintos géneros de una película entre los que se encuentran los de información y entretenimiento respectivamente.

³⁰ La ontología del dominio de las películas cinematográficas *The Movie Ontology* esta disponible en <http://www.movieontology.org>. Última consulta en Enero de 2015.

³¹ *Protégé* es un editor de ontologías gratuito, de código fuente abierto, y está disponible en <http://protege.stanford.edu/>. Última consulta en Enero de 2015.

³² La base de datos en el dominio cinematográfico *IMDb* tiene sus datos disponibles en <http://www.imdb.com>. Última consulta en Enero de 2015.

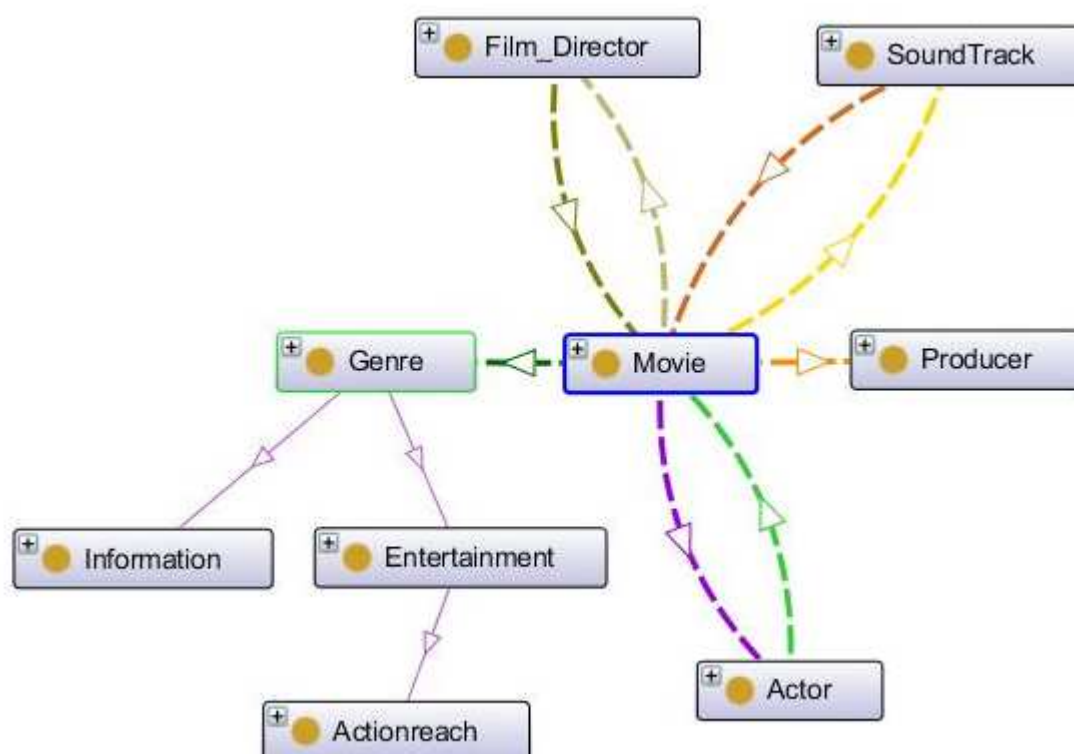


Figura 2.7. Extracto de la ontología del dominio ‘The Movie Ontology’.

Por ejemplo, si se toma como entrada el texto de la opinión del usuario que aparece en la figura 2.5 (texto con la opinión del usuario etiquetado morfológicamente) y la ontología del dominio *MO*, el proceso de identificación de características devuelve un listado con las características identificadas dentro de esa opinión. En la figura 2.8 aparecen resaltadas las características identificadas: *actors*, *Wes Bentley*, *American Beauty*, *Melissa Sagemiller*, *film*, *runtime*, *horror*, *flick*, *genre* y *kids*. Más concretamente, se identifican los términos *Wes Bentley* y *Melissa Sagemiller* como instancias de la clase *Actor*. El término *actors* es un término que representa a la clase *Actor* de la *MO*. Los términos *Film* y *flick* se identifican como características porque son términos sinónimos de la clase *Movie*. En este proceso de identificación de términos sinónimos el sistema usa los *labels* asociados a cada entidad en la ontología del dominio como veremos en detalle en el siguiente apartado. De igual manera, las palabras *kids* y *genre* son identificadas también como características porque están representadas en las clases *Kids* y *Genre* respectivamente.

“The **actors** are pretty good for the most part, although **Wes Bentley** just seemed to be playing the exact same character that he did in **American Beauty**, only in a new neighborhood. But my biggest kudos go out to **Melissa Sagemiller**, who holds her own throughout the entire **film**, and actually has you feeling her character's unraveling. Overall, the **film** doesn't stick because it doesn't entertain, it's confusing, it rarely excites and it feels pretty redundant for most of its **runtime**, despite a pretty cool ending and explanation to all of the craziness that came before it.

Oh, and by the way, this is not a **horror** or teen slasher **flick** . . . it's just packaged to look that way because someone is apparently assuming that the **genre** is still hot with the **kids**. It also wrapped production two years ago and has been sitting on the shelves ever since. Whatever... skip it!”

Figura 2.8. Características identificadas por el módulo de identificación de características en una opinión.

En este punto, gracias a la estructura semántica de las ontologías, las opiniones de usuarios que contengan clases que están directamente relacionadas con la clase *Movie*, como por ejemplo *Actor*, *Producer*, *Genre* y sus respectivas instancias, serán tenidas en cuenta a la hora de calcular la polaridad global de la opinión del usuario. Por ejemplo, en la figura 2.9 se puede apreciar que la palabra *horror* se identifica como una nueva característica porque es una instancia de la clase *Thrilling*, la cual es una subclase de la clase *Imaginational Entertainment*, la cual es una subclase de la clase *Entertainment*, que es una subclase de *Genre*. Finalmente, las clases *Movie* y *Genre* están relacionadas por la propiedad *belongsToGenre*. El mismo razonamiento es válido para la inclusión de la palabra *kids* como una nueva característica. La última característica identificada es *runtime*, ya que existe una relación entre clases en la ontología *MO* llamada *runtime* que implica a la clase *Movie*.

La figura 2.9 muestra gráficamente un extracto de la estructura de la ontología del dominio *MO*. En esta figura se puede observar un ejemplo gráfico de la potencia que ofrece una estructura ontológica de un dominio específico (en este caso el de las películas cinematográficas) a la hora de llevar a cabo el proceso de identificación de características que se propone en este trabajo. En esta figura aparecen representadas con círculos de color amarillo las clases y subclases de la ontología del dominio y con un rombo las instancias de dichas clases.

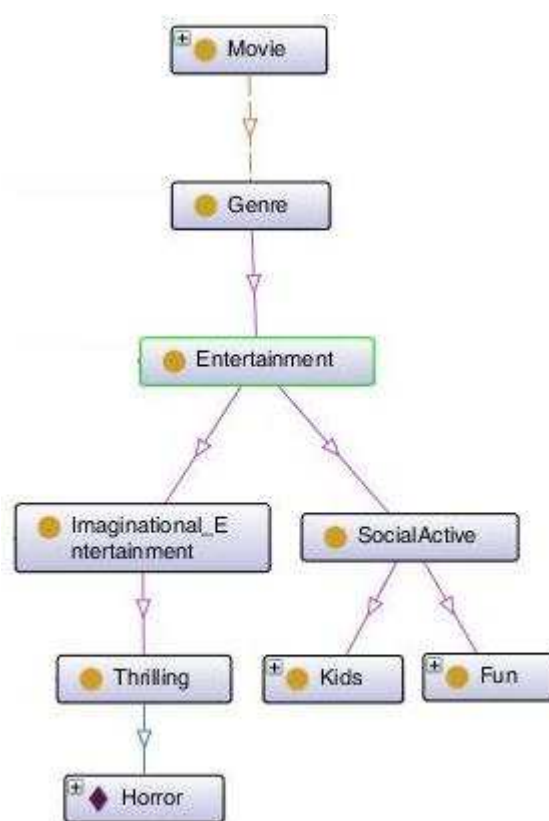


Figura 2.9. Estructura semántica para la ontología *The Movie Ontology*.

2.5.2 Búsqueda de sinónimos

El segundo proceso que tiene lugar dentro del módulo de identificación de características es el proceso de búsqueda de sinónimos. Este proceso se ejecuta en paralelo con el proceso anterior de identificación de características. Es decir, a la vez que se identifican aquellos términos lingüísticos del texto de opinión que tienen una correspondencia directa con uno de los objetos de la ontología del dominio (clases, instancias, atributos o relaciones) se trata de identificar también aquellos términos lingüísticos que son sinónimos de los objetos de la ontología del dominio.

Para llevar a cabo este proceso se necesita utilizar el conjunto de *labels* definido para cada uno de los conceptos de la ontología del dominio en el idioma en el que se esté trabajando. Para cada concepto de la ontología del dominio definir un conjunto de *labels* que contenga todos los posibles sinónimos en cada idioma puede llegar a ser una tarea ardua. De hecho la experiencia de los experimentos llevados a cabo con ontologías del dominio encontradas en la Web nos dice que es muy difícil encontrar una ontología

definida de esta manera. Esta es la razón por la que la metodología propuesta en esta tesis doctoral propone el uso opcional de una base de datos de sinónimos en el idioma en el que se esté trabajando como recurso lingüístico necesario para corregir la posible falta de completitud de la ontología del dominio.

En el ejemplo que se ha visto anteriormente, los términos *film* y *flick* son identificados como características porque son términos sinónimos de la clase *Movie*, es decir, los términos *film* y *flick* pertenecían al conjunto de *labels* definidos para la clase *Movie* en la ontología del dominio cinematográfico para el idioma inglés.

2.5.3 Cálculo del peso de las características

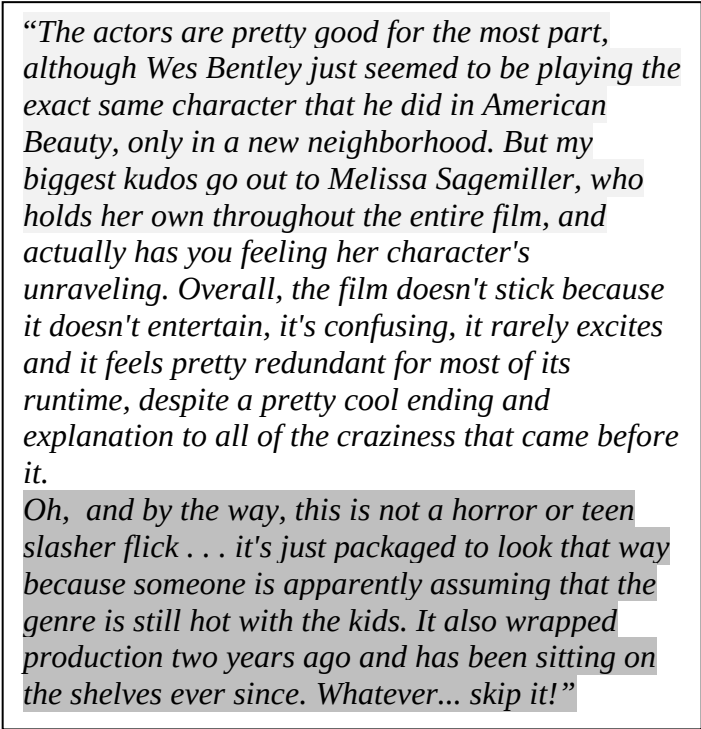
El tercer y último proceso que tiene lugar dentro del módulo de identificación de características es el proceso de cálculo del peso de las características identificadas anteriormente. El objetivo de este proceso es asignar un peso concreto a cada una de las características identificadas, es decir, ponderar cada característica en base a ciertos criterios que se van a estudiar en este apartado.

Al investigar sobre los métodos existentes de identificación de características para compararlos con la propuesta de este trabajo se ha visto que la mayoría de estos métodos asignan la misma importancia a todas las características identificadas en un texto (Baccianella et al., 2009; Zhou & Chaovalit, 2008; Zhao & Li, 2009). A diferencia de estos trabajos, en la metodología presentada en esta tesis doctoral se propone un sistema de ponderación a aplicar a cada una de las características identificadas en la opinión de los usuarios. En concreto, se asigna un peso específico para cada característica identificada en una opinión en base a las siguientes consideraciones:

- No todas las características relacionadas con la misma clase tienen la misma relevancia.
- Las características citadas mayoritariamente en las opiniones de los usuarios son las más relevantes.
- Desde un punto de vista del análisis del discurso periodístico la polaridad del último párrafo dentro de una opinión coincide, en la mayor parte de los casos, con la polaridad global de dicha opinión (Moreno et al., 2010). Empezar o acabar un discurso (u opinión) con un punto de vista positivo o negativo realza la polaridad positiva o negativa del discurso en cuestión. A este respecto, el lugar o lugares en los que una característica aparece en las opiniones debería ser

tenido en cuenta a la hora de calcular su peso en cada opinión. Por ejemplo, si la última frase de la opinión de un usuario sobre una película cinematográfica es la siguiente: “*undoubtedly a film with an amazing soundtrack*”, entonces esto realza la polaridad positiva de la característica *soundtrack* en la opinión del usuario y, por lo tanto, esa característica debería tener una puntuación o peso mayor a las demás.

En la figura 2.10 se muestra un extracto de la opinión de un usuario expresada en lenguaje natural. En ella se pueden observar en distinto color las áreas de influencia de las características identificadas en la opinión. Aparecen resaltadas las áreas de la opinión donde las características que se identifiquen allí tendrán mayor peso a la hora de calcular la polaridad global de la misma. Se muestra resaltada en un color más intenso la parte final de la opinión ya que, según el criterio de ponderación establecido para la metodología propuesta, el inicio y el final de la opinión son las zonas en las que las características tendrán mayor peso, especialmente en la parte final. Por este motivo se definen los parámetros del sistema Z_1 , Z_2 y Z_3 , que representan la importancia (ponderación) de las ocurrencias de las características en el inicio, la parte media y el final de la opinión del usuario. Estos parámetros se tendrán en cuenta a la hora de asignar un peso a cada una de las características identificadas en la opinión.



“The actors are pretty good for the most part, although Wes Bentley just seemed to be playing the exact same character that he did in American Beauty, only in a new neighborhood. But my biggest kudos go out to Melissa Sagemiller, who holds her own throughout the entire film, and actually has you feeling her character's unraveling. Overall, the film doesn't stick because it doesn't entertain, it's confusing, it rarely excites and it feels pretty redundant for most of its runtime, despite a pretty cool ending and explanation to all of the craziness that came before it.

Oh, and by the way, this is not a horror or teen slasher flick . . . it's just packaged to look that way because someone is apparently assuming that the genre is still hot with the kids. It also wrapped production two years ago and has been sitting on the shelves ever since. Whatever... skip it!”

Figura 2.10. Áreas de influencia de la posición relativa de características dentro de una opinión.

De lo anterior se deduce que una vez identificadas todas las características se calculan las puntuaciones o pesos asignados a cada una de ellas dentro de cada opinión. Por lo que es necesario disponer de una fórmula para calcular el peso asignado a una característica dentro de la opinión de un usuario. En esta fórmula se tiene que tener en cuenta la posición relativa de la expresión lingüística que representa a la característica dentro de la opinión. Para ello se divide el texto de la opinión en tres partes iguales: el inicio, la parte media y el final. El cálculo del peso de la característica f dentro de la opinión de usuario i ($opinion_i$) se define en la fórmula 2.1. A continuación explicamos las partes de la misma.

$$peso (f, opinion_i) = Z_1 * |O_1| + Z_2 * |O_2| + Z_3 * |O_3| \quad (2.1)$$

En la fórmula 2.1, los términos $|O_1|$, $|O_2|$ y $|O_3|$ representan el número de ocurrencias de la característica f en la opinión de usuario $opinion_i$. Concretamente:

- $|O_1|$: representa el número de ocurrencias de la característica f en la parte inicial de la opinión $opinion_i$.
- $|O_2|$: representa el número de ocurrencias de la característica f en la parte media de la opinión $opinion_i$.
- $|O_3|$: representa el número de ocurrencias de la característica f en la parte final de la opinión $opinion_i$.

Los términos Z_1 , Z_2 y Z_3 representan la importancia de las ocurrencias de la característica f en la opinión de usuario $opinion_i$. Concretamente:

- Z_1 : representa la importancia de las ocurrencias característica f en la parte inicial de la opinión $opinion_i$.
- Z_2 : representa la importancia de las ocurrencias característica f en la parte media de la opinión $opinion_i$.
- Z_3 : representa la importancia de las ocurrencias característica f en la parte final de la opinión $opinion_i$.

En la figura 2.10 se muestra mediante colores una representación del significado de los parámetros Z_1 , Z_2 y Z_3 . Los textos de opinión se dividen en tres partes iguales teniendo en cuenta el número de palabras del texto, y los valores para los parámetros del sistema Z_1 , Z_2 y Z_3 ponderan el impacto de cada característica en cada parte de la opinión. En los resultados experimentales que se muestran en el Capítulo IV se destaca la importancia de estos tres parámetros, haciendo una comparación entre los valores de los mismos y la

precisión obtenida a la hora de calcular la polaridad de las características identificadas por el sistema.

Por ejemplo, si $Z_1 = 80$, $Z_2 = 50$, $Z_3 = 100$, $opinion_x$ representa la opinión x y la característica *movie* aparece cuatro veces en el texto (una al inicio, otra en el medio del texto y dos veces al final de la opinión x) entonces el peso de la característica *movie* en la opinión x es el siguiente:

$$peso(movie, opinion_x) = 80 * 1 + 50 * 1 + 100 * 2 = 330$$

2.6 Módulo de cálculo de polaridades de características

El objetivo principal de este módulo es llevar a cabo el cálculo de la polaridad asociada a cada característica en las opiniones de los usuarios. Para lo que se hace uso de una base de datos léxica de sentimientos específica del idioma en el que se vaya a trabajar. Por ejemplo, en esta tesis doctoral para las opiniones escritas en inglés se utiliza la base de datos léxica de sentimientos *SentiWordNet* (Esuli y Sebastiani, 2006).

En la figura 2.11 se muestra un esquema gráfico de las entradas y salidas del módulo de cálculo de la polaridad de características así como de los procesos que lo componen. En la figura se puede ver que este módulo recibe como entrada un conjunto de características ponderadas para cada opinión. Se llevarán a cabo dos procesos internos: el proceso de cálculo de la polaridad de las características identificadas y la agrupación de polaridades. La salida de este módulo es un conjunto de características con sus polaridades asociadas en cada una de las opiniones.

El proceso de cálculo de la polaridad de las características identificadas por el sistema calculará la polaridad positiva, negativa o neutra de cada una de ellas en cada opinión de usuario. Este proceso utiliza las operaciones sobre espacios vectoriales explicadas en el Capítulo I de la presente tesis doctoral.

El proceso de agrupación de polaridades de las características se encarga de agrupar las polaridades de aquellas características identificadas por el sistema que hacen referencia al mismo objeto del dominio en el que se este trabajando. Para realizar este proceso el sistema utiliza, de nuevo, operaciones sobre espacios vectoriales.

A continuación se explica detalladamente cada uno de estos procesos.

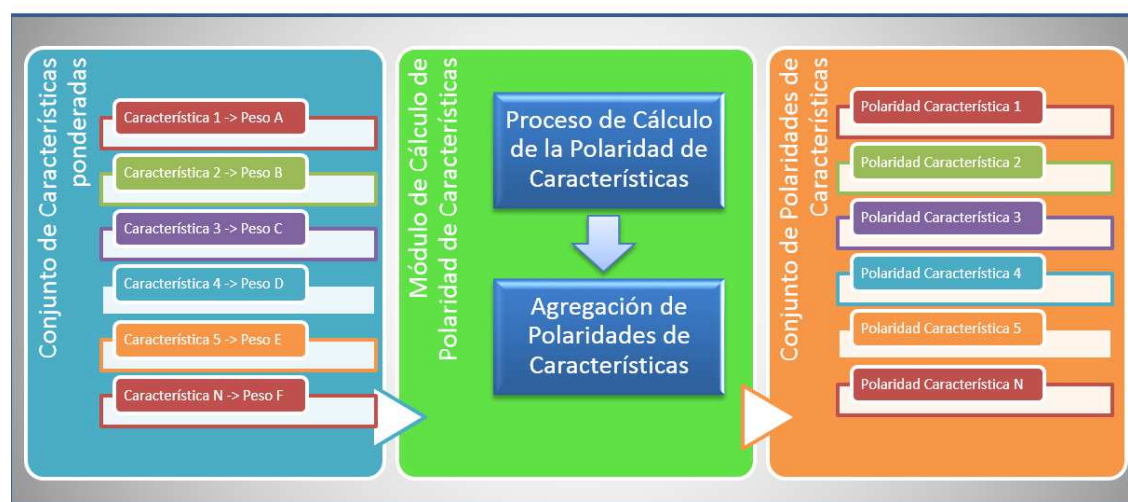


Figura 2.11. Módulo de cálculo de la polaridad de características.

2.6.1 Cálculo de la polaridad de las características identificadas

En este apartado se describe el proceso de cálculo de la polaridad de las características identificadas por el sistema en las opiniones. Este proceso utiliza una base de datos léxica de sentimientos. Como ya se ha dicho desgraciadamente no existe actualmente un recurso lingüístico de calidad con estas características para un idioma distinto del inglés, tal y como se refleja en el trabajo de (Balahur, 2011). Este es uno de los motivos por los que se han centrado las investigaciones realizadas en este trabajo en opiniones de usuarios escritas en inglés. A pesar de que la metodología aportada en esta tesis es independiente del idioma en el que se aplique, se necesita implementar un sistema de software que permita validar la misma sobre un idioma y dominio concretos. En adelante cuando se hable de la base de datos léxica de sentimientos que utiliza la metodología estaremos hablando implícitamente de *SentiWordNet (SWN)*.

Como se ha comentado en el Capítulo I, *SWN* (Esuli y Sebastiani, 2006) es una base de datos de sentimientos que proporciona valores de positividad y negatividad a nombres, adjetivos, verbos y adverbios en el idioma inglés. En la metodología propuesta en esta tesis doctoral es necesario recuperar estos valores para todas las palabras que están situadas cerca de la expresión lingüística que representa una característica dentro de una opinión.

Se ha visto también que *SWN* no proporciona el sentido neutral u objetivo de las palabras que aparecen en la base de datos por lo que se tiene que calcular. Se denomina *SentiNeu (palabra)* al sentido neutral del término lingüístico representado por *palabra*.

De manera análoga se llama *SentiPos (palabra)* y *SentiNeg (palabra)* a los sentidos positivo y negativo respectivamente. El sentido neutral de una palabra se calcula según se muestra en la fórmula 2.2.

$$SentiNeu (palabra) = 1 - SentiPos (palabra) - SentiNeg (palabra) \quad (2.2)$$

Los valores de sentimiento *SentiPos (palabra)* y *SentiNeg (palabra)* se calculan como la media de los sentidos positivos y negativos de la palabra dentro de la base de datos SWN. Los valores de sentimiento de SWN son siempre números positivos o cero. Las fórmulas 2.3 y 2.4 muestran la forma en la que se calculan los valores de *SentiPos (palabra)* y *SentiNeg (palabra)* respectivamente.

$$SentiPos (palabra) = \frac{\sum_{i \in WS_{pos}} WS(palabra)_i}{|WS(palabra)_{i \in WS_{pos}}|} \quad (2.3)$$

$$SentiNeg (palabra) = \frac{\sum_{i \in WS_{neg}} WS(palabra)_i}{|WS(palabra)_{i \in WS_{neg}}|} \quad (2.4)$$

En la fórmula 2.3 se representa mediante $|WS(palabra)_{i \in WS_{pos}}|$ el número de valores de sentimiento con polaridad positiva del término *palabra* en la base de datos de sentimientos SWN. Las siglas de la función WS son la abreviatura de las palabras inglesas *Word Senses* y representan los sentidos positivos o negativos de una palabra en SWN. El sumatorio $\sum_{i \in WS_{pos}} WS(palabra)_i$ representa la suma de todos los valores de sentimiento positivos del término *palabra* en SWN. De igual manera podemos interpretar los valores $|WS(palabra)_{i \in WS_{neg}}|$ y $\sum_{i \in WS_{neg}} WS(palabra)_i$ de la fórmula 2.4.

Para calcular los valores de polaridad asociados a una característica identificada dentro de la opinión de un usuario se necesitan conocer los valores de sentimiento de cada una de las palabras cercanas a la expresión lingüística de la característica dentro de dicha opinión. Se pueden obtener estas palabras de distintas maneras. En el presente trabajo se han implementado cuatro métodos distintos:

- **Método *N_GRAM Before*.** Este método obtiene las *N_GRAM* palabras anteriores a la expresión lingüística de la característica en la opinión del usuario a la hora de calcular la polaridad asociada a dicha característica.

- **Método *N_GRAM After*.** Este método obtiene las *N_GRAM* palabras posteriores a la expresión lingüística de la característica en la opinión del usuario a la hora de calcular la polaridad asociada a dicha característica.
- **Método *N_GRAM Around*.** Este método obtiene las *N_GRAM* palabras anteriores a la expresión lingüística de la característica en la opinión del usuario junto con las *N_GRAM* palabras posteriores a la expresión lingüística de la característica a la hora de calcular la polaridad asociada a dicha característica.
- **Método *All Phrase*.** Este método obtiene todas las palabras que se encuentran en la misma frase que la expresión lingüística de la característica en la opinión del usuario a la hora de calcular la polaridad asociada a dicha característica.

Como se ha visto, a excepción del método *All Phrase*, estos métodos requieren el uso del parámetro del sistema *N_GRAM*, el cual indica el número de palabras cercanas a la característica que son tenidas en cuenta en el proceso de cálculo de la polaridad de la misma. El parámetro *N_GRAM* es un parámetro utilizado en multitud de trabajos sobre minería de opiniones basada en características. En el trabajo de (Zhai et al., 2011) desempeña un papel fundamental ya que toda la investigación gira alrededor de este concepto.

Las figuras 2.12, 2.13, 2.14, 2.15 muestran gráficamente las palabras cercanas a la expresión lingüística de una de las características identificadas que denotaremos en verde como *{característica}*. En esas figuras se ha establecido el parámetro del sistema *N_GRAM* con valor igual a tres. Las palabras tenidas en cuenta para el cálculo de la polaridad de las características aparecen resaltadas en amarillo y, como se aprecia, cada método toma un conjunto de palabras distinto. De ahí que los resultados para el cálculo de las polaridades de las características difiera significativamente de un método a otro. En la figura 2.15 se aprecia como el método *All Phrase* no tiene en cuenta el parámetro del sistema *N_GRAM* y analiza todas las palabras que se encuentran en la misma frase que contiene a la expresión lingüística que representa a la característica.

...	Pal. 1	Pal. 2	Pal. 3	Pal. 4	{característica}	Pal. 5	Pal. 6	Pal. 7	Pal. 8	...

Figura 2.12. Cálculo de la polaridad de características con el método *N_GRAM Before* cuando *N_GRAM* es igual a tres.

...	Pal. 1	Pal. 2	Pal. 3	Pal. 4	{característica}	Pal. 5	Pal. 6	Pal. 7	Pal. 8	...

Figura 2.13. Cálculo de la polaridad de características con el método *N_GRAM After* cuando *N_GRAM* es igual a tres.

...	Pal. 1	Pal. 2	Pal. 3	Pal. 4	[característica]	Pal. 5	Pal. 6	Pal. 7	Pal. 8	...

Figura 2.14. Cálculo de la polaridad de características con el método *N_GRAM Around* cuando *N_GRAM* es igual a tres.

...	Pal. 1	Pal. 2	Pal. 3	Pal. 4	[característica]	Pal. 5	Pal. 6	Pal. 7	Pal. 8	...

Figura 2.15. Cálculo de la polaridad de características con el método *All Phrase*.

De lo anterior se deduce que se necesita una la función que devuelva el conjunto de *NGRAM* palabras alrededor de la característica *característica* dentro de la opinión *opinion_x* obtenidas aplicando el método *metodo*. Esta función se representa de la siguiente manera: *ConjuntoPalabras (metodo, N_GRAM, característica, opinion_x)*. En esta función el parámetro *metodo* tomará uno de los siguientes valores: *N_GRAM_Before*, *N_GRAM_After*, *N_GRAM_Around* o *All_Phrase*. El parámetro *N_GRAM* tomará un valor numérico positivo. El parámetro *característica* contendrá la expresión lingüística que representa una característica en la opinión del usuario. Por último el parámetro *opinion_x* contendrá el valor que identifica a la opinión que se esté analizando.

A continuación se muestra mediante un ejemplo el resultado de la función *ConjuntoPalabras (metodo, N_GRAM, característica, opinion_x)* al instanciarla con cada uno de los métodos citados en el párrafo anterior y aplicarla sobre una frase de una opinión concreta. Si se tiene la opinión *opinion_x* que contiene la siguiente frase: “*This film has good argument but what I liked most was the interpretation of Scarlett Johansson because she is great and makes a really good job*”, y se está calculando la polaridad de la característica *Scarlett Johansson*, entonces la función *ConjuntoPalabras (metodo, N_GRAM, característica, opinion_x)* devuelve los siguientes valores al instanciarla con cada uno de los métodos mostrados y el valor del parámetro *N_GRAM* igual seis:

- *ConjuntoPalabras (N_GRAM_Before, 6, Scarlett Johansson, opinion_x) = { liked, most, was, the, interpretation, of }*
- *ConjuntoPalabras (N_GRAM_After, 6, Scarlett Johansson, opinion_x) = { because, she, is, great, and, makes }*
- *ConjuntoPalabras (N_GRAM_Around, 6, Scarlett Johansson, opinion_x) = { liked, most, was, the, interpretation, of, because, she, is, great, and, makes }*
- *ConjuntoPalabras (All_Phrase, 6, Scarlett Johansson, opinion_x) = { This, film, has, good, argument, but, what, I, liked, most, was, the, interpretation, of, because, she, is, great, and, makes, a, really, good, job }*

Como se ha comentado anteriormente la polaridad de cada característica identificada en la opinión de un usuario viene representada por un vector euclídeo asociado a la característica en dicha opinión. La fórmula 2.5 define la expresión del vector de polaridad $\vec{V}$ asociado a la característica f dentro de la opinión $opinion_x$.

$$\vec{V}(f, opinion_x) = peso(f, opinion_x) * (SentiPos(f, opinion_x), SentiNeg(f, opinion_x), SentiNeu(f, opinion_x)) \quad (2.5)$$

En la fórmula 2.5 el valor $peso(f, opinion_x)$ representa el peso asignado a la característica f dentro de la opinión $opinion_x$ según se ha visto en la fórmula 2.1. Los valores $SentiPos(f, opinion_x)$, $SentiNeg(f, opinion_x)$ y $SentiNeu(f, opinion_x)$ representan los valores medios de sentimiento positivo, negativo y neutro de las palabras cercanas a la expresión lingüística que identifica a la característica f dentro de la opinión $opinion_x$. Los valores $SentiPos(f, opinion_x)$, $SentiNeg(f, opinion_x)$ y $SentiNeu(f, opinion_x)$ se calcularán según las ecuaciones (2.6), (2.7) y (2.8) que se muestran a continuación.

$$SentiPos(f, opinion_x) = \frac{\sum_{i \in \text{ConjuntoPa labras (metodo, NGRAM, f, opinion_x)}} SentiPos(palabra_i)}{|\text{ConjuntoPa labras (metodo, NGRAM, f, opinion_x)}|} \quad (2.6)$$

$$SentiNeg(f, opinion_x) = \frac{\sum_{i \in \text{ConjuntoPa labras (metodo, NGRAM, f, opinion_x)}} SentiNeg(palabra_i)}{|\text{ConjuntoPa labras (metodo, NGRAM, f, opinion_x)}|} \quad (2.7)$$

$$SentiNeu(f, opinion_x) = 1 - SentiPos(f, opinion_x) - SentiNeg(f, opinion_x) \quad (2.8)$$

En la fórmula 2.6 la expresión $\sum_{i \in \text{ConjuntoPa labras (metodo, NGRAM, f, opinion_x)}} SentiPos(palabra_i)$ representa la suma todos los valores de sentimiento positivos en SWN para las N_GRAM palabras cercanas a la expresión lingüística de la característica f en la opinión $opinion_x$ obtenidas según el método de obtención de palabras adyacentes *metodo*. En esta misma fórmula la expresión $|\text{ConjuntoPa labras (metodo, NGRAM, f, opinion_x)}|$ representa el número de palabras adyacentes a la característica f en la opinión $opinion_x$ utilizadas para calcular el valor de $SentiPos(f, opinion_x)$. De manera análoga se puede interpretar el valor

$\sum_{i \in \text{ConjuntoPa labras (metodo, NGRAM, f, opinion_x)}} SentiNeg(palabra_i)$ de la fórmula 2.7.

Según se ha visto en la expresión 2.5, el vector de polaridad de las características identificadas dentro de la opinión está representado mediante tres coordenadas. De esta manera, cada característica vendrá definida por su vector euclídeo $\vec{V}$ en R^3 . Un vector euclídeo se representa mediante dos puntos: el origen y el destino (Strang, 2009). En nuestra metodología el punto de origen es siempre $(0, 0, 0)$ ya que todos los vectores de polaridad partirán del origen de coordenadas en el espacio R^3 . Por este motivo la expresión del vector de polaridad asociado a una característica determinada se reduce a expresar el punto de destino. Un vector de estas características en el espacio R^3 también recibe el nombre de vector de posición (Strang, 2009). Un vector de posición se expresa de la siguiente manera: $\vec{V} = (x, y, z)$ donde la coordenada x denota el sentido positivo, la coordenada y el sentido negativo y la coordenada z el sentido neutro u objetivo. En la figura 2.16 se muestra una representación gráfica del vector $\vec{V} = (x, y, z)$. Una característica con sentido positivo estricto (aquella cuyas coordenadas de sentidos negativo y neutro son 0) vendrá determinada por el vector de posición $\vec{V} = (N, 0, 0)$, donde N es un número positivo. Por analogía, los vectores $\vec{V} = (0, N, 0)$ y $\vec{V} = (0, 0, N)$ representan los sentidos estrictamente negativo y neutro respectivamente.

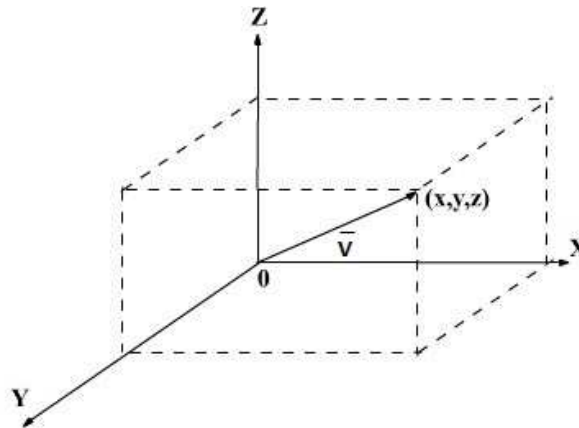


Figura 2.16. Vector de posición que representa la polaridad de una característica.

Para mayor aportar mayor claridad, sean A , B , C y D cuatro características identificadas dentro de la opinión de un usuario. En la figura 2.17 se muestra una representación gráfica de los vectores de posición que representan las polaridades de las características A , B , C y D dentro de la opinión de un usuario. Estos vectores de posición, definirán la polaridad de cada característica como positiva, negativa o neutra de la forma que se muestra a continuación.

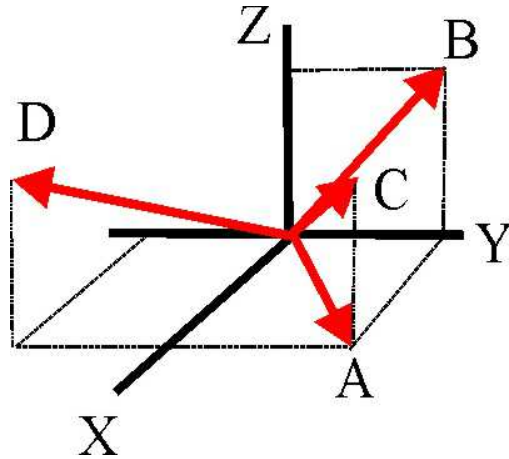


Figura 2.17. Vectores de posición que representan las polaridades de las características A, B, C y D.

La polaridad de una característica f dentro de la opinión de usuario $opinion_x$, llamada $Polaridad(f, opinion_x)$, se calcula según se describe en la fórmula 2.9.

$$Polaridad(f, opinion_x) = \begin{cases} +x \text{ (positiva)} & \text{si } estaContenido(P, P_{Pos}) = verdadero \\ -y \text{ (negativa)} & \text{si } estaContenido(P, P_{Neg}) = verdadero \\ 0 \text{ (neutra)} & \text{si } estaContenido(P, P_{Neu}) = verdadero \end{cases} \quad (2.9)$$

Los componentes que aparecen en la fórmula 2.9 tienen el siguiente significado:

- $P = (x, y, z)$ es el punto de destino del vector $\vec{V}(f, opinion_x)$.
- P_{Pos} es la pirámide geométrica cuyo volumen está compuesto de todos los puntos de destino de los vectores de posición con dirección positiva.
- P_{Neg} es la pirámide geométrica cuyo volumen está compuesto de todos los puntos de destino de los vectores de posición con dirección negativa.
- P_{Neu} es la pirámide geométrica cuyo volumen está compuesto de todos los puntos de destino de los vectores de posición con dirección neutra.
- ***estaContenido*** (P, P_x) es una función que devuelve *verdadero* si el punto P es un punto interior de la pirámide geométrica P_x y *falso* en caso contrario.

En la figura 2.18 se muestra una representación gráfica de un sistema de coordenadas en R^3 en el cual aparecen representadas las pirámides geométricas P_{Pos} , P_{Neg} y P_{Neu} .

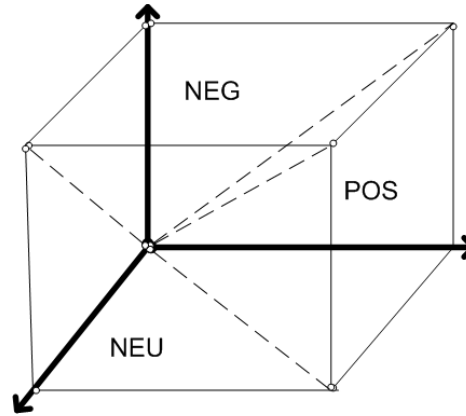


Figura 2.18. Pirámides geométricas de Polaridad en R^3 .

Para implementar la función $estaContenido(P, P_x)$ es necesario determinar cuando el punto dado $P = (x, y, z)$ es un punto interior de las pirámides geométricas P_{Pos} , P_{Neg} o P_{Neu} . Por lo que se necesitan los elementos de topología definidos en (Apostol, 2006) que se explican a continuación:

- **Punto interior.** Sea S un subconjunto de R^n y P un punto de R^n . Además $P \in S$. Entonces se llama a P punto interior de S si existe una N -bola centrada en P y contenida en S .
- **N -bola.** Sea P un punto de R^n y R un número positivo dado. El conjunto de todos los puntos X en R^n tal que $||X - P|| < R$ es llamado N -bola abierta de radio R y centro P . Este conjunto se expresa como $B(P)$ o $B(P, R)$. La figura 2.19 muestra la representación gráfica de una N -bola de centro en P y radio r .

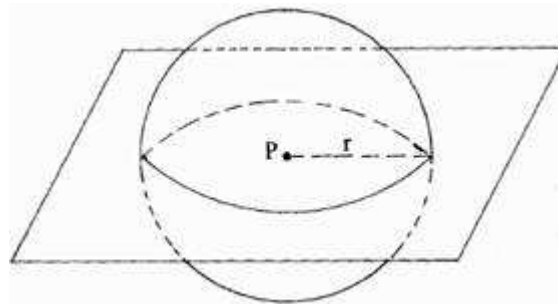


Figura 2.19. Representación gráfica de una N -bola en R^3 con centro en P y radio r .

En otras palabras, cada punto interior P del subconjunto S podría ser rodeado por una N -bola $B(P) \subseteq S$. El conjunto de puntos interiores de S se llama interior de S . Estos conceptos se manejan para calcular la polaridad de una característica teniendo en cuenta su vector de polaridad asociado en el espacio R^3 .

Partiendo de las definiciones anteriores, para saber si un punto del espacio R^3 es un punto interior de las pirámides geométricas P_{Pos} , P_{Neg} o P_{Neu} se necesita conocer si este punto pertenece al conjunto de puntos del espacio que determinan el volumen de la pirámide en cuestión. En la figura 2.20 se muestra una representación gráfica del volumen de una pirámide geométrica en el espacio R^3 . Este volumen viene determinado por la siguiente fórmula: $Volumen\ Pirámide = 1/3 (\text{Área Base} * \text{Altura})$. El área de la base de la pirámide es un cuadrado por lo que su área será: $\text{Área Cuadrado} = lado * lado$, donde *lado* será la longitud de cada uno de los lados que componen el cuadrado. Por último la altura de la pirámide es la longitud de la línea que aparece representada en rojo en la figura 2.20, y es fácilmente calculable teniendo en cuenta la longitud de la apotema de la pirámide (dibujada en color azul en la figura), ya que se toma como referencia el triángulo equilátero formado por la apotema, la altura de la pirámide y la mitad de la longitud de los lados de la base de la pirámide.

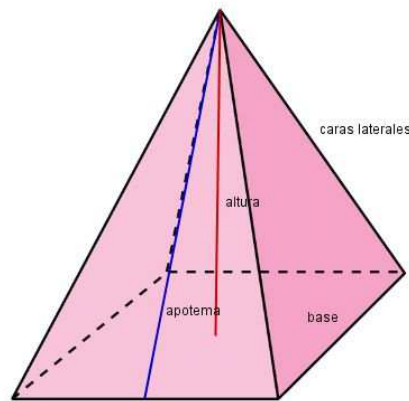


Figura 2.20. Representación gráfica del volumen de una pirámide geométrica.

En la figura 2.21 se muestra un algoritmo que explica el comportamiento de la función $estaContenido(P, P_x)$ donde P es un punto del espacio R^3 y P_x representa una pirámide geométrica en R^3 . En negrita aparecen las primitivas del lenguaje de pseudocódigo empleado. La función $estaContenido(P, P_x)$ devuelve *verdadero* si el punto P es un punto interior de la pirámide geométrica P_x y *falso* en caso contrario.

Funcion $estaContenido (Punto P, Piramide P_x) \{$

Si ($puntoInterior (P, P_x)$) {
 devuelve verdadero;
} **en caso contrario** {
 devuelve falso;
}

}

Figura 2.21. Algoritmo que muestra la implementación de la función $estaContenido(P, P_x)$.

En la figura 2.22 se muestra el algoritmo que representa a la función *puntoInterior* (P, P_x). Esta función recibe como parámetros un punto y una pirámide geométrica en R^3 . Devolverá *verdadero* si existe una N-bola centrada en P y contenida en P_x y *falso* en caso contrario.

Funcion *puntoInterior* (Punto P , Piramide P_x) {

```

    Si (N-Bola( $P, P_x$ ) ) {
        devuelve verdadero;
    } en otro caso {
        devuelve falso;
    }
}
```

Figura 2.22. Algoritmo que muestra la implementación de la función *puntoInterior*(P, P_x).

En la figura 2.23 se muestra la implementación de la función *N-Bola*(P, P_x). Esta función utiliza una constante llamada *Radius*, definida previamente y que representa el radio de la N-bola con centro en el punto P , para saber si un punto dado es un punto interior de la pirámide P_x . Se representa mediante $P_x.listaPuntos$ a un subconjunto de R^3 formado por todos los puntos interiores de la pirámide P_x . La función *N-Bola*(P, P_x) devolverá *verdadero* si $\| P_x.listaPuntos.obtener(i) - P \| < Radius$, donde $P_x.listaPuntos.obtener(i)$ representa a un punto interior de la pirámide geométrica P_x y P es el punto que estamos evaluando si se trata de un punto interior de P_x .

Funcion *N-Bola* (Point P , Pyramid P_x) {

```

    entero  $i = 0$ ;

    mientras (  $i < P_x.listaPuntos.tamaño$  ){

        Si (  $\| P_x.listaPuntos.obtener(i) - P \| < Radius$  ) {
            devuelve verdadero;
        }
         $i = i + 1$ ;
    }

    devuelve falso;
}
```

Figura 2.23. Algoritmo que muestra la implementación de la función *N-Bola*(P, P_x).

Para ilustrar el proceso de cálculo de las polaridades de las características identificadas en una opinión se seguirá el hilo conductor del ejemplo presentado en la sección anterior. Así, la polaridad de la característica *actors* en la opinión $opinion_x$ se calcula de la manera que se muestra a continuación. La frase dentro de la opinión del usuario $opinion_x$ que contiene la característica *actors* es la siguiente: “*The actors are pretty good for the most part*”. Se necesita aplicar la función *ConjuntoPalabras* (*metodo*, *N_GRAM*, *caracteristica*, $opinion_x$) que devuelve el conjunto de *N_GRAM* palabras alrededor de la *caracteristica* dentro de la $opinion_x$. Estas palabras se obtienen aplicando el *metodo* específico de obtención de palabras adyacentes. En este caso los parámetros toman los siguientes valores: *ConjuntoPalabras* (*All_Phase*, 0, *actors*, $opinion_x$), es decir, se aplicará el método *All_Phase* para obtener las palabras cercanas a la característica *actors*. El parámetro *N_GRAM* toma el valor 0 debido a que el método *All_Phase* no necesita utilizarlo. Por lo tanto, *ConjuntoPalabras* (*All_Phase*, 0, *actors*, $opinion_x$) = { *The*, *are*, *pretty*, *good*, *for*, *the*, *most*, *part* }. Teniendo en cuenta lo anterior, los valores de sentimiento en SWN para las palabras cercanas a la característica *actors* son los siguientes:

- La palabra *The* no tiene ningún valor de sentimiento asociado en la base de datos SWN. Por lo tanto, *SentiPos(The)* y *SentiNeg(The)* tienen valor 0.
- La palabra *are* tiene una entrada en la base de datos SWN. Los valores de la entrada son *SentiPos(are)* = 0 y *SentiNeg(are)* = 0.
- La palabra *pretty* tiene tres entradas en la base de datos SWN. Por tanto, calcularemos sus valores medios, tal y como se muestra en las fórmulas 2.3 y 2.4. Los valores finales serían *SentiPos(pretty)* = 0.333 y *SentiNeg(pretty)* = 0.417.
- La palabra *good* tiene veintisiete sentidos en la base de datos SWN con valores medios de *SentiPos(good)* = 0.597 y *SentiNeg* = 0.0050.
- La palabra *most* tiene cinco sentidos en la base de datos SWN con valores medios de *SentiPos(most)* = 0 y *SentiNeg* = 0.
- La palabra *part* tiene dieciocho sentidos en la base de datos SWN con valores medios de *SentiPos(part)* = 0 y *SentiNeg(part)* = 0.

Teniendo en cuenta toda la información anterior con respecto a los valores de sentimiento de las palabras alrededor de la expresión lingüística de la característica *actors*, el proceso de cálculo de la polaridad de la característica dentro de la opinión del usuario vendrá expresado como el vector de polaridad $\vec{V}$, cuya expresión es la siguiente:

$$\vec{V}(actors, opinion_x) = peso(actors, opinion_x) * (SentiPos, SentiNeg, SentiNeu) = peso(actors, opinion_x) * (0.465, 0.211, 0.162)$$

En la expresión anterior $peso(actors, opinion_x)$ representa el peso de la característica *actors* dentro de la $opinion_x$. Como se ha explicado en la sección anterior es un valor positivo. Por ejemplo, si $peso(actors, opinion_x) = 300$ el vector de polaridad asociado a la característica *actors* en la opinión de usuario $opinion_x$ vendrá expresado de la siguiente manera:

$$\vec{V}(actors, opinion_x) = 300 * (0.465, 0.211, 0.162) = (139.5, 63.3, 48.6)$$

Siguiendo la fórmula 2.9 de cálculo de la polaridad de las características: $Polaridad(actors, opinion_x) = 139.5$. Esto significa que $Polaridad(actors, opinion_x) \in P_{Pos}$ debido a que el punto $P = (139.5, 63.3, 48.6)$ es un punto interior de la pirámide geométrica cuyo volumen está compuesto de todos los puntos de destino de los vectores de posición con dirección positiva (P_{Pos}). Por tanto la función *estaContenido* (P, P_{Pos}) devuelve *verdadero* porque el punto P es un punto interior de la pirámide positiva P_{Pos} .

Por último, cabe destacar que en la metodología expuesta en esta tesis doctoral si se encuentra una cláusula negativa en una frase del texto, entonces los valores para *SentiPos (palabra)* y *SentiNeg (palabra)* se intercambian para la palabra de la que estamos analizando su valor de sentimiento dentro de esa frase. Por ejemplo, si aparece en la opinión de un usuario una frase del tipo: “*The interpretation of Denzel Washington does not seem good*”, el sistema detecta una cláusula negativa que afecta a la polaridad de la palabra *good* (con evidente sentido positivo) e invierte los valores de sentimiento *SentiPos (good)* y *SentiNeg (good)* asociados a la palabra *good* en *SWN*. Esto implica que cuando se quiera calcular la polaridad de la característica *Denzel Washington* dentro de la opinión, la palabra cercana *good*, que de no existir una cláusula negativa detectada para la frase aportaría un sentido claramente positivo a la característica, aportará un evidente sentido negativo al invertir los valores de sentimiento *SentiPos (good)* y *SentiNeg (good)*.

2.6.2 Agregación de polaridades de características

Como se ha expuesto anteriormente, el módulo de identificación de características lleva a cabo un proceso de búsqueda de sinónimos para identificar aquellos términos lingüísticos en las opiniones que tienen una relación de sinonimia con objetos de la

ontología del dominio. Estos términos son identificados como características por el sistema. Sin embargo, en este proceso no se tiene en cuenta si el conjunto de características identificado contiene características que son sinónimas entre ellas, es decir, hasta ahora todas las características identificadas por el sistema han sido tratadas de manera independiente. El sistema dispone de un conjunto de características con sus polaridades asociadas y tratará de ver si puede agregar las polaridades de aquellas características que hacen referencia al mismo concepto ontológico. Para ello se ayudará de una base de datos de términos sinónimos. Si trabajamos con opiniones de usuarios escritas en inglés, la base de datos de términos sinónimos empleada será *WordNet* (Miller, 1995).

El sistema buscará entre el conjunto de características identificadas aquellas que sean sinónimas y agregará sus vectores de polaridad asociados. El método de agregación de polaridades de características agrupará los términos lingüísticos que identifican a las características en un único término. El término elegido para dar nombre a todas las características agrupadas será el de uno de ellas, ya que cualquier término hace referencia al mismo concepto en la ontología del dominio. A continuación se calcula el vector de polaridad asociado a este término. Para calcular las coordenadas de este vector se calcularán los valores medios de la suma de cada una de las coordenadas de los vectores de polaridad asociados a las características que hacen referencia al mismo concepto ontológico. En la fórmula 2.5 se muestra la forma en la que se calculan las coordenadas del vector de polaridad asociado a la característica *característica_agrupadora* en la *opinion_x* cuando la característica tiene un número *n* de sinónimos entre las características identificadas por el sistema. El término *característica_agrupadora* representa al nombre de la característica que se convierte en representante de todas las características sinónimas agrupadas. Por otro lado, *f_i* representa a los términos asociados a cada una de las *n* características identificadas como sinónimas.

$$\vec{V}(\text{característica\_agrupadora}, \text{opinion}_x) = \frac{\sum_{i=1}^n \vec{V}_i(f_i, \text{opinion}_x)}{|n|} \quad (2.5)$$

Por ejemplo, si se tienen las características identificadas *movie*, *film* y *picture* con los vectores de polaridad asociados $\vec{V}_1(\text{movie}, \text{opinion}_x) = (60.30, 25.22, 10.35)$, $\vec{V}_2(\text{film}, \text{opinion}_x) = (28.66, 25.23, 18.35)$ y $\vec{V}_3(\text{picture}, \text{opinion}_x) = (14.35, 27.80, 16.43)$ respectivamente, el sistema agregará estas tres características en una sola característica

agrupadora llamada *movie* con vector de polaridad asociado $\vec{V}_4(movie, opinion_x) = (34.43, 26.08, 15.04)$.

2.7 Módulo de análisis de sentimientos de opiniones

El enfoque propuesto en esta tesis doctoral está basado en la creación de un mecanismo para realizar el análisis de sentimientos de opiniones. El módulo de minería de opiniones descrito en esta sección está basado en análisis vectorial y proporciona una clasificación de sentimientos basada en los vectores de polaridad asociados a las características identificadas dentro de las opiniones.

En la figura 2.24 se muestra un esquema gráfico de las entradas y salidas del módulo de análisis de sentimientos de opiniones. En esta figura se aprecia como la entrada de este módulo es un conjunto de características con sus polaridades asociadas para cada opinión. La salida del módulo serán las polaridades globales de cada opinión realizada por los usuarios.

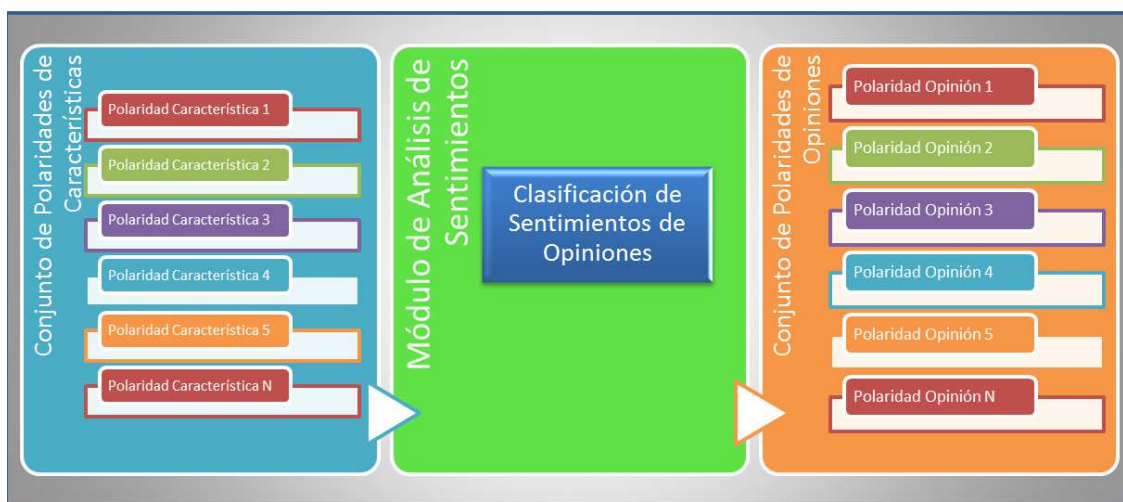


Figura 2.24. Módulo de análisis de sentimientos de opiniones.

Como se ha comentado anteriormente, la polaridad global de la opinión de un usuario vendrá determinada por las polaridades individuales de cada una de las características identificadas en dicha opinión. Además se ha visto que las características identificadas en cada opinión tienen asociado un vector euclídeo de polaridad. En la metodología de esta tesis se calcula el vector de polaridad de la opinión de un usuario como la suma de los vectores de polaridad de todas las características identificadas en dicha opinión. Este

proceso se denomina análisis de sentimientos y viene determinado por el vector de posición resultante de la suma de todos los vectores de polaridad, previamente ponderados, de cada característica identificada en la opinión del usuario. La expresión del vector de polaridad global de una $opinion_x$ se muestra en la fórmula 2.6, donde f_i representa cada una de las características identificadas en la $opinion_x$.

$$\vec{V}(opinion_x) = \sum_{i=1}^n \vec{V}(f_i, opinion_x) \quad (2.6)$$

El vector de polaridad global resultante será un vector euclídeo con tres coordenadas (x , y , z). Sea P el punto de destino del vector de posición resultante de la suma de los vectores de polaridad de las características, expresado como $P = (x, y, z)$. P_{Pos} , P_{Neg} y P_{Neu} son las pirámides geométricas cuyo volumen está compuesto de todos los puntos de destino de los vectores de posición con dirección positiva, negativa y neutra respectivamente. Entonces la polaridad global de la opinión de un usuario viene dada según se muestra en la fórmula 2.7.

$$Polaridad(opinion_x) = \begin{cases} +x \text{ (positiva)} & \text{si } estaContenido(P, P_{Pos}) = verdadero \\ -y \text{ (negativa)} & \text{si } estaContenido(P, P_{Neg}) = verdadero \\ 0 \text{ (neutra)} & \text{si } estaContenido(P, P_{Neu}) = verdadero \end{cases} \quad (2.7)$$

Por ejemplo, si se tiene una opinión de usuario con las siguientes características identificadas: *actors*, *music* y *movie* con los vectores de polaridad asociados: $\vec{V}_1(actors, opinion_x) = (139.5, 63.3, 48.6)$, $\vec{V}_2(music, opinion_x) = (83.4, 132.3, 26.2)$ y $\vec{V}_3(movie, opinion_x) = (36.5, 93.1, 14.2)$ entonces, según la fórmula 2.6, se obtiene: $\vec{V}(opinion_x) = \vec{V}_1 + \vec{V}_2 + \vec{V}_3 = (259.4, 288.7, 89)$. Aplicando la fórmula 2.7, $Polaridad(opinion_x) = -288.7$ ya que el punto $P = (259.4, 288.7, 89)$ es un punto interior de la pirámide geométrica cuyo volumen está compuesto de todos los puntos de destino de los vectores de posición con dirección negativa (P_{Neg}). Por lo que la función $estaContenido(P, P_{Neg})$ devuelve *verdadero*. Esto significa que la polaridad global de la opinión de usuario $opinion_x$ será negativa.

2.8 Diagrama general de procesos

Una vez analizados los módulos principales que componen la arquitectura del sistema y explicado el funcionamiento específico de cada uno de ellos, en esta sección se describe el funcionamiento conjunto del sistema mediante un diagrama de procesos.

Para mostrar un diagrama general de procesos que resuma gráficamente el funcionamiento de nuestra metodología de análisis de sentimientos se necesitan definir algunas notaciones gráficas. En la tabla 2.3 se muestran las convenciones gráficas que se van a utilizar para describir de manera robusta y homogénea el sistema desde el punto de vista de los procesos que lo componen.

Tabla 2.3. Definición de los elementos empleados en los diagramas de procesos del sistema.

Elemento	Descripción
 Elemento de Entrada	Se emplea para representar el elemento de entrada que es requerido por los procesos para poder llevar a cabo su labor. Se trata de un elemento exterior al sistema, pero necesario para su funcionamiento.
 Elemento Interno	Se emplea para representar los datos de entrada y salida que se intercambian internamente todos los procesos del sistema.
 Elemento de Salida	Se emplea para representar el elemento de salida del sistema. Se trata de un elemento que se exporta al exterior del sistema.
 Relaciones Básicas	Representan una relación entre un elemento de entrada o salida y un proceso del sistema. Su sentido define el flujo de los datos.
 Proceso del Sistema	Se emplea para representar la ejecución de un proceso del sistema. En su interior aparecerá el nombre del proceso.

La figura 2.25 muestra un diagrama general de todos los procesos que se llevan a cabo en el sistema para analizar los sentimientos de las opiniones de los usuarios. Un proceso especifica la tarea o conjunto de tareas que lleva a cabo el sistema y puede que necesite algún elemento de entrada para llevar a cabo su trabajo (base de datos léxica, ontología, etc.). Una vez finalizada su tarea proporciona uno o varios elementos de salida.

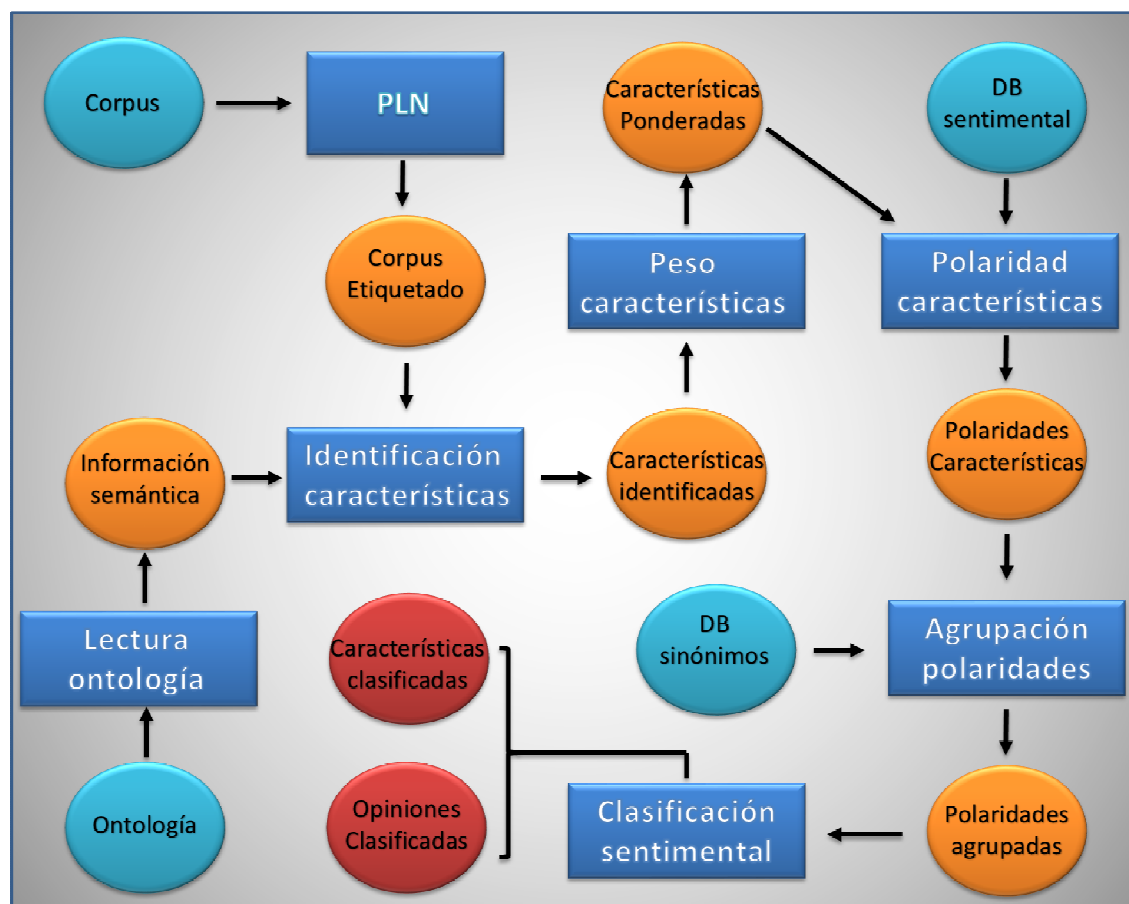


Figura 2.25. Diagrama general de procesos del sistema.

En el diagrama de procesos mostrado en la figura 2.25 inicialmente se requieren los elementos de entrada siguientes: un corpus con las opiniones de los usuarios expresadas en lenguaje natural y una ontología del dominio. El corpus es analizado por el proceso de *PLN* del sistema, que devuelve el corpus etiquetado morfológicamente.

Seguidamente un proceso del sistema procesa la ontología del dominio y devuelve la estructura semántica de la misma una vez que la ha procesado. En este punto, se han obtenido dos elementos de salida listos para poder utilizarse: un corpus etiquetado y la información semántica de la ontología del dominio.

El proceso de identificación de características utiliza estos dos elementos para llevar a cabo un proceso de búsqueda de las características relevantes presentes en las opiniones de los usuarios. Una vez que este proceso finaliza su trabajo devuelve una lista de características identificadas.

Partiendo del conjunto de características identificadas, el proceso de asignación de pesos de características devuelve este conjunto de características convenientemente ponderadas en función de sus posiciones relativas dentro de las opiniones de los usuarios.

Una vez que el sistema dispone del conjunto de características identificadas y ponderadas está listo para calcular las polaridades de estas características. Para realizar este trabajo, el proceso de cálculo de polaridades de características se ayudará de un elemento externo: una base de datos de sentimientos. Con estos elementos el proceso se ejecuta y devuelve un listado de polaridades de las características identificadas dentro de las opiniones de los usuarios.

En este conjunto de características con sus polaridades asignadas pueden existir características que tengan expresiones lingüísticas distintas pero que conceptualmente signifiquen lo mismo. Es decir, pueden aparecer características que son sinónimas. Un proceso del sistema llevará a cabo la labor de agrupar las polaridades de estas características equivalentes.

Con las características identificadas y sus polaridades calculadas, se ejecuta el proceso de análisis de sentimientos de opiniones, que utiliza los vectores de polaridad del conjunto de características identificadas para calcular la polaridad global de las opiniones de los usuarios.

2.9 Ejemplo de funcionamiento del sistema

En esta sección se muestra un ejemplo completo del funcionamiento del sistema. El objetivo de esta sección es explicar paso a paso, mediante un ejemplo completo, cada uno de los módulos de la arquitectura del sistema en los que se basa la metodología propuesta en esta tesis doctoral.

Como se ha comentado anteriormente, la metodología es independiente del dominio e idioma sobre los que se aplique. Para conseguir este objetivo ha sido diseñada de

manera modular y no tiene ningún tipo de ‘acoplamiento’ o dependencia de ninguna herramienta o utilidad externa. Sin embargo, aunque se ha descrito la metodología de manera genérica, para aplicarla sobre un dominio e idioma concretos se necesita del conjunto de utilidades y herramientas externas específicas que se citan a continuación:

- Un **corpus** con las opiniones de los usuarios escritas en el idioma concreto.
- Una **herramienta de procesamiento del lenguaje natural** que trabaje en el idioma específico y etiquete morfológicamente el corpus de opiniones.
- Una **ontología del dominio** con los *labels* de los conceptos traducidos al idioma objeto de estudio.
- Una **base de datos léxica** de sentimientos en el idioma concreto.
- Una **base de datos de sinónimos** en el idioma concreto.

Como se ha comentado antes, en este apartado se muestra el funcionamiento del sistema paso a paso. Para ello se ha elegido el dominio de las películas cinematográficas y el idioma inglés. La arquitectura general del sistema se configura de manera específica con las siguientes herramientas externas y utilidades (ver figura 2.26):

- El corpus con las **opiniones de los usuarios sobre películas** escritas en **inglés**.
- La herramienta de procesamiento del lenguaje natural **Stanford Log-linear Part-Of-Speech Tagger**, que soporta el inglés entre otros muchos idiomas.
- La ontología del dominio **The Movie Ontology** que contiene los *labels* de los conceptos traducidos al inglés.
- La base de datos léxica de sentimientos **SentiWordNet**.
- La base de datos de sinónimos **WordNet**.

La figura 2.26 muestra la arquitectura general del sistema de la figura 2.2 con las herramientas y utilidades externas configuradas apropiadamente para poder trabajar sobre el dominio de las películas cinematográficas e idioma inglés. Aparecen resaltadas en color rosa las partes que son ‘externas’ y ‘configurables’ en la arquitectura general del sistema que sustenta a la metodología propuesta en esta tesis doctoral.

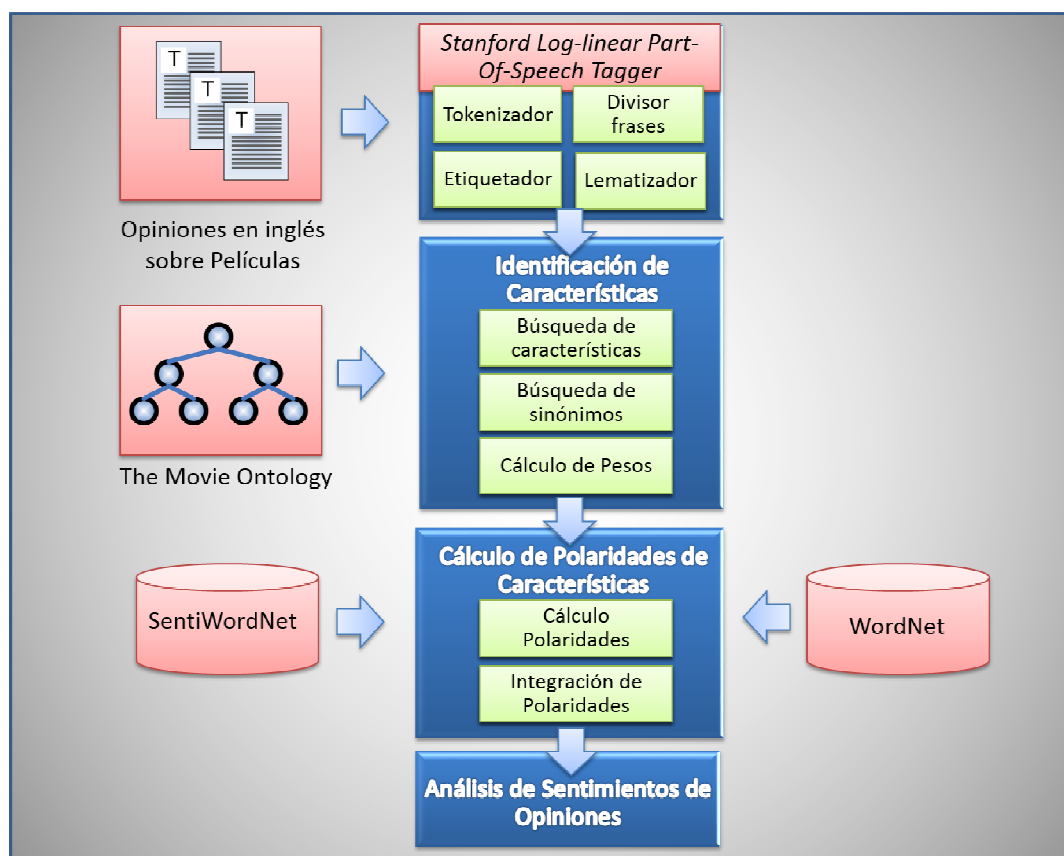


Figura 2.26. Arquitectura específica del sistema para un dominio e idioma concretos con las herramientas externas necesarias.

2.9.1 Estructura de la ontología del dominio y proceso de identificación de términos

La metodología propuesta en esta tesis necesita de una ontología del dominio correctamente estructurada e instanciada para obtener resultados óptimos. En este apartado se explica la estructura que debe tener la ontología del dominio y el mecanismo de detección de términos en las opiniones. Posteriormente, algunos de estos términos serán las características identificadas de las que habrá que calcular su polaridad, tal y como se ha explicado en los apartados anteriores.

La metodología propuesta en esta tesis doctoral utiliza una ontología del dominio definida en lenguaje *OWL* y no incluye ninguna restricción en la forma u orden de definición de los elementos típicos que componen la ontología. La estructura de una ontología del dominio típicamente utilizada por el sistema es la siguiente: inicialmente se definen los espacios de nombres por defecto de la ontología que se está describiendo, ya que esto sentará las bases del vocabulario *OWL* con el que se va a trabajar. Seguidamente se definen una serie de encabezamientos con información relativa a

comentarios, versión de la ontología, autor, etc. A continuación se definen los siguientes elementos básicos de *OWL*: declaración de clases simples, subclases, individuos de una clase y propiedades. Las propiedades se clasifican en dos tipos de relaciones: las relaciones entre tipos de datos y las relaciones entre individuos de clases. A las primeras se les denomina *datatype properties* y a las segundas *object properties*.

Con la definición de estos elementos el sistema tiene lo necesario para identificar las características presentes en las opiniones de los usuarios. Tras llevar a cabo el proceso de *PLN* en el corpus de opiniones, el sistema obtiene una serie de términos lingüísticos candidatos a ser características. Para ello buscará las correspondencias directas entre este conjunto de términos lingüísticos y el conjunto de *labels* asociados a cada una de las clases, individuos y propiedades de la ontología del dominio. En primer lugar se buscan correspondencias directas, es decir, aquellas correspondencias en las que el término lingüístico que aparece en la opinión del usuario es idéntico al término expresado en uno de los *labels* de la ontología del dominio.

Para ilustrar este punto se toma como ejemplo la definición de clase que aparece en la figura 2.27 para una ontología del dominio de las finanzas. La definición está hecha en lenguaje *OWL*. El nombre de la clase, a nivel conceptual, será *Empresa*. Seguidamente se definen el conjunto de *labels* que identificarán a este concepto en los idiomas inglés y español. En concreto, este concepto quedará identificado por los *labels* “Enterprise” y “Company” para el idioma inglés y por los *labels* “Empresa” y “Compañía” para el idioma español. Además se define que la clase *Empresa* será una subclase de la clase *Tipos_de_activos_para_invertir*. Por último se definen un par de comentarios en inglés y español para explicar la clase *Empresa*.

```
<owl:Class rdf:about="#Empresa">
  <rdfs:label xml:lang="en">Enterprise</rdfs:label>
  <rdfs:label xml:lang="en">Company</rdfs:label>
  <rdfs:label xml:lang="es">Empresa</rdfs:label>
  <rdfs:label xml:lang="es">Compañía</rdfs:label>
  <rdfs:subClassOf rdf:resource="#Tipos_de_activos_para_invertir"/>
  <rdfs:comment xml:lang="en">Organizational unit dedicated to
industrial, commercial or provide services for profit.
</rdfs:comment>
  <rdfs:comment xml:lang="es">Unidad de organización dedicada a
actividades industriales, mercantiles o de prestación de servicios con
fines lucrativos.
</rdfs:comment>
</owl:Class>
```

Figura 2.27. Ejemplo de definición en *OWL* de una clase en una ontología del dominio de las finanzas.

La figura 2.27 se toma como referencia para mostrar la manera en la que el sistema identifica las características presentes en las opiniones de los usuarios. De esta manera, cuando el sistema, por ejemplo, se encuentra la siguiente frase en la opinión de un usuario: “*This company does not provide a good service*” comprobará que existe una correspondencia directa entre el término lingüístico de la opinión *company* y uno de los *labels* definidos para la clase *Empresa* de la ontología del dominio. Por tanto, se identificará el término lingüístico *company* dentro de la opinión como una característica. Por otro lado, si el sistema se encuentra, por ejemplo, una frase como la siguiente: “*this firm treats its customers very well!*”, el sistema ve que, en primera instancia, no existe ninguna correspondencia directa entre el término lingüístico *firm* y ninguna de las *labels* definidas para la clase *Empresa* de la ontología del dominio. Sin embargo, en un proceso posterior llamado ‘agregación de polaridades de características’ (ver sección 2.6.2) se utilizará una base de datos de sinónimos (*WordNet*, si estamos trabajando con opiniones en inglés) que identificará al término lingüístico *firm* como una característica dentro de la opinión, ya que comprobará que este término es sinónimo de los *labels* *Enterprise* y *Company* definidos para el idioma inglés dentro de la clase *Empresa*. Por lo tanto se puede afirmar que, en la metodología propuesta en esta tesis, la base de datos de sinónimos utilizada por el sistema ‘complementa’ el trabajo realizado por la ontología del dominio en el proceso de identificación de características.

2.9.2 Configuración inicial del sistema

Para funcionar correctamente el sistema necesita que se le faciliten las siguientes configuraciones:

- **Ficheros con las opiniones.** Hay que indicarle al sistema cuáles son los ficheros de texto que contienen las opiniones expresadas en lenguaje natural. El conjunto de estos ficheros forma el corpus del sistema.
- **Base de datos que contenga un léxico de sentimientos.** Hay que indicarle al sistema dónde se encuentra la base de datos con los valores de sentimiento positivos y negativos que se asocian a las palabras que expresan alguna emoción.
- **Base de datos de sinónimos.** Hay que indicarle al sistema dónde se encuentra la base de datos de sinónimos para que pueda identificar como el mismo concepto aquellos términos lingüísticos que son equivalentes.

- **Ontología del dominio.** Hay que indicarle al sistema dónde se encuentra la ontología del dominio. Esta ontología deberá estar correctamente estructurada e instanciada.
- **Método de clasificación de sentimientos.** Como se ha visto a lo largo de este capítulo, el sistema dispone de distintos métodos de cálculo de la polaridad de características y opiniones de usuarios. Estos métodos son *N_GRAM Before*, *N_GRAM After*, *N_GRAM Around* y *All Phrase*. Hay que indicarle al sistema el método de clasificación que se va a utilizar.
- **Valores de los parámetros Z_1 , Z_2 , Z_3 y N_GRAM .** Hay que indicarle al sistema los pesos que asignaremos a las características identificadas en las opiniones de los usuarios en función de su posición relativa dentro de dichas opiniones (parámetros Z_1 , Z_2 y Z_3). Además hay que indicarle el número de palabras cercanas a la expresión lingüística que identifica a la característica que se tendrán en cuenta a la hora de calcular la polaridad de la misma (parámetro N_GRAM).

2.9.3 Análisis morfológico del corpus de opiniones

La primera fase de este proceso consiste en la utilización de herramientas de procesamiento del lenguaje natural para realizar un procesado léxico y morfológico del texto de entrada. Estas herramientas analizan el texto y obtienen información lingüística. Se comienza por separar el texto en palabras e identificar cada una de las frases, también se identifica a los nombres propios y se obtiene la forma normal de cada palabra. Por último, se identifica la categoría gramatical de cada palabra, proporcionando información morfológica.

Para ilustrar este proceso en este apartado se va a partir de una opinión real de un usuario expresada en inglés (ver tabla 2.4). Se ha elegido una opinión de forma aleatoria del corpus de opiniones en el dominio de las películas cinematográficas. Se trata de una opinión que se encuentra en un fichero de texto dentro del directorio *NEG* del corpus, lo que indica que es una opinión etiquetada manualmente por un experto como negativa. En este trabajo no sólo se quiere calcular la polaridad global de la opinión de los usuarios, sino que, además, se quiere calcular la polaridad de cada una de las características que aparecen en la opinión.

Tabla 2.4. Opinión de un usuario en lenguaje natural.

plot : two teen couples go to a church party , drink and then drive .
they get into an accident .
one of the guys dies , but his girlfriend continues to see him in her life ,
and has nightmares .
what's the deal ?
watch the movie and " sorta " find out . . .
critique : a mind-fuck movie for the teen generation that touches on a very
cool idea , but presents it in a very bad package .
which is what makes this review an even harder one to write , since i
generally applaud films which attempt to break the mold , mess with your head
and such (lost highway & memento) , but there are good and bad ways of
making all types of films , and these folks just didn't snag this one
correctly .
they seem to have taken this pretty neat concept , but executed it terribly .
so what are the problems with the movie ?
well , its main problem is that it's simply too jumbled .
it starts off " normal " but then downshifts into this " fantasy " world in
which you , as an audience member , have no idea what's going on .
there are dreams , there are characters coming back from the dead , there are
others who look like the dead , there are strange apparitions , there are
disappearances , there are a looooot of chase scenes , there are tons of
weird things that happen , and most of it is simply not explained .
now i personally don't mind trying to unravel a film every now and then , but
when all it does is give me the same clue over and over again , i get kind of
fed up after a while , which is this film's biggest problem .
it's obviously got this big secret to hide , but it seems to want to hide it
completely until its final five minutes .
and do they make things entertaining , thrilling or even engaging , in the
meantime ?
not really .
the sad part is that the arrow and i both dig on flicks like this , so we
actually figured most of it out by the half-way point , so all of the
strangeness after that did start to make a little bit of sense , but it still
didn't the make the film all that more entertaining .
i guess the bottom line with movies like this is that you should always make
sure that the audience is " into it " even before they are given the secret
password to enter your world of understanding .
i mean , showing melissa sagemiller running away from visions for about 20
minutes throughout the movie is just plain lazy ! !
okay , we get it . . . there
are people chasing her and we don't know who they are .
do we really need to see it over and over again ?
how about giving us different scenes offering further insight into all of the
strangeness going down in the movie ?
apparently , the studio took this film away from its director and chopped it
up themselves , and it shows .
there might've been a pretty decent teen mind-fuck movie in here somewhere ,
but i guess " the suits " decided that turning it into a Music video with
little edge , would make more sense .
the actors are pretty good for the most part , although wes bentley just
seemed to be playing the exact same character that he did in american beauty
 , only in a new neighborhood .
but my biggest kudos go out to sagemiller , who holds her own throughout the
entire film , and actually has you feeling her character's unraveling .

overall , the film doesn't stick because it doesn't entertain , it's confusing , it rarely excites and it feels pretty redundant for most of its runtime , despite a pretty cool ending and explanation to all of the craziness that came before it .
 oh , and by the way , this is not a horror or teen slasher flick . . . it's just packaged to look that way because someone is apparently assuming that the genre is still hot with the kids .
 it also wrapped production two years ago and has been sitting on the shelves ever since .
 whatever . . . skip
 it !
 where's joblo coming from ?
 a nightmare of elm street 3 (7/10) - blair witch 2 (7/10) - the crow (9/10) - the crow : salvation (4/10) - lost highway (10/10) - memento (10/10) - the others (9/10) - stir of echoes (8/10)

El resultado que se obtiene después del procesamiento del lenguaje natural es el del texto totalmente etiquetado, en el que se han identificado cada una de las categorías gramaticales y significado morfológico asociados a los términos lingüísticos que aparecen en las opiniones de los usuarios. Como se ha comentado, la herramienta que se ha utilizado para ello es *Stanford Log-linear Part-Of-Speech Tagger* y el resultado se puede observar en la tabla 2.5.

Tabla 2.5. Opinión etiquetada de un usuario después de analizarla.

NN plot : : CD two JJ teen NNS couples VBP go TO to DT a NN church NN party , , NN
 drink CC and RB then VB drive . .
 PRP they VBP get IN into DT an NN accident . .
 CD one IN of DT the NNS guys VBZ dies , , CC but PRP\$ his NN girlfriend VBZ continues
 TO to VB see PRP him IN in PRP\$ her NN life , , CC and VBZ has NNS nightmares . .
 WP what VBZ 's DT the NN deal . ?
 VB watch DT the NN movie CC and `` " NN sorta " " VBP find RP out
 NN critique : : DT a JJ mind-fuck NN movie IN for DT the JJ teen NN generation IN that
 NNS touches IN on DT a RB very JJ cool NN idea , , CC but VBZ presents PRP it IN in
 DT a RB very JJ bad NN package . .
 WDT which VBZ is WP what VBZ makes DT this NN review DT an RB even JJR harder
 CD one TO to VB write , , IN since FW i RB generally VBP applaud NNS films WDT which
 VBP attempt TO to VB break DT the NN mold , , NN mess IN with PRP\$ your NN head
 CC and JJ such -LRB- (NN lost_highway CC & NNS memento -RRB-) , , CC but EX there
 VBP are JJ good CC and JJ bad NNS ways IN of VBG making DT all NNS types IN of
 NNS films , , CC and DT these NNS folks RB just VBD did RB n't NN snag DT this CD
 one RB correctly . .
 PRP they VBP seem TO to VB have VBN taken DT this RB pretty JJ neat NN concept , ,
 CC but VBD executed PRP it RB terribly . .
 IN so WP what VBP are DT the NNS problems IN with DT the NN movie . ?
 UH well , , PRP\$ its JJ main NN problem VBZ is IN that PRP it VBZ 's RB simply RB too
 JJ jumbled . .
 PRP it VBZ starts RP off `` " JJ normal " " CC but RB then VBZ downshifts IN into DT
 this `` " NN fantasy " " NN world IN in WDT which PRP you , , IN as DT an NN
 audience NN member , , VBP have DT no NN idea WP what VBZ 's VBG going IN on . .
 EX there VBP are NNS dreams , , EX there VBP are NNS characters VBG coming RB back
 IN from DT the NN dead , , EX there VBP are NNS others WP who VBP look IN like DT
 the NN dead , , EX there VBP are JJ strange NNS apparitions , , EX there VBP are NNS
 disappearances , , EX there VBP are DT a NN looooot IN of NN chase NNS scenes , , EX
 there VBP are NNS tons IN of JJ weird NNS things WDT that VBP happen , , CC and JJS

most IN of PRP it VBZ is RB simply RB not VBD explained . .
 RB now FW i RB personally VBP do RB n't VB mind VBG trying TO to VB unravel DT a
 NN film DT every RB now CC and RB then , , CC but WRB when DT all PRPit VBZ does
 VBZ is VB give PRP me DT the JJ same NN clue IN over CC and INover RB again , ,
 FW i VBP get NN kind IN of VBN fed RP up IN after DT a NNwhile , , WDT which VBZ
 is DT this NN film POS 's JJS biggest NN problem . .
 PRP it VBZ 's RB obviously VBD got DT this JJ big NN secret TO to VB hide , , CCbut
 PRP it VBZ seems TO to VB want TO to VB hide PRP it RB completely IN until PRP\$ its
 JJ final CD five NNS minutes . .
 CC and VBP do PRP they VBP make NNS things JJ entertaining , , JJ thrilling CCor RB
 even VBG engaging , , IN in DT the NN meantime . ?
 RB not RB really . .
 DT the JJ sad NN part VBZ is IN that DT the NN arrow CC and FW i DT both NNdig IN
 on NNS flicks IN like DT this , , IN so PRP we RB actually VBD figured JJSmost IN of
 PRP it IN out IN by DT the JJ half-way NN point , , RB so DT all INof DT the NN
 strangeness IN after DT that VBD did VB start TO to VB make DT a JJ little NN bit IN of
 NN sense , , CC but PRP it RB still VBD did RB n't DT the VBP make DT the NN film DT
 all IN that RBR more JJ entertaining . .
 FW i VB guess DT the JJ bottom NN line IN with NNS movies IN like DT this VBZis IN
 that PRP you MD should RB always VB make JJ sure IN that DT the NNaudience VBZ is
 `` " IN into PRP it " " RB even IN before PRP they VBP are VBN given DT the NN secret
 NN password TO to VB enter PRP\$ your NN world INof VBG understanding . .
 FW i VB mean , , VBG showing NN melissa_sagemiller VBG running RB away INfrom
 NNS visions IN for IN about CD 20 NNS minutes IN throughout DT the NNmovie VBZ is
 RB just RB plain JJ lazy . ! . !
 JJ okay , , PRP we VBP get PRP it EX there
 VBP are NNS people VBG chasing PRP her CC and PRP we VBP do RB n't VB know WP
 who PRP they VBP are . .
 VBP do PRP we RB really VBP need TO to VB see PRP it IN over CC and IN over RB
 again . ?
 WRB how RB about VBG giving PRP us JJ different NNS scenes VBG offering RBRfurther
 VB insight IN into DT all IN of DT the NN strangeness VBG going RP down IN in DT the
 NN movie . ?
 RB apparently , , DT the NN studio VBD took DT this NN film RB away IN from PRP\$ its
 NN director CC and VBN chopped PRP it RP up PRP themselves , , CCand PRP it VBZ
 shows . .
 EX there MD might VB 've VBN been DT a RB pretty JJ decent NN teen JJ mind-fuck NN
 movie IN in RB here RB somewhere , , CC but FW i VBP guess `` " DTthe NNS suits "
 " VBN decided IN that VBG turning PRP it IN into DT a NN music NN video IN with JJ
 little NN edge , , MD would VB make JJR more NN sense . .
 DT the NNS actors VBP are RB pretty JJ good IN for DT the JJS most NN part , , IN
 although NN wes_bentley RB just VBD seemed TO to VB be VBG playing DTthe JJ exact
 JJ same NN character IN that PRP he VBD did IN in NNAmerican_beauty , , RB only IN
 in DT a JJ new NN neighborhood . .
 CC but PRP\$ my JJS biggest NNS kudos VBP go RP out TO to JJR sagemiller , , WP who
 VBZ holds PRP her JJ own IN throughout DT the JJ entire NN film , , CCand RB actually
 VBZ has PRP you VBG feeling PRP\$ her NN character POS 's NNunraveling . .
 JJ overall , , DT the NN film VBZ does RB n't VB stick IN because PRP it VBZdoes RB n't
 VB entertain , , PRP it VBZ 's JJ confusing , , PRP it RB rarely VBZexcites CC and PRP it
 VBZ feels RB pretty JJ redundant IN for JJS most IN of PRP\$ its NN runtime , , IN
 despite DT a RB pretty JJ cool NN ending CC and NNexplanation TO to DT all IN of DT
 the NN craziness WDT that VBD came IN before PRP it . .
 UH oh , , CC and IN by DT the NN way , , DT this VBZ is RB not DT a NNhorror CC or
 JJ teen JJR slasher RB flick PRP it VBZ 's
 RB just VBN packaged TO to VB look DT that NN way IN because NN someone VBZ is
 RB apparently VBG assuming IN that DT the NN genre VBZ is RB still JJhot IN with DT
 the NNS kids . .
 PRP it RB also VBD wrapped NN production CD two NNS years RB ago CC and VBZ has
 VBN been VBG sitting IN on DT the NNS shelves RB ever IN since . .
 WDT whatever VB skip
 PRP it . !

```
WRB where POS 's NN joblo VBG coming IN from . ?
NN a_nightmare_of_elm_street_3 -LRB- ( NN 7/10 -RRB- ) : - NN blair_witch_2 -LRB- (
NN 7/10 -RRB- ) : - NN the_crow -LRB- ( NN 9/10 -RRB- ) : - NN the_crow_ :: NN
_salvation -LRB- ( NN 4/10 -RRB- ) : - NN lost_highway -LRB- ( NN 10/10 -RRB- ) : -
NN memento -LRB- ( NN 10/10 -RRB- ) : - NNS the_others -LRB- ( NN 9/10 -RRB- ) : -
NNS stir_of_echoes -LRB- ( NN 8/10 -RRB- )
```

A partir del resultado del análisis morfológico de la opinión del usuario, el sistema elige una serie de términos lingüísticos como candidatos a ser características. Los resultados que aporta el sistema en esta fase se pueden ver en la tabla 2.6. En este caso, el sistema ha identificado 125 términos lingüísticos en la opinión del usuario que son candidatos a características.

Tabla 2.6. Términos candidatos a características identificados por el sistema.

```
1) NLP PROCESS
=====

->      Results:
[|plot||couples||church||party||drink||drive||accident||guys||girlfriend||lif
e||nightmares||deal||movie||sorta||critique||movie||generation||touches||idea
||package||review||films||mold||mess||head||lost_highway||memento||ways||type
s||films||folks||snag||concept||problems||movie||problem||fantasy||world||aud
ience||member||idea||dreams||characters||others||apparitions||disappearances|
|loooooo||scenes||tons||things||film||clue||kind||while||film||problem||secre
t||minutes||things||meantime||part||arrow||dig||flicks||half-
way||point||strangeness||bit||sense||film||line||movies||audience||password||
world||understanding||mean||visions||minutes||movie||people||scenes||insight|
|strangeness||movie||studio||film||director||movie||guess||suits||music||vide
o||edge||sense||actors||part||character||neighborhood||kudos||sagemiller||fil
m||character||unraveling||film||runtime||explanation||craziness||way||horror|
|slasher||flick||way||someone||genre||kids||production||years||shelves||joblo
||salvation||lost_highway||memento||the_others||stir_of_echoes|]
```

2.9.4 Identificación de características

Una vez que se ha llevado a cabo el análisis morfológico de la opinión y se han identificado los términos que son candidatos a convertirse en características, se necesita que el sistema filtre entre estos términos lingüísticos y descubra cuáles de ellos son verdaderas características relevantes en el dominio en el que se esté trabajando, que en este ejemplo es el dominio de las películas cinematográficas.

El sistema analizará la estructura de la ontología del dominio y extraerá toda la información que necesita para la fase de identificación de características. Básicamente esta información son clases, instancias de clases, relaciones y atributos.

Una vez que el sistema ha procesado el corpus de opiniones y ha identificado el listado de términos lingüísticos candidatos a ser características, se apoyará en la información recogida desde la ontología del dominio para seleccionar aquellos términos candidatos que finalmente se convierten en características.

En la tabla 2.7 aparece el listado de características identificadas por el sistema partiendo de los términos candidatos de la tabla 2.6 y utilizando la información semántica de la ontología. El sistema ha identificado 28 características en la opinión del usuario de entre los 125 términos candidatos de los que se partía, además aporta información acerca del motivo por el que ha seleccionado cada una de las características. Por ejemplo, se añade la característica *movie* porque el sistema la identifica como una clase en la ontología. Por el mismo motivo añade la característica *wes bentley* porque es una instancia en la ontología.

En esta fase no solamente se identifican características, entre los términos candidatos, que tengan una correspondencia directa con los términos de la ontología del dominio de manera literal, sino que se lleva a cabo un proceso de tratamiento de sinónimos por parte del sistema utilizando una base de datos de sinónimos y el conjunto de *labels* asociados a cada término de la ontología. Por este motivo el sistema identifica la característica *films*, que es un término lingüístico que aparece identificado anteriormente en el proceso de *PLN* y es candidato a ser una característica del sistema. El término *film* (que es el lema de *films*) aparece de forma literal en el conjunto de *labels* asociados a la clase *Movie* de la ontología del dominio para el idioma inglés, pero si no apareciese, el sistema utiliza una base de datos de sinónimos para tratar de complementar la información de la ontología. Este es el caso de la característica *years* cuya expresión lingüística es sinónima de la clase de la ontología *age*, pero en este caso no pertenecía a su conjunto de *labels*.

Tabla 2.7. Resultados de las características identificados por el sistema.

2) FEATURES SEARCH
=====
*Add feature: 'movie' (ONTOLOGY CLASS)
*Add feature: 'films' (ONTOLOGY CLASS) -- synonymous of movie
*Add feature: 'lost_highway' (Proper Noun - ONTOLOGY INSTANCE)
*Add feature: 'memento' (Proper Noun - ONTOLOGY INSTANCE)
*Add feature: 'fantasy' (ONTOLOGY INSTANCE)
*Add feature: 'chase' (ONTOLOGY INSTANCE)
*Add feature: 'things' (ONTOLOGY CLASS)
*Add feature: 'film' (ONTOLOGY CLASS) -- synonymous of movie
*Add feature: 'thrilling' (ONTOLOGY CLASS)

```

*Add feature: 'melissa_sagemiller' (Proper Noun - ONTOLOGY INSTANCE)
*Add feature: 'music' (ONTOLOGY INSTANCE)
*Add feature: 'edge' (ONTOLOGY INSTANCE)
*Add feature: 'actors' (ONTOLOGY CLASS)
*Add feature: 'wes_bentley' (Proper Noun - ONTOLOGY INSTANCE)
*Add feature: 'american_beauty' (Proper Noun - ONTOLOGY INSTANCE)
*Add feature: 'sagemiller' (Proper Noun - ONTOLOGY INSTANCE)
*Add feature: 'runtime' (ONTOLOGY PROPERTY)
*Add feature: 'horror' (ONTOLOGY INSTANCE)
*Add feature: 'someone' (ONTOLOGY CLASS) -- synonymous of person
*Add feature: 'genre' (ONTOLOGY CLASS)
*Add feature: 'kids' (ONTOLOGY CLASS)
*Add feature: 'years' (ONTOLOGY CLASS) -- synonymous of age
*Add feature: 'a_nightmare_of_elm_street_3' (Proper Noun - ONTOLOGY
INSTANCE)
*Add feature: 'blair_witch_2' (Proper Noun - ONTOLOGY INSTANCE)
*Add feature: 'the_crow' (Proper Noun - ONTOLOGY INSTANCE)
*Add feature: 'the_crow:_salvation' (Proper Noun - ONTOLOGY
INSTANCE)
*Add feature: 'the_others' (Proper Noun - ONTOLOGY INSTANCE)
*Add feature: 'stir_of_echoes' (Proper Noun - ONTOLOGY INSTANCE)

->      Results:
[|movie||films||lost_highway||memento||fantasy||chase||things||film||thrillin
g||melissa_sagemiller||music||edge||actors||wes_bentley||american_beauty||sag
emiller||runtime||horror||someone||genre||kids||years||a_nightmare_of_elm_str
eet_3||blair_witch_2||the_crow||the_crow:_salvation||the_others||stir_of_ech
oes|]

```

2.9.4.1 Cálculo del peso de las características identificadas

Una vez que el sistema ha identificado las características de la opinión, a continuación calcula el peso asociado a cada una de ellas. Como se ha explicado anteriormente, el sistema asigna un peso distinto a cada característica en función de la parte de la opinión en la que aparezca. Si hay más de una aparición para una misma característica, entonces se pondera de manera independiente cada aparición y al final se hace el cálculo global del peso asignado a esa característica en función de su número total de apariciones.

La tabla 2.8 muestra el peso que se ha asignado a cada una de las características identificadas en el apartado anterior. En esta tabla los parámetros del sistema Z_1 , Z_2 y Z_3 tienen valores 100, 25 y 100 respetivamente. Esto significa que cada aparición de una característica en la parte inicial de la opinión se ponderará con un peso de valor igual a 100, si la aparición es en la parte media se ponderará con 25, y si aparece en la parte final de la opinión se pondera con 100.

Por ejemplo, en el cálculo del peso global asociado a la primera característica el sistema asigna un peso de 450 a la característica *movie*. Esto es así porque en la tabla 2.4, que

recoge la opinión en lenguaje natural del usuario, se muestra que la característica *movie* aparece 3 veces al principio de la opinión, 2 veces en la parte media y una vez al final de la opinión del usuario.

Tabla 2.8. Peso asociado a cada una de las características identificadas.

3) SCORE EVALUATION OF IDENTIFIED FEATURES
=====
-> Parameters: Z1=100, Z2=25, Z3=100
*feature: 'movie' Score: $(100*3 + 25*2 + 100*1) = 450$
*feature: 'films' Score: $(100*2 + 25*0 + 100*0) = 200$
*feature: 'lost_highway' Score: $(100*1 + 25*0 + 100*1) = 200$
*feature: 'memento' Score: $(100*1 + 25*0 + 100*1) = 200$
*feature: 'fantasy' Score: $(100*1 + 25*0 + 100*0) = 100$
*feature: 'chase' Score: $(100*1 + 25*0 + 100*0) = 100$
*feature: 'things' Score: $(100*1 + 25*1 + 100*0) = 125$
*feature: 'film' Score: $(100*0 + 25*3 + 100*3) = 375$
*feature: 'thrilling' Score: $(100*0 + 25*1 + 100*0) = 25$
*feature: 'melissa_sagemiller' Score: $(100*0 + 25*1 + 100*0) = 25$
*feature: 'music' Score: $(100*0 + 25*0 + 100*1) = 100$
*feature: 'edge' Score: $(100*0 + 25*0 + 100*1) = 100$
*feature: 'actors' Score: $(100*0 + 25*0 + 100*1) = 100$
*feature: 'wes_bentley' Score: $(100*0 + 25*0 + 100*1) = 100$
*feature: 'american_beauty' Score: $(100*0 + 25*0 + 100*1) = 100$
*feature: 'sagemiller' Score: $(100*0 + 25*0 + 100*1) = 100$
*feature: 'runtime' Score: $(100*0 + 25*0 + 100*1) = 100$
*feature: 'horror' Score: $(100*0 + 25*0 + 100*1) = 100$
*feature: 'someone' Score: $(100*0 + 25*0 + 100*1) = 100$
*feature: 'genre' Score: $(100*0 + 25*0 + 100*1) = 100$
*feature: 'kids' Score: $(100*0 + 25*0 + 100*1) = 100$
*feature: 'years' Score: $(100*0 + 25*0 + 100*1) = 100$
*feature: 'a_nightmare_of_elm_street_3' Score: $(100*0 + 25*0 + 100*1)$
= 100
*feature: 'blair_witch_2' Score: $(100*0 + 25*0 + 100*1) = 100$
*feature: 'the_crow' Score: $(100*0 + 25*0 + 100*1) = 100$
*feature: 'the_crow_salvation' Score: $(100*0 + 25*0 + 100*1) = 100$
*feature: 'the_others' Score: $(100*0 + 25*0 + 100*1) = 100$
*feature: 'stir_of_echoes' Score: $(100*0 + 25*0 + 100*1) = 100$

2.9.5 Cálculo de la polaridad de las características

Una vez que el sistema ha calculado el peso total que asigna a cada una de las características identificadas en la opinión del usuario, ahora puede calcular la polaridad asociada a cada una de ellas.

Como se ha visto a lo largo de este capítulo, para calcular la polaridad asociada a una característica existen distintos métodos. Estos métodos son: *N_GRAM Before*, *N_GRAM*

After, *N_GRAM Around* y *All Phrase*. La forma en la que cada uno de estos métodos lleva a cabo la clasificación de sentimientos de las características identificadas en la opinión del usuario ha sido descrita anteriormente en el apartado 2.6.1 de este capítulo.

En la tabla 2.9 aparece el proceso de cálculo de la polaridad de cada una de las características identificadas anteriormente. En este ejemplo, el método elegido para la clasificación de sentimientos o cálculo de la polaridad de las características identificadas es *N_GRAM Before*. Además el parámetro del sistema *N_GRAM* tiene valor igual a tres, lo que significa que se tomarán en consideración los valores de sentimiento de las tres palabras que precedan a la expresión lingüística de cada característica para calcular su polaridad.

Tabla 2.9. Extracto del proceso de cálculo de la polaridad de cada una de las características identificadas.

```
4) FEATURES POLARITY IDENTIFICATION
=====

-> Parmeter: N_GRAM = 3, ANALISIS_METHOD = N_GRAM_Before

* feature: 'genre'
->Adjacent words in file: [|assuming|that|the|]
*Word: 'assuming' SWN values: 0.0 0.5 num. senses: 1.0

* feature: 'american_beauty'
->Adjacent words in file: [|did|in|]
*Word: 'in' SWN values: 0.0 0.0 num. senses: 7.0

* feature: 'runtime'
->Adjacent words in file: [|most|of|its|]
*Word: 'most' SWN values: 0.0 0.0 num. senses: 5.0

* feature: ' melissa_sagemiller'
->Adjacent words in file: [|showing|out|to|]
*Word: 'showing' SWN values: 0.0 0.0 num. senses: 2.0
*Word: 'out' SWN values: 0.0 0.14 num. senses: 17.0

* feature: 'wes_bentley'
->Adjacent words in file: [|,||although|] -> Sorry, they don't obtain any
sentimental classification!!
->Alternative adjacent words (file line):
[|,||although|wes_bentley|just||seemed||to||be||playing||the||exact||same||
character||that||he||did||in||american_beauty|]
*Word: 'just' SWN values: 0.312 0.0 num. senses: 10.0
-->Investing sentimental values of : 'just', actual values = (0.0,0.312),
previous word: although wes_bentley
*Word: 'seemed' SWN values: 0.156 0.031 num. senses: 4.0 -> this word comes
from the verb: seem
*Word: 'be' SWN values: 0.027 0.018 num. senses: 14.0
*Word: 'playing' SWN values: 0.0 0.0 num. senses: 3.0
*Word: 'exact' SWN values: 0.031 0.0 num. senses: 4.0
*Word: 'same' SWN values: 0.083 0.104 num. senses: 6.0
```

```
*Word: 'character' SWN values: 0.238 0.0 num. senses: 10.0
*Word: 'he' SWN values: 0.0 0.0 num. senses: 2.0
*Word: 'in' SWN values: 0.0 0.0 num. senses: 7.0

* feature: 'actors'
->Adjacent words in file: [|the|] -> Sorry, they don't obtain any sentimental
classification!!
->Alternative adjacent words (file line):
[|the||actors||are||pretty||good||for||the||most||part|]
*Word: 'actors' SWN values: 0.0 0.0 num. senses: 2.0 -> this word is the
plural of: actor
*Word: 'are' SWN values: 0.0 0.0 num. senses: 1.0
*Word: 'pretty' SWN values: 0.333 0.417 num. senses: 3.0
*Word: 'good' SWN values: 0.597 0.0050 num. senses: 27.0
*Word: 'most' SWN values: 0.0 0.0 num. senses: 5.0
*Word: 'part' SWN values: 0.0 0.0 num. senses: 18.0
```

Como se muestra en el ejemplo, cuando en las *N_GRAM* palabras que preceden a una característica no existe ningún valor de sentimiento asignado en la base de datos de sentimientos, el sistema analiza los sentimientos expresados en la frase entera que contiene la expresión lingüística que identifica a la característica. Se hace así para poder encontrar palabras cercanas a la característica que tengan valores de sentimiento asociados y se pueda clasificar la característica de forma correcta. Esto sucede en aquellos casos puntuales en los que el sistema no tiene información suficiente para clasificar una característica concreta y, en este caso, se comporta de igual manera que cuando aplica el método *All Phrase*. Esta convención se utiliza también en el resto de métodos de clasificación de sentimientos cuando se da el mismo caso.

En la tabla 2.9 se ve que la primera característica identificada es *genre*, que fue identificada como característica porque es una clase de la ontología del dominio. En la opinión la característica *genre* aparece en el párrafo siguiente: “*it's just packaged to look that way because someone is apparently assuming that the genre is still hot with the kids*”. Por lo que el sistema cogerá las tres palabras precedentes a la característica (*N_GRAM* igual a tres) para mirar en la base de datos de sentimientos si existe algún valor de sentimiento para ellas. En este caso las palabras son *assuming*, *that* y *the*. Y solamente existe valor de sentimiento para la palabra *assuming*, quedando las palabras *that* y *the* sin valor de sentimiento asociado. En la tabla se ve que existe un valor de sentimiento para la palabra *assuming* y que éste es de (0.0, 0.5) siendo el primer valor de la dupla el sentido positivo y el segundo valor el sentido negativo. Si hubiese habido más de un sentido asociado a la palabra en la base de datos de sentimientos se hubiese calculado la media aritmética de todos los sentidos, tanto positivos como negativos.

En el ejemplo mostrado este caso se da con la característica *actors*. Esta característica aparece en el siguiente párrafo: “*the actors are pretty good for the most part, although wes bentley just seemed to be playing the exact same character that he did in american beauty, only in a new neighborhood*”. Por lo tanto, cuando aplicamos el método de cálculo de la polaridad *N_GRAM Before* para calcular la polaridad de la característica *actors* con *N_GRAM* igual a 3, se ve que en las *N_GRAM* palabras precedentes a la característica solamente existe la palabra *the*. Como se ha comentado en el párrafo anterior, la palabra *the* no tiene asignado ningún valor de sentimiento (positivo o negativo) en la base de datos de valores de sentimientos, por lo que el sistema analiza la frase entera en la que se encuentra la característica, que en este caso es: “*the actors are pretty good for the most part*”. Esta frase sí contiene palabras con valores de sentimiento asignados en la base de datos de sentimientos, como *pretty* o *good*, lo cual permite al sistema calcular la polaridad de la característica *actors* de manera correcta, ya que ahora sí dispone de información para ello.

Además, en este proceso de cálculo de la polaridad de las características identificadas, el sistema dispone de un mecanismo de detección de cláusulas negativas. Este mecanismo se ha implementado para el idioma inglés y básicamente consiste en detectar aquellas palabras etiquetadas en el proceso de *PLN* que invierten el valor de sentimiento de los adjetivos próximos a ellas. Por tanto, si el sistema detecta que en las palabras analizadas existe una cláusula negativa entonces invierte los valores positivo y negativo de los valores de sentimiento asociados a la palabra que es objeto de la cláusula negativa. Por ejemplo, si el sistema analiza la frase: “*the actress is not good*”, cuando vea que existe una cláusula de negatividad para la palabra *good* (en este caso dada por la palabra *not*) invertirá los valores de sentimiento asociados a la palabra *good* en la base de datos de valores de sentimientos.

Por último, en la tabla 2.10 se muestran los resultados del cálculo de las polaridades asociadas a cada una de las características identificadas en la opinión del usuario. La polaridad asociada a cada característica viene expresada por tres coordenadas, esto es, viene definida como el punto de destino de un vector de posición en el espacio R^3 . Como ya se ha comentado, la primera coordenada expresa el valor de positividad asociado a la característica, la segunda el sentido negativo y la tercera el sentido neutro u objetivo.

En el ejemplo, la característica *genre* con vector de polaridad asociado $\vec{V} = (0.0, 50.0, 0.5)$ tendrá una polaridad claramente negativa. La característica *american beauty*, con vector de polaridad asociado $\vec{V} = (0.0, 0.0, 0.1)$ tendrá una polaridad neutral u objetiva.

La característica *actors* con vector de polaridad asociado $\vec{V} = (15.5, 7.033, 0.129)$ tendrá polaridad positiva.

Tabla 2.10. Resultados de las polaridades asignada a las características identificadas.

```
->Feature: genre, Polarity Vector: (0.0,50.0,0.5)
->Feature: american_beauty, Polarity Vector: (0.0,0.0,0.1)
->Feature: lost_highway, Polarity Vector: (0.0,12.4,0.187)
->Feature: runtime, Polarity Vector: (0.0,0.0,0.1)
->Feature: chase, Polarity Vector: (1.8,3.6,0.946)
->Feature: the_crow, Polarity Vector: (0.0,0.0,0.0)
->Feature: fantasy, Polarity Vector: (2.42,8.37,0.892)
->Feature: the_crow_:salvation, Polarity Vector: (0.0,0.0,0.0)
->Feature: someone, Polarity Vector: (1.0,0.0,0.99)
->Feature: sagemiller, Polarity Vector: (0.0,7.0,0.46)
->Feature: wes_bentley, Polarity Vector: (5.944,5.167,0.987)
->Feature: memento, Polarity Vector: (0.4,10.733,0.626)
->Feature: actors, Polarity Vector: (15.5,7.033,0.129)
->Feature: music, Polarity Vector: (0.9,1.8,0.486)
->Feature: film, Polarity Vector: (7.25,14.792,0.392)
->Feature: edge, Polarity Vector: (1.25,11.9,0.434)
->Feature: melissa_sagemiller, Polarity Vector: (0.0,0.0,0.25)
->Feature: movie, Polarity Vector: (29.495,15.014,0.368)
->Feature: films, Polarity Vector: (0.0,20.8,0.597)
->Feature: kids, Polarity Vector: (17.9,11.9,0.702)
->Feature: blair_witch_2, Polarity Vector: (0.0,0.0,0.0)
->Feature: the_others, Polarity Vector: (0.0,0.0,0.0)
->Feature: thrilling, Polarity Vector: (6.806,3.512,0.367)
->Feature: stir_of_echoes, Polarity Vector: (0.0,0.0,0.0)
->Feature: a_nightmare_of_elm_street_3, Polarity Vector: (0.0,0.0,0.0)
->Feature: years, Polarity Vector: (2.767,3.3,0.313)
->Feature: things, Polarity Vector: (7.469,16.812,0.251)
->Feature: horror, Polarity Vector: (1.8,32.15,0.33)
```

2.9.5.1 Agregación de polaridades de características

Una vez que el sistema ha calculado las polaridades de las características identificadas en la opinión del usuario, comprueba si existen características que puedan ser sinónimas, es decir, que aunque tengan expresiones lingüísticas distintas se refieran al mismo concepto. De ser así, el sistema agregará sus polaridades.

Para calcular las coordenadas del vector de polaridad resultante de la agregación de polaridades, se calculan los valores medios de la suma de cada una de las coordenadas de los vectores de polaridad asociados a las características que hacen referencia al mismo concepto ontológico.

En la tabla 2.11 se muestra el resultado de la agregación de polaridades de características sinónimas partiendo de los vectores de polaridad obtenidos en la tabla 2.10. Por ejemplo, el sistema ha detectado que las características *film*, *movie* y *films* hacen referencia al mismo concepto, por lo tanto, agrega sus polaridades correspondientes en un único concepto que llamará *movie*. Se elige el término *movie* como el término representante de las tres características porque es el término original que aparece en la clase de la ontología del dominio. Esto también se aplica en el caso de los nombres compuestos, ya que el sistema también es capaz de detectar que las características *sagemiller* y *melissa sagemiller* hacen referencia al mismo concepto, por lo que actúa de igual manera que en el caso anterior.

Como se muestra en la tabla 2.10, las características identificadas por el sistema *sagemiller* y *melissa sagemiller* en la $opinion_x$ tienen vectores de polaridad asociados $\vec{V}_1$ (*sagemiller*, $opinion_x$) = (0.0, 7.0, 0.46) y $\vec{V}_2$ (*melissa sagemiller*, $opinion_x$) = (0.0, 0.0, 0.25) respectivamente. El sistema agrega los vectores de polaridad de estas dos características en uno solo, expresado por $\vec{V}_3$ (*melissa sagemiller*, $opinion_x$) = (0.0, 3.5, 0.355), tal y como se muestra en la tabla 2.11 de agrupación de polaridades de características que hacen referencia al mismo concepto.

Tabla 2.11. Resultados de la agregación de polaridades de características sinónimas.

* Grouping synonymous polarities:

```
->Feature: genre, Polarity Vector: (0.0,50.0,0.5)
->Feature: american_beauty, Polarity Vector: (0.0,0.0,0.1)
->Feature: lost_highway, Polarity Vector: (0.0,12.4,0.187)
->Feature: runtime, Polarity Vector: (0.0,0.0,0.1)
->Feature: chase, Polarity Vector: (1.8,3.6,0.946)
->Feature: the_crow, Polarity Vector: (0.0,0.0,0.0)
->Feature: fantasy, Polarity Vector: (2.42,8.37,0.892)
->Feature: the_crow_:salvation, Polarity Vector: (0.0,0.0,0.0)
->Feature: someone, Polarity Vector: (1.0,0.0,0.99)
->Feature: wes_bentley, Polarity Vector: (5.944,5.167,0.987)
->Feature: actors, Polarity Vector: (15.5,7.033,0.129)
->Feature: memento, Polarity Vector: (0.4,10.733,0.629)
->Feature: music, Polarity Vector: (0.9,1.8,0.486)
->Feature: edge, Polarity Vector: (1.25,11.9,0.434)
->Feature: melissa_sagemiller, Polarity Vector: (0.0,3.5,0.355)
->Feature: movie, Polarity Vector: (7.291,16.312,0.452)
->Feature: kids, Polarity Vector: (17.9,11.9,0.702)
->Feature: blair_witch_2, Polarity Vector: (0.0,0.0,0.0)
->Feature: the_others, Polarity Vector: (0.0,0.0,0.0)
->Feature: thrilling, Polarity Vector: (6.806,3.512,0.367)
->Feature: stir_of_echoes, Polarity Vector: (0.0,0.0,0.0)
->Feature: a_nightmare_of_elm_street_3, Polarity Vector: (0.0,0.0,0.0)
->Feature: things, Polarity Vector: (7.469,16.812,0.251)
->Feature: years, Polarity Vector: (2.767,3.3,0.313)
->Feature: horror, Polarity Vector: (1.8,32.15,0.33)
```

2.9.6 Análisis de sentimientos

Una vez que el sistema ha agregado las polaridades de características sinónimas de entre el conjunto de características identificadas en la opinión del usuario, el último paso a realizar es el proceso de clasificación de sentimientos, también llamado análisis de sentimientos.

Como se ha visto anteriormente, el vector de polaridad global asociado a una opinión se calculará como la suma de los vectores de polaridad de las características identificadas en la misma. Estos vectores tendrán unas coordenadas concretas en el espacio R^3 y han sido ponderados anteriormente en función de la importancia de la característica a la que van asociados dentro de la opinión que se está analizando.

Siguiendo con el ejemplo, y tomando los vectores de polaridad asociados a las características que aparecen en la tabla 2.11, el vector de polaridad asociado a la opinión $opinion_x$ se calculará según la fórmula 2.5. Por tanto, tomando los valores concretos de los vectores de polaridad de las características, el vector de polaridad global de la opinión del usuario $\vec{V}(userop_x)$ se calcula de la siguiente manera:

$$\begin{aligned}\vec{V}(opinion_x) = \sum_{i=1}^n \vec{V}(f_i, opinion_x) &= \vec{V}_1(genre, opinion_x) + \vec{V}_2(american\_beauty, \\ &opinion_x) + \vec{V}_3(lost\_highway, opinion_x) + \vec{V}_4(runtime, opinion_x) + \vec{V}_5(chase, opinion_x) \\ &+ \vec{V}_6(the\_crow, opinion_x) + \vec{V}_7(fantasy, opinion_x) + \vec{V}_8(the\_crow:salvation, opinion_x) \\ &+ \vec{V}_9(someone, opinion_x) + \vec{V}_{10}(wes\_bentley, opinion_x) + \vec{V}_{11}(actors, opinion_x) \\ &+ \vec{V}_{12}(memento, opinion_x) + \vec{V}_{13}(music, opinion_x) + \vec{V}_{14}(edge, opinion_x) \\ &+ \vec{V}_{15}(melissa\_sagemiller, opinion_x) + \vec{V}_{16}(movie, opinion_x) + \vec{V}_{17}(kids, opinion_x) \\ &+ \vec{V}_{18}(blair\_witch\_2, opinion_x) + \vec{V}_{19}(the\_others, opinion_x) + \vec{V}_{20}(thrilling, opinion_x) \\ &+ \vec{V}_{21}(stir\_of\_echoes, opinion_x) + \vec{V}_{22}(a\_nightmare\_of, opinion_x) + \vec{V}_{23}(elm\_street\_3, \\ &opinion_x) + \vec{V}_{24}(things, opinion_x) + \vec{V}_{25}(years, opinion_x) + \vec{V}_{26}(horror, opinion_x) = \\ &(4.309, 11.676, 0.354).\end{aligned}$$

Por lo tanto, aplicando la fórmula 2.6, $Polaridad(opinion_x) = -11.676$. Esto implica que la polaridad de la $opinion_x$ es negativa, ya que el punto $P = (4.309, 11.676, 0.354)$ es un punto interior de la pirámide geométrica cuyo volumen está compuesto de todos los puntos de destino de los vectores de posición con dirección negativa (P_{Neg}). Esto es así porque la función $estaContenido(P, P_{Neg})$ devuelve *verdadero*, y significa que, en términos generales, la polaridad global de la opinión de usuario $opinion_x$ es negativa.

En la tabla 2.12 se muestran los resultados obtenidos para la clasificación de sentimientos de la opinión del ejemplo y de cada una de las características identificadas en ella. El sistema concluye que la polaridad global de la opinión del usuario es negativa y luego detalla la clasificación de sentimiento de cada una de las características identificadas en ella.

Tabla 2.12. Resultados del análisis de sentimientos de la opinión del usuario y de sus características identificadas.

5) OPINION MINING

=====

-> Opinion: neg1.txt, Global Polarity Vector: (4.309, 11.676, 0.354)

-> Opinion Global Sentiment Classification: NEG

-> Opinion Detailed Features Sentiment Classification:

feature	polarity
genre	NEG
american_beauty	NEU
lost_highway	NEG
runtime	NEU
chase	NEG
the_crow	NEU
fantasy	NEG
the_crow:salvation	NEU
someone	POS
wes_bentley	POS
actors	POS
memento	NEG
music	NEG
edge	NEG
melissa_sagemiller	NEG
movie	NEG
kids	POS
blair_witch_2	NEU
the_others	NEU
thrilling	POS
stir_of_echoes	NEU
a_nightmare_of_	
elm_street_3	NEU
things	NEG
years	NEG
horror	NEG

2.10 Resumen

En este capítulo se presenta la arquitectura propuesta para cumplir el principal objetivo de esta tesis doctoral: desarrollar un nuevo método para la minería de opiniones basada en características que utiliza técnicas de la Web Semántica, en concreto las ontologías.

La arquitectura expuesta en este capítulo se divide en cuatro módulos principales: módulo de procesamiento del lenguaje natural, módulo de identificación de características, módulo de cálculo de polaridades de características y módulo de análisis de sentimientos.

El módulo de procesamiento del lenguaje natural lleva a cabo el análisis morfológico de las opiniones contenidas en el corpus. El módulo de identificación de características tiene como objetivo la extracción de todas las características contenidas en las opiniones de los usuarios utilizando para ello una ontología del dominio. El módulo de cálculo de la polaridad de las características tiene como objetivo la obtención de los vectores de polaridad asociados a las características identificadas en las opiniones. Por último, el módulo de análisis de sentimientos lleva a cabo el proceso de clasificación de sentimientos de las opiniones de los usuarios y de las características identificadas en ellas.

El proceso de análisis de sentimientos de opiniones que se propone en este trabajo está dividido en dos vertientes:

- **Análisis de sentimientos de las características.** Se analiza la orientación de sentimiento de cada una de las características identificadas en la opinión de los usuarios. El sistema clasifica cada característica identificada como positiva, negativa o neutra.
- **Análisis del sentimiento de las opiniones.** Se analizan los sentimientos de la opinión expresada por un usuario en su totalidad, es decir, clasificando la opinión globalmente como positiva, negativa o neutra.

En el análisis de sentimientos de las opiniones el sistema utiliza los valores de los vectores de polaridad asociados a cada una de las características dentro de las opiniones. Realiza cálculos vectoriales con ellos para obtener un vector de polaridad global en R^3 . Este vector determina la clasificación del sentimiento de las opiniones de los usuarios.

La metodología propuesta es independiente del dominio e idioma en los que se aplique. Sin embargo, se han empleado recursos lingüísticos propios del idioma inglés para implementarla en un sistema de software. Esto implica que también se haya utilizado un corpus con opiniones de usuarios escritas en inglés.

Al final de este capítulo se muestra un ejemplo práctico de cómo funciona la aplicación de la metodología en el dominio de las películas cinematográficas utilizando opiniones de usuarios escritas en inglés.

Capítulo III

Un sistema de software para la
minería de opiniones de usuarios
a partir de texto en lenguaje
natural

3.1 *Introducción*

En este capítulo se muestra el diseño e implementación de un sistema de software construido en base a los requisitos de funcionalidad de la metodología propuesta en esta tesis doctoral. El objetivo general de la aplicación es servir como entorno para la ejecución de procesos de minería de opiniones basada en características a partir de texto libre.

En primer lugar, se identifican las herramientas y librerías que han sido fundamentales para el desarrollo de la aplicación. La elección de éstas ha marcado la pauta en el trabajo desarrollado a la hora de evaluar la metodología que se propone. Los criterios que se han tenido en cuenta para su selección han sido los siguientes:

- **Disponibilidad en la comunidad investigadora.** Se han elegido herramientas *Open Source* disponibles para estudiantes e investigadores con el objetivo de poder explotar al máximo sus funcionalidades.
- **Fiabilidad y robustez.** Las herramientas elegidas aportan suficientes garantías con respecto a la fiabilidad y robustez de las mismas, ya que han sido empleadas en anteriores trabajos de investigación con éxito.
- **Actualizaciones.** Para conseguir que la aplicación construida en este trabajo evolucione con el paso del tiempo es necesario que las librerías y paquetes seleccionados sean de fácil actualización. De esta forma el sistema no quedará obsoleto.
- **Rendimiento.** Se han elegido herramientas de trabajo que facilitan el desarrollo de software de forma ágil y optimizada en el tiempo.

Para finalizar el capítulo se muestra un manual de usuario que explica detalladamente cada una de las vistas que conforman la interfaz de usuario de la aplicación construida.

3.2 *Fundamentos tecnológicos*

Para el desarrollo de la aplicación de software que implementa la metodología propuesta en esta tesis se han empleado diversas herramientas y librerías. En esta sección se describen las características de estos elementos, que han sido fundamentales para la construcción de la aplicación.

3.2.1 Herramientas

Las herramientas más importantes que han sido utilizadas para el desarrollo de la aplicación de software mostrada en este capítulo son las siguientes: *Java* (JDK 1.6), *Eclipse* (versión *Luna*, Release 4.4), *Stanford Log-linear Part-Of-Speech Tagger* (Version 3.0.3), etiquetas *Penn Treebank*, *JENA* (versión 2.6.2), *Protégé* 3.4, *MySQL* 5.5, *Apache Tomcat* 6.0, *JUnit* 3.8, *Apache Maven* 3.2, *Apache Ant* 1.9.4, *HTML* 5, *JavaScript*, *JSP*, *CVS*, *Jenkins* 1.580.1 y *Sonar* 4.5.2.

Además, como herramientas complementarias al desarrollo de la aplicación, se han utilizado *SentiWordNet* (versión 3.0.0), *WordNet* (versión 2.1) y *JAWS* (versión 1.3). Herramientas específicas del idioma inglés para que la metodología pueda aplicarse en textos escritos en este idioma.

Por otro lado, se ha seguido la metodología de desarrollo *RUP* (*Rational Unified Process*) para el desarrollo de la aplicación (Kruchten, 2004). La metodología *RUP* incluye como lenguaje de modelado y diseño de aplicaciones a *UML* (*Unified Modelling Language*) (Rumbaugh et al., 2004) que es el lenguaje de modelado de sistemas de software orientado a objetos más conocido y utilizado en la actualidad.

A continuación se realiza una breve descripción de algunas de las herramientas citadas atendiendo a su importancia dentro de la implementación del sistema construido.

3.2.1.1 Java

Java es una tecnología orientada al desarrollo de software con la que se puede llevar a cabo cualquier tipo de programa (Deitel & Deitel, 2006). Hoy en día la tecnología *Java* ha cobrado mucha importancia en el ámbito de Internet gracias a su plataforma *JEE*³³. La tecnología *Java* está compuesta por dos elementos: el lenguaje *Java* y su plataforma (entendiendo plataforma como la máquina virtual de *Java*) [(*JVM*)³⁴]. Una de las principales características que ha favorecido el crecimiento y difusión del lenguaje *Java* es su capacidad multiplataforma, es decir, que el código escrito con este lenguaje funcione sobre cualquier plataforma de software y hardware como *Windows* y *Linux*.

El lenguaje de programación empleado para la implementación de la aplicación de software construida en este trabajo ha sido *Java*. Esta elección se basó en las numerosas

³³ **JEE** proveniente de las siglas en inglés de *Java Enterprise Edition*.

³⁴ **JVM** proveniente de las siglas en inglés de *Java Virtual Machine*.

propiedades y ventajas que tiene *Java* con respecto a otros lenguajes de programación (Flanagan, 1999). De entre sus ventajas destacan: orientado a objetos (otorga una gran flexibilidad a la hora de presentar jerarquías complejas), distribuido (aporta mecanismos para interactuar con otras máquinas mediante protocolo *TCP/IP*), robusto (tipado estáticamente, generación de excepciones, etc.), seguro (comprobaciones de privilegios), soporte Web (numerosas *APIs* para el desarrollo de aplicaciones Web), en expansión (lenguaje muy popular que no deja de evolucionar). Otra razón por la que se ha seleccionado este lenguaje es porque algunas de las librerías para el manejo de ontologías, como *JENA*, estaban desarrolladas en este lenguaje.

3.2.1.2 Eclipse

Para la implementación de la aplicación se ha utilizado el entorno de desarrollo *Eclipse*³⁵ en su versión *Luna*. Eclipse es una plataforma de software de código abierto que incorpora un entorno integrado de desarrollo (*IDE*) para la construcción, despliegue y gestión de software en su ciclo de vida. Este entorno ha sido considerado como el mejor del mercado en el año 2005 (Geer, 2005) y, actualmente, es la plataforma de desarrollo de software más utilizada. Eclipse está estructurado como una colección de *plug-ins* donde cada módulo contiene el código que proporciona algunas de las funcionalidades del producto. Entre las características principales de Eclipse se encuentran: el editor de texto, resaltado de sintaxis, compilación en tiempo real, pruebas unitarias con *JUnit*, control de versiones con *CVS* (*Concurrent Versions System*), integración con *Ant*, distintos asistentes para creación de proyectos, refactorización, etc.

3.2.1.3 Stanford Log-linear Part-Of-Speech Tagger

El *Stanford Log-linear Part-Of-Speech Tagger*¹⁶ (Toutanova & Manning, 2000; Toutanova et al., 2003) es un conjunto de herramientas que se utilizan para desarrollar y desplegar los componentes de software que procesan el lenguaje natural. Este software ayuda a científicos y desarrolladores de dos maneras: mediante la especificación de una arquitectura para procesar el lenguaje natural mediante software y proporcionando un entorno que implementa la arquitectura y se puede utilizar para incrustar capacidades de procesamiento del lenguaje en diversas aplicaciones.

³⁵ **Eclipse**, disponible en: <http://www.eclipse.org/>. Última consulta realizada en Mayo de 2015.

Un etiquetador *Part-Of-Speech* (en inglés, *POS Tagger*) es un software que, a partir de un texto en un idioma determinado, etiqueta cada palabra con su categoría gramatical (sustantivo, verbo, adjetivo, etc.). El software *Stanford Log-linear Part-Of-Speech Tagger* está implementado en *Java*, y fue originalmente desarrollado por Kristina Toutanova. Desde entonces varios investigadores se han centrado en la mejora de su velocidad, rendimiento, facilidad de uso y soporte para otros idiomas. En la actualidad este etiquetador soporta los idiomas: árabe, chino, francés, español y alemán (última consulta realizada en mayo de 2015).

Stanford Log-linear Part-Of-Speech Tagger es un paquete de código abierto que proporciona un conjunto de herramientas para el análisis del lenguaje natural. Está diseñado para ser utilizado como una biblioteca externa de cualquier aplicación que requiera este tipo de servicios. Las características principales que ofrece este paquete son:

- Tokenización del texto.
- Separación de oraciones.
- Análisis morfológico.
- Reconocimiento de conceptos compuestos.
- Predicción probabilística de categorías de palabras desconocidas.
- Detección de nombres de entidades.
- Reconocimiento de fechas, números, relaciones, dinero y magnitudes físicas (velocidad, peso, temperatura, densidad, etc.).
- Clasificación de nombres de entidades.
- Anotación y desambiguación del sentido basado en *WordNet*.
- Resolución de correferencia nominal.

La razón por la que se eligió el etiquetador *Stanford Log-linear Part-Of-Speech Tagger* sobre otros etiquetadores libres como *FreeLing*³⁶ (Padró, 2011; Padró & Stanilovsky, 2012) o *GATE*³⁷ (Cunningham et al., 2011; Cunningham et al., 2013) es porque, en los experimentos llevados a cabo en este trabajo, se comportó de manera más estable y robusta en un entorno basado en *J2EE* que los otros etiquetadores.

³⁶ <http://nlp.lsi.upc.edu/freeling/>, última consulta realizada en Febrero de 2015.

³⁷ <https://gate.ac.uk/>, última consulta realizada en Febrero de 2015.

3.2.1.4 Etiquetas Penn Treebank

TreeBank (Santorini, 1990), denominado también *Penn Treebank*, es un corpus lingüístico en el que cada frase ha sido anotada con su estructura sintáctica. La estructura sintáctica se ha representado generalmente como una estructura arbórea que recibe la denominación de *TreeBank*.

Como ejemplo en la tabla 3.1 se muestra el análisis sintáctico de la frase: “*Juan ama a Sara*”. Este análisis sintáctico puede ser representado con un sistema anidado de paréntesis en el texto siguiendo la notación *Penn Treebank*.

Tabla 3.1. Ejemplo de análisis sintáctico utilizando notación *Penn Treebank*.

(S (NP (NNP Juan)) (VP (VPZ ama) (PP (TO a) (NP (NNP Sara)))) (. .))

Los corpus *Treebanks* se pueden crear a mano mediante un grupo de lingüistas que anotan cada frase con una estructura sintáctica, o mediante procedimientos semi-automáticos donde un analizador sintáctico (parser, en inglés) asigna la estructura bajo la supervisión de un lingüista. En la práctica, el completo control del *parseado* del lenguaje natural es una labor intensiva en la que varios equipos de lingüistas pueden dedicar varios años.

Una ventaja de usar las etiquetas *Penn TreeBank* es la reducción de etiquetas respecto a otros sistemas de etiquetado, por ejemplo *FreeLing* o *Brown Corpus*. Esta reducción se establece a partir de tres criterios:

- **Recuperación:** se trata de eliminar redundancia tomando en cuenta conjuntamente la información léxica y la información sintáctica. Por ejemplo, si en el *Brown Corpus* se distingue una etiqueta para lo que llaman pre-cualificadores (*quite, rather, such*) y pre-cuantificadores (*all, half, many*), en el sistema de codificación del *Penn Treebank* se asigna a todas estas palabras una etiqueta única *PDT* (*predeterminer*).
- **Consistencia:** reducir el tamaño del conjunto de etiquetas reduce las posibilidades de inconsistencias en el etiquetado. El *Brown Corpus*, por ejemplo,

utiliza para los adverbios *there* and *now* siempre la etiqueta *RB*, mientras que *here* and *then* se etiquetan en algunas ocasiones como *RB* y en otras como *RN*.

- **Función sintáctica:** puesto que el objetivo del *Penn TreeBank* es preparar textos para la fase de análisis, se asigna función sintáctica siempre que sea posible. Por ejemplo, el *Brown Corpus* etiqueta *both* como *ABX* (pre-cuantificador, conjunción doble), sin tener en cuenta si funciona como un modificador prenominal (*both the boys*) o como un postnominal (*the boys both*), como un núcleo de una frase nominal (*both of the boys*) o como parte de una conjunción coordinada compleja (*both boys and girls*). El sistema de etiquetado del *Penn Tree Bank* diferencia todos estos contextos sintácticos como *PDT* (predeterminante), *RB* (adverbio), *NNS* (nombre común plural) y *CC* (conjunción coordinada), respectivamente.

3.2.1.5 WordNet

WordNet (Miller, 1995) es una base de datos léxica del idioma inglés. Agrupa palabras en conjuntos de sinónimos llamados *synsets*, proporcionando definiciones cortas y generales y almacenando las relaciones semánticas entre los conjuntos de sinónimos. Su propósito es doble: producir una combinación de diccionario y tesaurus cuyo uso sea más intuitivo y soportar análisis automático de texto para aplicaciones de inteligencia artificial. La base de datos y las herramientas software se han liberado bajo una licencia *BSD* y pueden ser descargadas y usadas libremente. Además la base de datos puede consultarse online.

WordNet fue creado, y es mantenido, por el *Cognitive Science Laboratory* de la Universidad de Princeton bajo la dirección del profesor de psicología George A. Miller en 1985. A través de los años, el proyecto ha recibido financiación de agencias gubernamentales interesadas en el procesamiento del lenguaje natural.

Desde noviembre de 2012 la última versión de *WordNet* es la 3.1 (anunciada en junio del 2011), pero la última versión liberada fue la 3.0 (liberada en diciembre del 2006). La base de datos 3.0 contiene 155.287 palabras organizadas en 117.659 *synsets* para un total de 206.941 pares de palabras, y en forma comprimida tiene 12 megabytes de tamaño.

WordNet distingue entre sustantivos, verbos, adjetivos y adverbios siguiendo distintas reglas gramaticales. Cada *synset* contiene un grupo de palabras que son sinónimos o

collocations (secuencia de palabras que unidas toman un significado específico). Diferentes significados de una palabra están en distintos *synsets*. El significado de los *synsets* queda más claro con la definición de *glosses* (definición y/o sentencias de ejemplo). Un ejemplo de *synset* con *gloss* es el que aparece en la tabla 3.2.

Tabla 3.2. Ejemplo de *synset* con *gloss* dentro de *WordNet*.

good, right, ripe – (*most suitable or right for a particular purpose; "a good time to plant tomatoes"; "the right time to act"; "the time is ripe for great sociological changes"*)

La mayoría de los *synsets* están conectados a otros *synsets* mediante numerosas relaciones semánticas. Estas relaciones varían según el tipo de palabra. Aunque las relaciones semánticas se aplican a todos los miembros de un *synset* porque comparten significado. Las palabras también pueden estar conectadas a otras palabras a través de relaciones léxicas, incluyendo la antonimia.

Además, *WordNet* también provee el *polysemy count* de una palabra (el número de *synsets* que contienen la palabra). Si una palabra participa en varios *synsets* es porque tiene varios significados. Las funciones morfológicas del software distribuidas con la base de datos tratan de deducir el lema o raíz de una palabra de un texto. Solo la raíz de cada palabra es almacenada en la base de datos.

WordNet estructura el conocimiento de la siguiente manera: los sustantivos y los verbos están organizados en jerarquías. Por ejemplo, en la tabla 3.3 aparece la jerarquía *hypernym* del primer significado de la palabra *dog*; las palabras en el mismo nivel son sinónimos unas de otras. Cada conjunto de sinónimos (*synset*) tiene un único identificador y comparte sus propiedades, además de una definición *gloss*.

Tabla 3.3. Ejemplo de jerarquía *hypernym* en *WordNet*.

dog, domestic dog, Canis familiaris
 => canine, canid
 => carnivore
 => placental, placental mammal, eutherian, eutherian mammal
 => mammal
 => vertebrate, craniate
 => chordate
 => animal, animate being, beast, brute, creature, fauna
 => ...

El objetivo de *WordNet* era desarrollar un sistema que fuese consistente con el conocimiento adquirido a través de los años para conocer la forma en la que los humanos procesan el lenguaje. Por ejemplo, los individuos pueden verificar rápidamente que los canarios pueden volar porque es un ave, pero no identificar si tiene piel. Esto muestra que los humanos almacenamos información semántica en una forma muy parecida a como lo hace *WordNet*, porque solo retenemos la información más específica que necesitamos para diferenciar un concepto en particular de otros similares.

A diferencia de otros diccionarios, *WordNet* no incluye información sobre la etimología, pronunciación y la forma de los verbos irregulares. La información lexicográfica y semántica actual se mantiene en *lexicographical files*, que son procesados por una herramienta llamada *grind* para producir la base de datos distribuida. Ambos, el *grind* y los *lexicographer files*, están disponibles libremente en una distribución separada, pero la modificación y mantenimiento de la base de datos requiere experiencia.

A pesar de que *WordNet* contiene un rango suficientemente amplio de palabras comunes, no cubre todo el vocabulario de un dominio específico. Como está diseñado para actuar como capa subyacente para diferentes aplicaciones, esas aplicaciones no pueden ser usadas en dominios específicos que no son cubiertos por *WordNet*.

Por último, cabe destacar que *WordNet* es el lexicón computacional en inglés más usado para desambiguar el significado de las palabras (*Word Sense Disambiguation*, *WSD*) una tarea que tiene como objetivo asignar el significado más apropiado a las palabras en un contexto específico.

3.2.1.6 SentiWordNet

SentiWordNet (Esuli y Sebastiani, 2006) es un recurso para aplicaciones de minería de opinión obtenido a partir de *WordNet 3.0*. El objetivo de *SentiWordNet* es proporcionar una extensión para *WordNet*, de tal manera que todos los *synsets* pueden estar asociados con un valor relativo a la connotación negativa, positiva o neutra. *SentiWordNet 3.0* es la versión mejorada de *SentiWordNet 1.0* y públicamente disponible gratuitamente para fines de investigación con una interfaz Web (Kreutzer & Witte, 2014).

SentiWordNet etiqueta cada *synset* con un valor para cada categoría de entre 0.0 y 1.0. La suma de los tres valores es siempre 1.0, de modo que cada *synset* puede tener un valor distinto de cero para cada sentimiento, porque algunos *synsets* pueden ser positivos, negativos o neutros en función del contexto en el que se utilizan.

La interfaz Web permite al usuario buscar cualquier *synset* perteneciente a *WordNet* para obtener la información de sentimiento en *SentiWordNet* asociada a ese *synset* de manera visual. Cada categoría está vinculada a un color. Así, el color rojo corresponde a la negatividad, azul para la objetividad (o neutralidad) y verde para representar los valores positivos. En la figura 3.1 podemos ver un ejemplo de la visualización de la *synset* *good#5*. La ventaja de usar *synsets* en lugar de términos, es ofrecer diferentes puntuaciones de sentimiento para cada sentido de una palabra, porque las connotaciones pueden diferir en una palabra dependiendo del sentido.



Figura 3.1. Ejemplo de visualización de un *synset* con puntuaciones en *SentiWordNet*.

3.2.1.7 Protégé 3.4.1

Protégé es un editor de ontologías gratuito y de código fuente abierto. Provee un conjunto de herramientas para construir modelos de dominio y aplicaciones basadas en conocimiento. Además implementa un amplio conjunto de estructuras de modelado del conocimiento y acciones que permiten la creación, visualización y manipulación de ontologías en varios formatos de representación.

Protégé se puede personalizar para crear un entorno amigable para facilitar la creación de modelos de conocimiento y la introducción de información. Además, puede extenderse por medio de una arquitectura de *plug-ins* y *APIS* para *Java* que faciliten la construcción de herramientas y aplicaciones basadas en conocimiento.

La plataforma *Protégé* soporta actualmente dos formas principales de modelado de ontologías:

- El editor **Protégé-Frames** permite a los usuarios construir y poblar las ontologías basadas en *frames*, en conformidad con el *Open Knowledge Base Connectivity* (*OKBC*³⁸).

³⁸ <http://www.ai.sri.com/~okbc/>. Última consulta realizada en Mayo de 2015.

- El editor **Protégé-OWL** permite a los usuarios construir ontologías para la Web Semántica, en particular, para el lenguaje *OWL* de la *W3C*.

Se ha elegido este software por ser uno de los paquetes en el que se puede representar la ontología de forma gráfica, además de ser uno de los editores más extendido para desarrollar ontologías.

3.2.2 Librerías

En la implementación del sistema se han utilizado una gran cantidad de librerías *Java*. Se pueden destacar dos: la librería que nos permite interactuar con las ontologías del dominio y la librería que nos permite trabajar con el conjunto de sinónimos de un concepto dado. Estas son las librerías *JENA* y *JAWS*, respectivamente. A continuación las explicamos en detalle.

3.2.2.1 JENA

*Apache Jena*³⁹ es un entorno *Java* para construir aplicaciones basadas en ontologías que se desarrolló el año 2000 en *HP Labs*. En 2009 *HP* cedió el proyecto a la fundación *Apache* que decidió adoptarlo en noviembre del siguiente año.

La arquitectura de *Jena* incluye las siguientes funcionalidades:

- *API* para trabajar (leer, procesar, escribir) con ontologías escritas en *RDF* y *OWL*.
- Motor de inferencia para razonar sobre ontologías *RDF* y *OWL*.
- Estrategias de almacenamiento flexible para almacenar tripletas *RDF* en memoria o fichero.
- Motor de consultas compatible con especificación *SPARQL*⁴⁰.

En la implementación del sistema de análisis de sentimientos propuesta en esta tesis se ha utilizado el *API* de *Jena* para trabajar con ontologías escritas en *OWL*. En concreto, se ha empleado la *Jena API 2.6.2*.

³⁹ *JENA*, framework *Java* para construir aplicaciones basadas en ontologías. <https://jena.apache.org/>. Última consulta realizada en Abril de 2015.

⁴⁰ *SPARQL Query Language for RDF*, <http://www.w3.org/TR/rdf-sparql-query/>. Última consulta realizada en Mayo de 2015.

3.2.2.2 JAWS

*JAWS*⁴¹ es una *API* de *Java* para llevar a cabo búsquedas en *WordNet*. Es una *API* simple y rápida, y es compatible con *WordNet 2.1* y *WordNet 3.0*. Se puede utilizar con versiones de *Java* 1.4 y posteriores.

JAWS fue creado y es mantenido por Spell Brett, miembro adjunto de la facultad en el departamento de *Computer Science and Engineering (CSE)* de la *Southern Methodist University*.

En la implementación de la metodología de análisis de sentimientos propuesta en esta tesis, se utiliza *JAWS* a la hora de buscar los sinónimos de las características identificadas en el proceso de identificación de características del sistema. Esta *API* es de fácil manejo y proporciona mucha potencia a la hora de trabajar con *WordNet*. En concreto, se ha empleado la *JAWS API 1.3* cuando se ha trabajado con opiniones de usuarios escritas en inglés y se ha utilizado una base de datos de sinónimos en este idioma.

En la figura 3.2 se recoge un extracto de código *Java* que muestra cómo se pueden recuperar *synsets* de la base de datos *WordNet*. Este código recupera todos *synsets* sustantivo y recorre cada uno de ellos para imprimir su primera forma de palabra, su descripción y el número de hipónimos asociados con ese *synset*.

```
NounSynset nounSynset;
NounSynset[] hyponyms;

WordNetDatabase database = WordNetDatabase.getFileInstance();
Synset[] synsets = database.getSynsets("fly", SynsetType.NOUN);
for (int i = 0; i < synsets.length; i++) {
    nounSynset = (NounSynset)(synsets[i]);
    hyponyms = nounSynset.getHyponyms();
    System.err.println(nounSynset.getWordForms()[0] +
        ": " + nounSynset.getDefinition() + ") has " +
        hyponyms.length + " hyponyms");
}
```

Figura 3.2. Extracto de código Java que usa el API JAWS.

⁴¹ *JAWS*, del inglés: *Java API for WordNet Searching*. <http://lyle.smu.edu/~tspell/jaws/>. Última consulta realizada en Enero de 2015.

3.3 *Herramienta software para la clasificación de sentimientos de opiniones de usuarios expresadas en lenguaje natural*

3.3.1 Introducción

En este apartado se explica el desarrollo de una herramienta de software que sirve de entorno de ejecución para la minería de opiniones basada en características guiada por ontologías a partir de texto en lenguaje natural. Esta herramienta se ha implementado siguiendo los requerimientos de funcionalidad de la metodología propuesta en esta tesis doctoral. Toma un corpus de opiniones como entrada, lo procesa y como resultado final proporciona un informe con el análisis de sentimientos de las opiniones y de las características identificadas en ellas. A su vez proporciona una serie de ficheros intermedios en los que se pueden consultar los resultados de cada uno de los procesos que se llevan a cabo dentro de la arquitectura del sistema.

La aplicación ha sido diseñada utilizando los patrones de diseño que aparecen en el trabajo de (Gamma, 2003). Ha resultado fundamental el uso de clases adaptadoras (patrón de diseño *Adapter*) para desacoplar la implementación del software del sistema de los recursos lingüísticos necesarios para implementar la metodología. Además, a la hora de diseñar la aplicación software que implementa la metodología propuesta en esta tesis doctoral, se han tenido en cuenta los criterios de construcción de software orientado a objetos que aparecen en (Meyer, 1998).

El sistema de software construido es una aplicación Web, por lo que no requiere una instalación ni configuración previa de aplicaciones locales. Dando la posibilidad al usuario, ya no sólo de poder trabajar con el sistema en cualquier parte del mundo, sino de mantener una copia de seguridad de sus datos en el servidor Web sin la necesidad de tener que mantener una copia de seguridad permanente en su dispositivo personal.

La interfaz de esta herramienta software se ha diseñado para que pueda ser usada tanto por usuarios expertos en procesamiento del lenguaje natural (interesados en obtener el análisis de sentimientos de las opiniones de los usuarios) como por expertos del dominio que no tengan muchos conocimientos en las tecnologías utilizadas en esta tesis doctoral.

3.3.2 Una sesión con el sistema implementado en esta tesis

En esta sección se explica la manera de utilizar la herramienta desarrollada y sus interfaces. La aplicación ofrece las distintas opciones de las que disponen los usuarios para utilizar y configurar la herramienta, como: crear un proyecto nuevo, abrir un proyecto existente, establecer los parámetros iniciales, seleccionar el método de clasificación de sentimientos, configurar cada módulo del método o seleccionar un idioma para las interfaces.

A continuación se explica en detalle cada una de las posibilidades que nos ofrece la herramienta.

3.3.2.1 Acceso al sistema e interfaz inicial

En la figura 3.3 se muestra la pantalla de acceso al sistema. El usuario tiene que identificarse antes de que el sistema le permita hacer uso de las funcionalidades que ofrece. La aplicación se ha diseñado así porque se pretende llevar un registro individual de cada usuario para que éste pueda crear y trabajar en sus propios proyectos. De esta manera cada usuario tendrá su propio entorno de trabajo personalizado y podrá guardar sus propias configuraciones sin que esto influya para nada en el trabajo llevado a cabo por otros usuarios en el sistema.

UNIVERSIDAD DE MURCIA

Departamento de Informática y Sistemas

Sistema basado en Minería de Opiniones basada en Características guiada por Ontologías

LOGIN

Acceso Restringido

Accede al sistema de clasificación sentimental de opiniones basado en características a través de tu usuario y contraseña.

ACCESO AL SISTEMA

Usuario: isidro.penalver.martinez@um.es

Contraseña:

Borrar Acceder

W3C XHTML 1.0 W3C CSS

Universidad de Murcia, 2014
Departamento de Informática y Sistemas

Figura 3.3. Pantalla de acceso al sistema software de clasificación de sentimientos de opiniones.

En la figura 3.4 se muestra la interfaz inicial del sistema que permite el acceso a todas las opciones disponibles a través de sus cinco menús: inicio, configuración, procesos del sistema, idioma y ayuda.

En la pantalla de inicio también se aporta información al usuario correspondiente a las últimas novedades en las tecnologías de minería de opiniones o Web Semántica. Esta información se actualiza dinámicamente cada vez que se publica nueva información en los canales *RSS* de los que carga información de interés. Además cada usuario podrá elegir la lista de canales de los que desea recibir información en tiempo real, es decir, podrá suscribirse o eliminar la suscripción a los distintos canales ofertados según sus preferencias.



Figura 3.4. Pantalla de inicio del sistema software de clasificación de sentimientos de opiniones.

3.3.2.2 Configuración del sistema

Con el menú de configuración se pueden configurar los proyectos del sistema y las opciones generales. En la figura 3.5 se muestra el desplegable con ambas opciones.



Figura 3.5. Pantalla del menú de configuración del sistema software de clasificación de sentimientos de opiniones.

Dentro de la configuración de proyectos es donde cada usuario puede dar de alta los proyectos en los que va a trabajar. Es decir, un usuario puede estar trabajando en paralelo sobre proyectos enmarcados en distintos dominios cada uno de ellos con sus configuraciones pertinentes.

En la parte de configuración de proyectos se pueden dar de alta nuevos proyectos y se puede obtener un listado de todos los proyectos en los que está implicado un determinado usuario. En la figura 3.6 se introduce un nuevo proyecto en el que se va a trabajar sobre opiniones de los usuarios enmarcadas en el ámbito cinematográfico. Se introduce un nombre para identificar al proyecto, junto a una descripción del mismo y la fecha en la que se da de alta, para así poder clasificar los proyectos con facilidad. Posteriormente, se introducen las opciones de configuración comunes a todos los procesos que forman parte de cada proyecto.

En la figura 3.7 aparece una pantalla de ejemplo con el listado de proyectos en los que está trabajando un usuario concreto. Los proyectos aparecen ordenados por fecha y pueden editarse y borrarse. Así un usuario puede trabajar en distintos proyectos dentro de un mismo dominio.

Configuración de Proyectos
Añade los proyectos en los que vas a trabajar.

NUEVO PROYECTO
LISTAR MIS PROYECTOS

AÑADIR NUEVO PROYECTO

Nombre Proyecto 1 - Opiniones sobre películas cinematográficas

Descripción Proyecto consistente en la clasificación sentimental de opiniones de usuarios vertidas en redes sociales, blogs y foros sobre películas cinematográficas de actualidad.

Fecha 03/10/2014

Cancelar Guardar

Figura 3.6. Pantalla del menú de configuración de proyectos para añadir un nuevo proyecto.

Configuración de Proyectos
Lista el conjunto de proyectos personales en los que estás trabajando.

NUEVO PROYECTO
LISTAR MIS PROYECTOS

LISTADO DE MIS PROYECTOS

01/07/2014	Proyecto sobre opiniones en finanzas - En este proyecto pretendemos llevar a cabo un estudio sobre las opin...
28/06/2014	Proyecto sobre opiniones en coches - En este proyecto pretendemos llevar a cabo un estudio sobre las opini...
15/06/2014	Proyecto sobre opiniones en smartphones - En este proyecto pretendemos llevar a cabo un estudio sobre las ...
12/05/2014	Proyecto sobre opiniones en películas - En este proyecto pretendemos llevar a cabo un estudio sobre las opi...
01/04/2014	Proyecto sobre opiniones en hoteles - En este proyecto pretendemos llevar a cabo un estudio sobre las opini...

« 1 2 3 »

Figura 3.7. Pantalla del menú de configuración de proyectos para listar los proyectos de un usuario.

La otra opción del menú de configuración es la definición de las opciones generales para cada proyecto. En esta opción de menú será donde se indique para cada proyecto concreto el conjunto de informaciones y valores de parámetros comunes a todos los procesos que formaran parte de este proyecto. En la figura 3.8 se muestra la pantalla de configuración de opciones generales de proyectos. Como se puede apreciar en la figura, lo primero que se selecciona es el proyecto para el que se van a establecer sus opciones de configuración generales. A continuación, se introduce la ruta en la que se encuentra la ontología del dominio que se va a emplear para obtener las relaciones semánticas entre los conceptos del dominio. Seguidamente, se introduce la ruta en la que se encuentra el corpus con las opiniones de los usuarios que se quieren clasificar. A continuación se indica dónde están alojadas las bases de datos de sentimientos y de sinónimos, que el sistema empleará como recursos lingüísticos dentro del proceso de

análisis de sentimientos de opiniones basado en características. Por último, se introduce la ruta en la que el sistema irá dejando un registro del estado de todas las acciones que lleva a cabo.

OPCIONES DE CONFIGURACIÓN

Proyecto Proyecto 1 - Opiniones sobre películas cinematográficas

Descripción Proyecto Proyecto consistente en la clasificación sentimental de opiniones de usuarios vertidas en redes sociales, blogs y foros sobre películas cinematográficas de actualidad.

Ontología C:\DATOS\Ontologías\MovieOntology.owl Examinar...

Descripción Ontología Ontología que modela los conceptos y relaciones existentes en el dominio de las películas cinematográficas. Esta ontología ha sido validada por expertos del dominio cinematográfico y está correctamente poblada.

Corpus dir C:\DATOS\Corpus\DominioPelículas\ Examinar... Idioma

Descripción Corpus Corpus sobre opiniones de películas cinematográficas que recoge 10.000 opiniones de usuarios, de las cuales 5000 están clasificadas como positivas y 5000 como negativas.

SentiWordNet dir C:\DATOS\BasesDatos\SentiWordNet\ Examinar...

WordNet dir C:\DATOS\BasesDatos\WordNet\ Examinar...

Logs dir C:\DATOS\LogsSistema\ Examinar...

Cancelar Guardar

Figura 3.8. Pantalla del menú de configuración de proyectos para establecer las opciones generales de configuración asociadas a cada proyecto.

3.3.2.3 Procesos del sistema

La siguiente opción de menú es la pestaña de procesos del sistema. En ella se pueden llevar a cabo cada uno de los procesos de análisis de sentimientos de opiniones que se han propuesto en la metodología expuesta en esta tesis doctoral. Estos procesos son los siguientes: análisis morfológico, identificación de características, cálculo de polaridades de características y análisis de sentimientos de opiniones. La figura 3.9 muestra el desplegable de la pestaña de procesos del sistema.

En la figura 3.9, en el desplegable de la opción de menú de procesos del sistema, se incluye también un apartado destinado a la inicialización de dichos procesos. El sentido de esta opción es configurar los parámetros comunes a todos los procesos del sistema para que no haya necesidad de repetirlos después dentro de cada proceso. Esto permite usar el sistema con agilidad sin la necesidad de estar introduciendo los mismos datos para cada uno de los procesos en los que participa el sistema. Asimismo, dentro de la pantalla de configuración de cada proceso del sistema se muestran seleccionadas las opciones por defecto del sistema con la posibilidad de cambiarlas.



Figura 3.9. Pantalla del menú de procesos del sistema software de clasificación de sentimientos de opiniones.

La figura 3.10 muestra el aspecto de la pantalla de inicialización de procesos del sistema. En esta pantalla se presenta el proyecto actual en el que se esté trabajando permitiendo seleccionar otro proyecto. En la misma pantalla se configuran los valores para los parámetros Z_1 , Z_2 y Z_3 , que establecen la importancia de cada característica identificada por el sistema en función de la zona de la opinión del usuario en la que aparece (Z_1 para las ocurrencias en la parte inicial de la opinión, Z_2 para las ocurrencias en la parte media y Z_3 para las ocurrencias al final de la opinión). Seguidamente se establece también el valor para el parámetro del sistema N_GRAM , que será común a todos los procesos del sistema para ese proyecto. Este parámetro indica el número de

palabras próximas al término lingüístico que identifica cada característica en la opinión de un usuario a la hora de llevar a cabo la clasificación de sentimientos, tanto de las características identificadas como de las opiniones en su totalidad. Por último, se elige el método de clasificación de sentimientos: *N_GRAM Before*, *N_GRAM After*, *N_GRAM Around* y *All Phrase*.

Figura 3.10. Pantalla del menú de procesos del sistema para establecer el valor de los parámetros comunes a los procesos del sistema.

El primer proceso que se lleva a cabo en el sistema es el análisis morfológico de las opiniones de los usuarios. En la figura 3.11 se muestra el aspecto de la pantalla con los datos que requiere el sistema para ejecutar este proceso. En primer lugar, aparece seleccionado el corpus de opiniones configurado por defecto para el proyecto actual en el que estamos trabajando (y que se guardará en la sesión de trabajo del usuario hasta que éste decida seleccionar otro). Una vez elegido el corpus con el que se va a trabajar, se introduce la ruta en la que se guardarán los ficheros con los resultados concernientes a este proceso de análisis morfológico, esto es, un corpus con las opiniones de los usuarios etiquetadas. El proceso arrancará cuando el usuario pinche en el botón ‘Comenzar’. El sistema notificará mediante un mensaje en pantalla cuando el proceso haya finalizado y los resultados estén listos para su consulta.

Procesos

Lleva a cabo el proceso de análisis morfológico y sintáctico del corpus de opiniones de entrada.

Inicialización

Proceso - Análisis Morfológico

Proceso - Identificación de Características

Proceso - Cálculo Polaridad de Características

Proceso - Análisis Sentimental de Opiniones

Proceso - Análisis Morfológico

Directorio de Entrada

C:\DATOS\Corpus\DominioPelículas\

Examinar...

Idioma ▼

Descripción

Corpus sobre finanzas que recoge 10.000 opiniones de usuarios, de las cuales 5000 están clasificadas como positivas y 5000 como negativas.

Directorio de Salida

C:\DATOS\Salida\DominioPelículas\Proceso-Análisis-Morfológico-Sintáctico\

Examinar...

Cancelar

Comenzar

Figura 3.11. Pantalla del menú de procesos del sistema para llevar a cabo el proceso de análisis morfológico.

El siguiente proceso que puede llevar a cabo el sistema es el de identificación de características. En la figura 3.12 se muestra la pantalla con los datos se necesitan para llevarlo a cabo. En primer lugar, el sistema requiere un directorio de entrada en el que se encuentre el corpus etiquetado de las opiniones de los usuarios (el corpus inicial del proyecto después de aplicarle el proceso de análisis morfológico). A continuación, debemos indicarle al sistema dónde encontrar la ontología del dominio (aparecerá seleccionada por defecto la ontología del dominio asociada al proyecto en el que se esté trabajando). Por último, se le indica al sistema el directorio en el que colocar los resultados del proyecto. El proceso arrancará cuando el usuario pinche en el botón ‘Comenzar’. El sistema notificará mediante un mensaje en la pantalla cuando el proceso haya finalizado y los resultados estén listos para su consulta.

Procesos

Lleva a cabo el proceso de identificación de características en el corpus etiquetado.

Inicialización

Proceso - Análisis Morfológico

Proceso - Identificación de Características

Proceso - Cálculo Polaridad de Características

Proceso - Análisis Sentimental de Opiniones

Proceso - Identificación de Características

Directorio de Entrada

C:\DATOS\Salida\DominioPelículas\Proceso-Análisis-Morfológico-Sintáctico\

▼

Ontología

C:\DATOS\Ontologías\MovieOntology.owl

Examinar...

Descripción Ontología

Ontología que modela los conceptos y relaciones existentes en el dominio de las películas cinematográficas. Esta ontología ha sido validada por expertos del dominio cinematográfico y está correctamente poblada.

Directorio de Salida

C:\DATOS\Salida\DominioPelículas\Proceso-Identificación-Características\

Examinar...

Cancelar

Comenzar

Figura 3.12. Pantalla del menú de procesos del sistema para llevar a cabo el proceso de identificación de características.

En la figura 3.13 se muestra la pantalla con los datos que le hacen falta al sistema para ejecutar el proceso de cálculo de la polaridad de las características identificadas en el proceso anterior. Es decir, el sistema va a proceder a clasificar los sentimientos asociados a cada una de las características encontradas en las opiniones. En primer lugar, tendremos que indicar al sistema el directorio de entrada en el que se encuentran los resultados de la identificación de características. A continuación, el sistema necesitará conocer la ubicación de las bases de datos de sentimientos y de sinónimos, que se utilizan en el cálculo de la polaridad de características y en la agrupación de características sinónimas respectivamente. En este caso, aparecen seleccionadas por defecto las bases de datos asociadas al proyecto en el que estamos trabajando para ahorrarnos el tener que introducir las rutas de nuevo. De manera análoga aparecen establecidos los valores de los parámetros Z_1 , Z_2 y Z_3 asociados al proyecto actual y cuyo significado se ha explicado antes. Estos valores se pueden variar al igual que el siguiente parámetro requerido por el sistema (N_GRAM). Por último, se elige el método de clasificación de sentimientos de características a emplear y el directorio donde almacenar los resultados de la clasificación. El proceso arrancará cuando el usuario pinche en el botón ‘Comenzar’. El sistema notificará mediante un mensaje en la pantalla cuando el proceso haya finalizado y los resultados estén listos para su consulta.

Figura 3.13. Pantalla del menú de procesos del sistema para llevar a cabo el proceso de clasificación de sentimientos de características.

En la figura 3.14 se puede ver un ejemplo de consulta de los resultados obtenidos por el sistema para el proceso de cálculo de polaridades de características. En primer lugar aparece el nombre del proyecto actual, y a continuación un área de texto que carga los resultados.

El último proceso a llevar a cabo dentro del sistema es el proceso de análisis de sentimientos de las opiniones de los usuarios. En la figura 3.15 aparecen las informaciones que se debe proporcionar al sistema para que éste pueda llevar a cabo este proceso. En primer lugar se proporciona al sistema el directorio que contiene los resultados del cálculo de la polaridad de características llevado a cabo en el proceso anterior. Una vez hecho esto, el sistema cargará de forma dinámica el conjunto de características identificadas en cada opinión de cada usuario junto a su vector de polaridad asociado. Esto se hace a modo de consulta, por si se desea hacer un cambio de última hora en procesos anteriores antes de llevar a cabo el proceso de análisis de sentimientos de la totalidad de las opiniones de los usuarios. Por último, elegiremos el directorio donde queremos guardar los resultados de la clasificación de sentimientos de las opiniones de los usuarios expresadas en lenguaje natural. El proceso arranca cuando el usuario pulsa en el botón ‘Comenzar’, y el sistema notifica mediante un mensaje en la pantalla cuando el proceso haya finalizado y los resultados estén listos para su consulta.

Por último, en la figura 3.16 aparece un ejemplo de resultados obtenidos por el sistema para el proceso de análisis de sentimientos de opiniones basado en características. En este análisis de sentimientos se clasifican los sentimientos de cada una de las opiniones presentes en el corpus y se proporciona una salida detallada de todo el proceso.

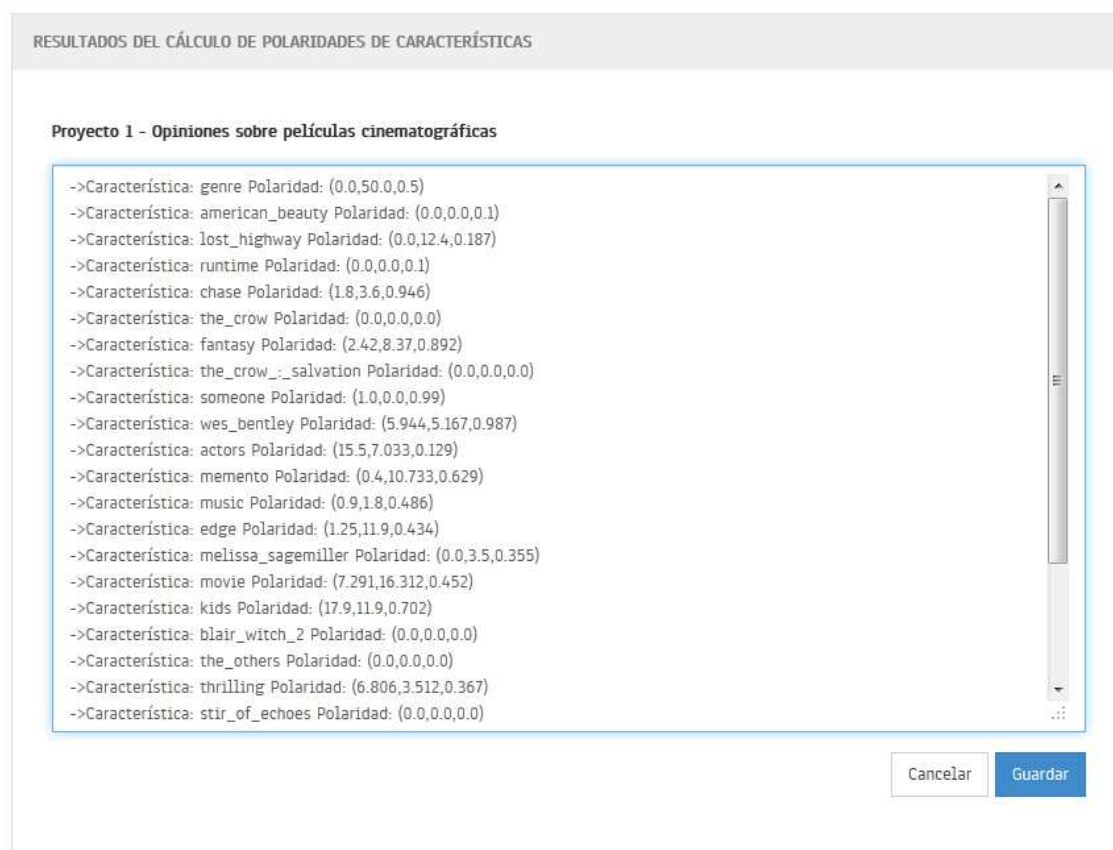


Figura 3.14. Pantalla del menú de procesos del sistema para consultar los resultados obtenidos en la clasificación de sentimientos de características.



Figura 3.15. Pantalla del menú de procesos del sistema para llevar a cabo el proceso de análisis de sentimientos de opiniones.

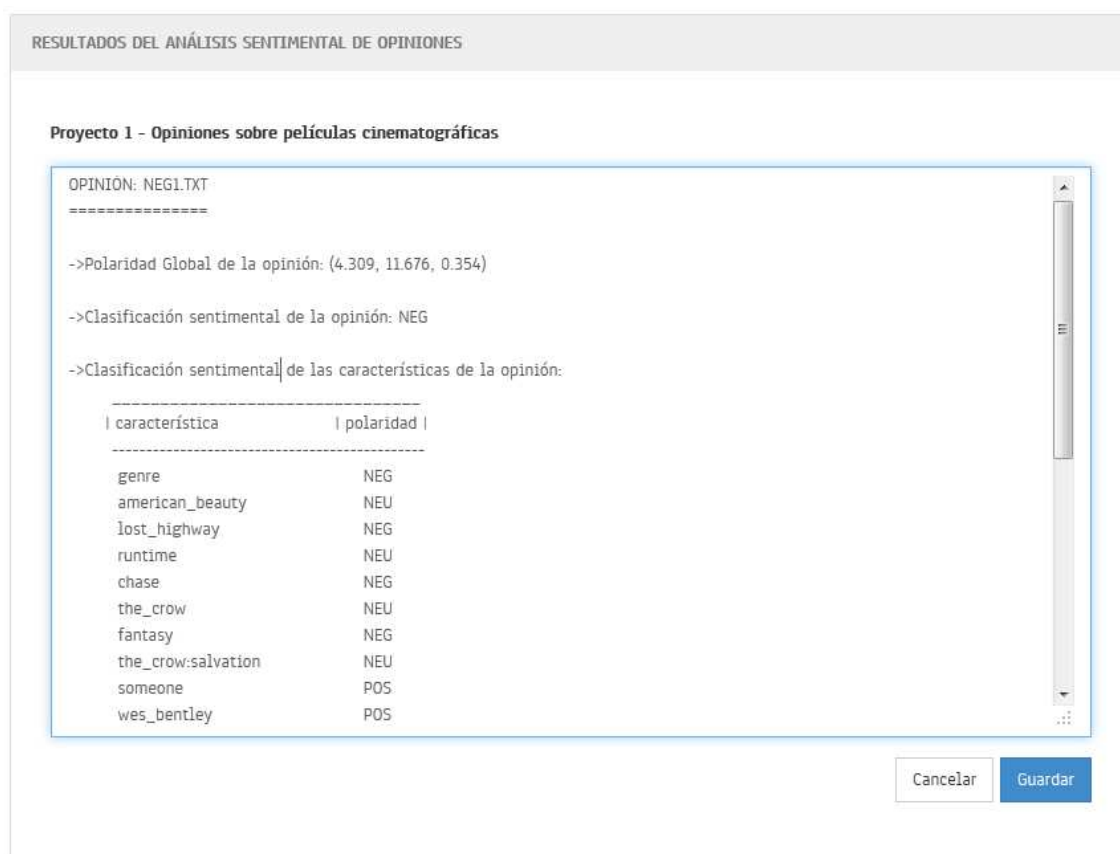


Figura 3.16. Pantalla del menú de procesos del sistema para consultar los resultados obtenidos en la clasificación de sentimientos de opiniones de usuarios.

3.3.2.4 Idiomas de la interfaz del sistema

La interfaz del sistema ha sido implementada para los idiomas español e inglés. Los idiomas en los que se presenta la interfaz (lenguajes en los que es posible presentar al usuario las informaciones para utilizar el sistema) han sido elegidos para dar la mayor cobertura posible a usuarios de cualquier parte del mundo que quieran utilizar la herramienta.

Además, como se ha apuntado anteriormente, el software empleado en el proceso de análisis morfológico (*Stanford Log-linear Part-Of-Speech Tagger*) soporta los idiomas: árabe, chino, francés, español y alemán.

3.3.2.5 Ayuda del sistema

La ayuda del sistema proporcionará al usuario un tutorial dinámico de uso del sistema con ejemplos prácticos de funcionamiento y una sección *How To*. El usuario podrá

descargar manuales de uso de la herramienta. También se proporciona información sobre la versión del software que hay desplegada (submenú ‘Acerca de...’) y una parte de agradecimientos hacia todas las personas y entidades que han facilitado la implementación de este sistema de clasificación de opiniones de usuarios expresadas en lenguaje natural.

3.4 Resumen

En este capítulo se han presentado los fundamentos tecnológicos necesarios para crear un sistema de software que lleve a cabo la minería de opiniones de usuarios a partir de texto en lenguaje natural. Dicha herramienta cumple exhaustivamente con los requisitos de la metodología de análisis de sentimientos propuesta en el Capítulo II de esta tesis doctoral.

En el análisis, diseño e implementación de la herramienta cuyo funcionamiento se ha explicado en este capítulo, se han utilizado herramientas libres disponibles para la comunidad investigadora. Además, se han seguido las buenas prácticas del desarrollo de software orientado a objetos y se han utilizado patrones de diseño, lo cual facilita la adaptabilidad a los cambios y el mantenimiento de la herramienta.

A continuación se resumen las partes más destacadas del sistema de análisis de sentimientos basado en características y guiado por ontologías implementado en este capítulo:

- **Acceso al sistema y control de usuarios.** El sistema requiere la autenticación de cada usuario que vaya a trabajar con los procesos del mismo. De esta manera, se abre una sesión individual por cada usuario autenticado en el sistema y se permite el trabajo en paralelo con la herramienta. Esta forma de trabajar permite a los usuarios el acceso a la herramienta desde cualquier parte del mundo y sin necesidad de instalación de la misma. Por último, esta forma de diseñar la aplicación permite que cada usuario gestione sus propios proyectos y pueda guardar sus propias configuraciones para trabajar con los procesos del sistema.
- **Configuración del sistema.** Representa la configuración inicial necesaria para ejecutar todos los procesos del sistema. Donde se realiza la gestión de los proyectos personales de cada usuario del sistema y se configuran las rutas de acceso a los distintos recursos lingüísticos necesarios para llevar a cabo las funcionalidades del sistema. Se debe establecer la ubicación del corpus de

opiniones de los usuarios, la ubicación de la base de datos con valores de sentimientos asociados a palabras, la ubicación de la base de datos de sinónimos, y la ubicación de la ontología del dominio, para cada uno de los proyectos que se definan.

- **Proceso de análisis morfológico.** Representa la funcionalidad de procesamiento del lenguaje natural sobre las opiniones de los usuarios. Se trata de etiquetar morfológicamente el corpus de opiniones de partida. En este proceso se identificarán los nombres, adjetivos, adverbios, verbos, etc. de las opiniones de los usuarios y la salida del mismo será un corpus de opiniones etiquetado.
- **Proceso de identificación de características.** Representa la búsqueda de las características presentes en las opiniones de los usuarios. Para llevar a cabo este proceso el sistema se ayudará de la estructura semántica de la ontología del dominio. La salida de este proceso será un conjunto de características identificadas en cada una de las opiniones de los usuarios.
- **Proceso de clasificación de sentimientos de características.** Representa el proceso de cálculo de la polaridad de cada una de las características identificadas en cada opinión de usuario. Para llevar a cabo este proceso el sistema utiliza una base de datos de valores de sentimientos y (opcionalmente) una base de datos de sinónimos (para agrupar características sinónimas). La salida de este proceso será un conjunto de características clasificadas en función de sus vectores de polaridad asociados.
- **Proceso de análisis de sentimientos de opiniones de usuarios.** Representa el proceso de clasificación de sentimientos de las opiniones de los usuarios expresadas en lenguaje natural. Para llevar a cabo este proceso el sistema partirá de la lista de características clasificadas en el punto anterior. La salida de este proceso será la clasificación de sentimientos de cada opinión vertida por los usuarios.

Para todos los procesos existe una funcionalidad común: guardar los resultados obtenidos permitiendo recuperarlos al inicio de una nueva sesión de trabajo. El sistema presenta un modo de trabajo individualizado para cada usuario de tal manera que cada uno de ellos se identifica al inicio, y a partir de ahí puede trabajar con la herramienta definiendo sus propios proyectos y configuraciones. Por último, cabe destacar que los idiomas de la interfaz de la herramienta son dos, el español y el inglés. Sin embargo, la aplicación permite trabajar en cualquier idioma o dominio, siempre que se le configuren apropiadamente los recursos lingüísticos necesarios.

Capítulo IV

Validación de la metodología de
minería de opiniones basada en
características guiada por
ontologías

4.1 Introducción

Este capítulo se centra en la validación de la metodología de anotación desarrollada en esta tesis. Para la validación se han utilizado las medidas de precisión, exhaustividad y medida-F. Estas medidas, que se describen de forma detallada en este capítulo, se utilizan en la validación de sistemas de procesamiento del lenguaje natural y recuperación de información (Rodríguez-García et al., 2014).

En la elaboración de esta tesis doctoral se ha hecho hincapié en el perfeccionamiento de la precisión de la solución planteada. Para ello, en la validación de la metodología, se determina la precisión en la identificación de las características, en el cálculo de la polaridad de cada una de ellas y en el cálculo de la polaridad global de la opinión de un usuario.

La validación se ha realizado en dos dominios bien diferenciados: el cinematográfico y el financiero. Con esto se persigue demostrar la utilidad de la metodología desarrollada para la resolución de los distintos tipos de problemas que se pueden presentar en dominios tan divergentes como los indicados y resaltar una de las propiedades más importantes de esta metodología: la independencia del dominio.

Para el experimento realizado en el dominio de las películas cinematográficas ha sido necesario recopilar un corpus de opiniones de los usuarios sobre películas de cine. Este corpus incluye referencias sobre los actores que intervienen, la calidad de la música de la película, la calidad de la trama, la capacidad de sorprender del director, el guion, la comparación con otras películas del mismo género, etc. El problema principal radica en que las opiniones de los usuarios suelen ser muy variables y, en muchas ocasiones, llenas de ironía. En términos generales el lenguaje empleado por los usuarios para este dominio suele ser bastante informal, incluidas muchas licencias gramaticales. La dificultad está en captar el sentido exacto que quiere dar el usuario a la frase empleando técnicas de procesamiento del lenguaje natural y los recursos lingüísticos de los que se dispone actualmente.

El segundo experimento está relacionado con el dominio financiero donde se ha elaborado el corpus a partir de columnas de opinión dadas por expertos en economía en sitios Web del ámbito de la economía, los negocios y las finanzas. El problema principal en un dominio de estas características es que los conceptos no están del todo claros y no suelen ser muy específicos. En particular, en este dominio se suelen utilizar conceptos que requieren interpretación adicional.

El capítulo comienza con la definición de las medidas de evaluación empleadas para validar la metodología propuesta. A continuación se describen los procesos que constituyen el módulo de validación de resultados construido para validar el sistema de software implementado. Seguidamente se muestran los resultados obtenidos en los dominios cinematográfico y financiero. Finalmente, se presentan las conclusiones de la validación de forma individualizada para cada uno de los dominios presentados y una comparativa entre ambos dominios.

4.2 Medidas de evaluación

La evaluación del rendimiento del sistema propuesto en esta tesis doctoral ha sido llevada a cabo mediante la utilización de un conjunto de métricas de evaluación estándar: precisión, exhaustividad y medida-F. Estas métricas, que son de aplicación muy común en los procesos de evaluación de sistemas de procesamiento del lenguaje natural y de recuperación/extracción de información, fueron propuestas inicialmente por Salton en 1983 (Salton & McGill, 1983).

Aunque estas métricas estaban destinadas a la medida del rendimiento de sistemas de búsqueda y recuperación de información y reconocimiento de patrones, actualmente son varias las áreas de investigación que las han adaptado para medir el rendimiento en otros contextos como la oftalmología (Miliotis et al., 2003), biomedicina (Krauthammer & Nenadic, 2004) y bioquímica (Alexopoulou et al., 2008), entre otros.

4.2.1 Definición de las medidas de evaluación estándar

Se puede definir la precisión como *“la fracción de una salida de búsqueda que es relevante para una consulta determinada. Su cálculo, por lo tanto, requiere el conocimiento de los éxitos relevantes y no relevantes en el conjunto de documentos evaluados”* (Clarke & Willett, 1997).

Según Salton y McGill (1983) la precisión se define como *“una medida de exactitud”* y *“determina la fracción de entidades relevantes de todas las entidades recuperadas en un sistema de extracción de información”*. Teniendo en cuenta esta definición la precisión se puede calcular a partir de la fórmula 4.1.

$$precisión = \frac{|entidades\ extraídas| \cap |entidades\ relevantes|}{|entidades\ extraídas|} \quad (4.1)$$

La exhaustividad se define como “la capacidad que un sistema de recuperación tiene para obtener todos o la mayoría de los documentos relevantes en una colección. Por lo tanto, se requiere del conocimiento de los documentos no sólo relevantes y recuperados, sino también, de aquellos que no fueron recuperados” (Clarke & Willett, 1997).

Sin embargo en el contexto del presente estudio, la exhaustividad es una medida de la integridad y determina la fracción de entidades relevantes recuperadas de todas las entidades relevantes. De acuerdo con esta definición, se puede calcular a partir de la fórmula 4.2.

$$exhaustividad = \frac{|entidades\ extraídas| \cap |entidades\ relevantes|}{|entidades\ relevantes|} \quad (4.2)$$

Mientras la precisión representa una medida de exactitud o fidelidad, la exhaustividad representa una medida de integridad.

El rendimiento global de estos dos métodos es evaluado por la tercera métrica, que se conoce como medida-F y se define como “la media armónica de los valores de precisión y exhaustividad” (Yang & Liu, 1999).

En concreto, la fórmula de la medida-F se utiliza para evaluar el rendimiento global de las dos métricas propuestas anteriormente. De acuerdo con esta definición, la medida-F se puede obtener a partir de la fórmula 4.3. El resultado de la fórmula proporciona un valor real entre 0 y 1.

$$medida - F = 2 * \frac{precisión * exhaustividad}{precisión + exhaustividad} \quad (4.3)$$

El sistema propuesto en esta tesis doctoral se va a validar utilizando estas tres métricas en cada una de las fases que forman parte de la metodología: la fase de identificación de características, de cálculo de la polaridad de las características identificadas y de clasificación de sentimientos de las opiniones de los usuarios.

4.2.2 Medidas de evaluación para la fase de identificación de características

Para la fase de identificación de características los elementos que estarán sujetos al proceso de validación son los siguientes: el número de ‘características relevantes’ y el número de características identificadas. Las características relevantes hacen referencia a las características que serían identificadas por un experto humano dentro de una opinión, mientras que las ‘características identificadas’ son aquellas que el sistema identifica cuando se ejecuta.

El objetivo es conocer si el sistema está proporcionando buenos resultados, si está identificando buenas características. Si el sistema obtiene un buen número de características que resultan relevantes se podrá concluir que la ontología del dominio que se ha empleado está muy bien construida y que ha sido poblada de manera correcta. Para ello, se adaptarán las métricas de precisión y exhaustividad anteriormente definidas para poder obtener los valores asociados con la identificación de características.

La fórmula matemática para la precisión en el proceso de identificación de características quedaría representada en esta fase como aparece en la fórmula 4.4.

$$precisión = \frac{|características\ identificadas| \cap |características\ relevantes|}{|características\ identificadas|} \quad (4.4)$$

Por su parte, la fórmula matemática para el exhaustividad quedaría representada tal y como aparece en la fórmula 4.5.

$$exhaustividad = \frac{|características\ identificadas| \cap |características\ relevantes|}{|características\ relevantes|} \quad (4.5)$$

El sistema determina si cada palabra de las que aparecen en el corpus de opiniones de los usuarios puede ser una posible característica. Para ello, el sistema comprueba si la expresión lingüística asociada a la característica está presente en el conjunto de clases de la ontología, o es una instancia de una clase de la ontología del dominio (también llamada *individual*). En ambos casos la metodología propuesta tiene en cuenta el uso de las expresiones lingüísticas asociadas a las posibles características en singular y en plural, así como haciendo uso de las expresiones sinónimas.

4.2.3 Medidas de evaluación para la fase de cálculo de la polaridad de las características

Para evaluar el cálculo de la polaridad asociada a cada característica identificada por el sistema dentro de la opinión expresada por un usuario, será necesario conocer la polaridad correcta asignada de manera manual a cada característica por un experto humano. El proceso de validación llevará a cabo una comparación entre las polaridades asignadas a las características por el experto y las polaridades de las características obtenidas usando nuestra metodología. Por tanto, se necesita un *corpus* con las opiniones de los usuarios etiquetadas manualmente, de tal manera que se disponga de unas polaridades de características asignadas de manera manual. Una vez hecho esto, el proceso automático de validación contará el número de características clasificadas correctamente con respecto al número total de características identificadas.

La fórmula matemática para calcular la precisión en el proceso de identificación de características quedaría adaptada para esta fase como aparece en la fórmula 4.6.

$$\text{precisión} = \frac{| \text{características con polaridad correcta} |}{| \text{características identificadas} |} \quad (4.6)$$

4.2.4 Medidas de evaluación para la fase de cálculo de la polaridad global de las opiniones

De manera análoga, en la fase validación del cálculo de la polaridad global asociada a las opiniones de los usuarios, se partirá de la polaridad global asignada a cada opinión por un experto humano. El proceso de validación llevará a cabo una comparación entre las polaridades asignadas por el experto a las opiniones y las polaridades de las opiniones obtenidas usando el sistema. Por tanto, el proceso automático de validación contará el número de aciertos que obtiene el sistema a la hora de clasificar la polaridad de las opiniones con respecto al número total de opiniones.

La fórmula matemática para calcular la precisión en el proceso de cálculo de la polaridad global de las opiniones de los usuarios queda representada en la fórmula 4.7.

$$\text{precisión} = \frac{| \text{opiniones con polaridad correcta} |}{| \text{opiniones} |} \quad (4.7)$$

4.3 Módulo de validación de resultados

La figura 4.1 muestra un diagrama de procesos (siguiendo con las convenciones gráficas mostradas en el Capítulo II) donde aparecen cada uno de los elementos de entrada, elementos internos y elementos de salida que juegan un papel relevante en el proceso de validación de resultados. Como se puede apreciar en la figura, el proceso de validación de resultados parte de los resultados obtenidos por el proceso principal del sistema, esto es, el proceso de análisis de sentimientos. Una vez que el proceso de validación tiene en su poder las características identificadas y clasificadas, además de la clasificación global de cada una de las opiniones de los usuarios, está preparado para llevar a cabo su trabajo. En concreto, el proceso de validación utiliza un corpus de opiniones con las características identificadas y clasificadas de manera manual por un experto. Con esta información de entrada, se obtienen los datos de validación siguientes: resultados para la identificación de características, resultados para la clasificación de características y resultados para la clasificación de opiniones.

Además, destacar que es necesario tomar en cuenta el dominio en el que se está realizando la investigación. Los resultados de la identificación de características dependerán, en buena medida, de que los conceptos estén mayoritariamente bien definidos en el dominio en cuestión y que, por lo tanto, sean fácilmente asociables a los conceptos de la ontología que modela ese dominio. Por ejemplo, identificar las características en el dominio financiero será muy complejo, ya que es un dominio sujeto a muchas interpretaciones y, por lo tanto, una mayor ambigüedad.

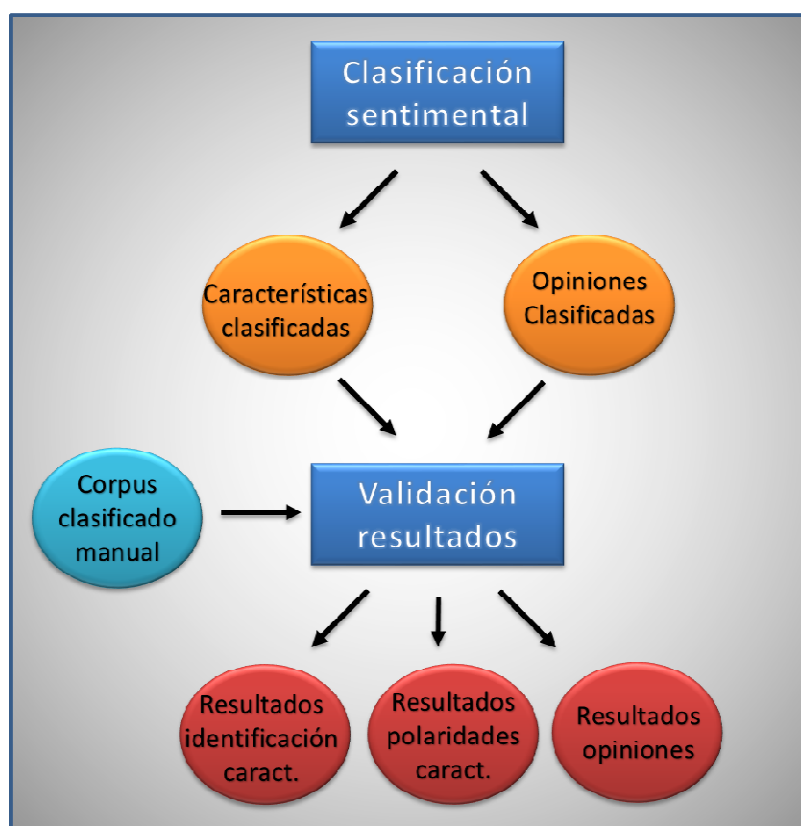


Figura 4.1. Diagrama de procesos de la validación del sistema.

4.4 Validación en el dominio de las películas cinematográficas

4.4.1 Introducción

“Las imágenes son lo primero, y con imágenes, al igual que con la música, la principal reacción es emocional”, dijo en una ocasión el difunto director de cine Richard Brooks.

El cine es arte. El cine compendia y se basa en todas las artes: las utiliza, trasvasa y recrea, necesita de ellas, las mejora y las difunde. Sin la literatura y los escritores, sean de novela, cuento, guion o poesía, el cine no tendría argumentos. Sin la fotografía, la pintura, la escultura y la arquitectura, no tendrían soporte estético ni justificación teórica. Sin la música y la danza, la luz o el color no podría expresarse en su plenitud. Sin las ciencias, la física y la química, la tecnología o la informática, el cine no tendría base material en que sustentarse. El cine, además, conduce a la tecnología hacia el arte, reproduce la luz y el color y eleva el movimiento y el ritmo a las alturas de las artes llamadas «nobles», para generar la fantasía, la ficción y la realidad (Gubern, 2014).

Las películas no son sólo entretenimiento, son una vuelta en la montaña rusa de las emociones que vive el espectador en la oscuridad de la sala de cine. Una de las realidades evidentes que a menudo se pasan por alto es que una película primero triunfa o fracasa a la luz candente del mercado (Fernández, 2014). ¿Pagará o no el público para ir a verla? Esa es la cuestión que se preguntan los magnates de la industria cuando consideran un proyecto de película, y es la clave para entender la industria cinematográfica.

En este contexto, se hace patente la utilidad de un sistema que analice los sentimientos o emociones presentes en las opiniones de los usuarios cuando han visto una película. En una industria (que mueve alrededor de 90 millones de dólares cada año, según *el economista*⁴²) es fundamental conocer la opinión de sus clientes para seguir creciendo cada año, facilitar la toma de decisiones por parte de productores y directores y ayudar a ofrecer productos mejor valorados.

Actualmente, existen varios trabajos sobre *opinión mining* que han centrado sus investigaciones en el ámbito de las películas cinematográficas (Deng et al., 2014; Martín-Valdivia et al., 2013; Singh et al., 2013). En el trabajo de Martín-Valdivia (2013) se propone el uso de meta-clasificadores que combinan aprendizaje supervisado y no supervisado para construir un sistema de clasificación de sentimientos. Sin embargo, esa propuesta se comprende de un sistema de clasificación de sentimientos semi-automático que requiere la intervención de un experto con el fin de obtener resultados. El enfoque presentado en esta tesis doctoral no requiere de intervención humana por lo que puede considerarse y clasificarse con un método totalmente automático.

Una de las razones principales por las que se ha elegido este dominio es su complejidad y completitud. En particular, el conocimiento del ámbito cinematográfico es esencial para comprender la terminología y las estructuras cinematográficas.

El análisis de toda la información vertida en la Web con respecto a las opiniones de los usuarios sobre películas cinematográficas es un trabajo muy costoso y, por esta razón, el objetivo del sistema aquí presentado es ayudar a extraer de forma automática conocimiento relevante a partir de opiniones en lenguaje natural que puedan influir en las decisiones que los productores tienen que tomar en este dominio.

⁴² <http://www.eleconomista.es>, última consulta realizada en Abril de 2015.

4.4.2 Corpus

Para llevar a cabo la validación del sistema en este dominio se ha utilizado un corpus disponible en la Web compuesto de 200 documentos del dominio cinematográfico. Estos documentos contienen opiniones de usuarios sobre películas cinematográficas en inglés. A veces una opinión contiene una crítica sobre una película específica pero, en otras ocasiones, se vierten opiniones sobre varias películas en el mismo documento.

El corpus del experimento contiene 48.822 palabras distribuidas a lo largo de estos 200 ficheros de opinión. La mitad de ellos contiene opiniones que han sido etiquetadas por expertos humanos como positivas, y la otra mitad opiniones negativas. Esta clasificación obedece a un criterio experto en cuanto a la totalidad de la opinión de cada usuario, es decir, la polaridad global de esa opinión. No se hace ningún análisis de la polaridad a nivel de características. Por tanto, este trabajo adicional tendremos que aportarlo nosotros con el objetivo de validar la metodología de identificación de características y cálculo de polaridades de las mismas.

Se ha utilizado el corpus que aparece en (Pang & Lee, 2004)⁴³, un corpus muy completo y bien construido que se ha empleado en varios trabajos de investigación relevantes (Airoldi et al., 2005; Alekh et al., 2005; Beineke et al., 2004; Durbin et al., 2003; Gabrilovich et al., 2004; Hutton et al., 2012; Kim et al., 2012; Li et al., 2012; Mullen et al., 2004; Peng, 2003; Salvetti et al., 2004; Saif et al., 2012; Sista et al., 2004). Este corpus se ha afianzado como una especie de *benchmark* para probar metodologías de minería de opiniones. Los documentos están clasificados en dos directorios *POS* y *NEG*, lo que evita tener que etiquetar manualmente los documentos como paso previo a la validación de las metodologías.

Para este experimento no se ha considerado oportuno emplear todos los documentos de opinión de este corpus. El corpus es muy extenso y para validar la metodología propuesta es preciso llevar a cabo un proceso manual de etiquetado de cada una de las características que aparecen en cada fichero de opinión. Este proceso de etiquetado manual se ha llevado a cabo para 200 documentos, elegidos aleatoriamente del corpus completo de (Pang & Lee, 2004), lo que nos permitirá reunir los resultados de referencia. Estos resultados se emplearán para evaluar los resultados obtenidos con el sistema automático de análisis de sentimientos que proponemos en esta tesis.

⁴³ Corpus disponible en: <http://www.cs.cornell.edu/people/pabo/movie-review-data/>, última consulta realizada en Abril de 2015.

4.4.3 Ontología del dominio cinematográfico

La mejor opción para desarrollar la ontología para un dominio determinado es reutilizar una ontología existente y completarla e instanciarla adecuadamente. Para este experimento se ha reutilizado la ontología *The Movie Ontology (MO)*⁴⁴. Esta ontología tiene como objetivo proporcionar un vocabulario controlado con el que describir semánticamente las relaciones entre conceptos dentro del dominio de las películas cinematográficas. Conceptos como película, género, director, actor, etc., forman parte de esta ontología.

Los contenidos de la ontología *MO* vienen especificados originalmente en lenguaje *OWL*. Además, esta ontología ha sido revisada y validada por expertos en el dominio, que se han encargado de contrastar los conceptos, complementarlos y llevar a cabo las modificaciones pertinentes en las relaciones de los mismos, con el objetivo de obtener una ontología del dominio completa, bien estructurada y correctamente especificada. Para fines experimentales, se han insertado un total de 1.000 películas diferentes en la ontología *MO* como *individuals* utilizando la base de datos *Internet Movie Database (IMDB)*⁴⁵. Para llevar a cabo este trabajo de completar, modificar, integrar y poblar la ontología apropiadamente, se ha usado la herramienta de edición de ontologías *Protégé*.

4.4.4 Resultados para la búsqueda de características

Para evaluar el proceso de búsqueda de características se han utilizado las métricas de validación de precisión, exhaustividad y medida-F.

El proceso de búsqueda o identificación de características se lleva a cabo justo después del procesamiento del lenguaje natural de la opinión en texto libre hecha por el usuario. El proceso de PLN devuelve una lista de términos lingüísticos debidamente etiquetados (sustantivos, adjetivos, verbos, etc.). A continuación, el proceso de identificación de características discrimina entre todos estos términos aquellos que son características del dominio cinematográfico. Es en este punto donde entra en juego la ontología del dominio que, gracias a su estructura semántica y a su completitud en la población de la misma, aporta la información precisa para obtener todas las características relevantes dentro de las opiniones de los usuarios.

⁴⁴ La ontología del dominio de las películas cinematográficas *The Movie Ontology* esta disponible en <http://www.movieontology.org>. Última consulta en Enero de 2015.

⁴⁵ La base de datos en el dominio cinematográfico *IMDb* tiene sus datos disponibles en <http://www.imdb.com>. Última consulta en Enero de 2015.

Esta es una de las grandes ventajas de la metodología propuesta. Al utilizar una ontología del dominio validada por expertos, completa y correctamente poblada, y utilizar su potencia semántica en el proceso de identificación de características, se obtiene que las características identificadas son prácticamente iguales a las características relevantes en cada opinión de usuario.

Las tablas 4.1, 4.2 y 4.3 muestran los resultados experimentales obtenidos para el proceso de identificación de características. La primera fila en cada tabla muestra el número de opiniones utilizadas, la segunda fila el número de características que el sistema identificó, y la tercera fila el número de características reconocidas por un experto humano al leer las opiniones de los usuarios. Estas últimas características se llaman características relevantes.

Tabla 4.1. Resultados experimentales para la identificación de características en opiniones positivas en el dominio cinematográfico.

Opiniones Positivas	
Numero de Opiniones	100
Características identificadas por el sistema	1855
Características relevantes	1883
Precisión	1
Exhaustividad	0.985
Medida - F	0.992

Tabla 4.2. Resultados experimentales para la identificación de características en opiniones negativas en el dominio cinematográfico.

Opiniones Negativas	
Numero de Opiniones	100
Características identificadas por el sistema	1871
Características relevantes	1911
Precisión	1
Exhaustividad	0.979
Medida - F	0.989

Tabla 4.3. Resultados experimentales totales para la identificación de características en el dominio cinematográfico.

Opiniones Totales	
Numero de Opiniones	200
Características identificadas por el sistema	3726
Características relevantes	3794
Precisión	1
Exhaustividad	0.982
Medida - F	0.990

Como se puede apreciar en estas tablas, en términos generales el sistema obtiene buenos resultados para el proceso de identificación de características. A continuación, se detallan los resultados de precisión, exhaustividad y medida-F según han sido definidos anteriormente.

La precisión alcanzada tanto para el conjunto de opiniones de usuarios etiquetadas como positivas como las etiquetadas como negativas (y, por tanto, también para la unión de ambas) es igual a 1 dado que todas las características identificadas por el sistema son una clase, relación o instancia de la ontología del dominio y, por tanto, relevantes.

Se obtiene 0.985 de exhaustividad para opiniones positivas, lo cual es un resultado bastante prometedor. Algo parecido sucede con las opiniones etiquetadas como negativas, donde se obtiene una exhaustividad de 0.979. Por lo tanto, la medida-F es de 0.992 y 0.989 para las opiniones positivas y negativas respectivamente. Esto indica que el método de identificación de características utilizado es bueno, ya que, al igual que para la exhaustividad, los valores para la medida-F varían entre 0 y 1. Por último, juntando todo el corpus del experimento y calculando las métricas en cuestión, se obtiene una exhaustividad de 0.982 y una medida-F de 0.990, que confirma lo dicho anteriormente para opiniones positivas y negativas, es decir, usar una ontología del dominio en el proceso de identificación de características obtiene buenos resultados para este proceso.

En la figura 4.2 se muestra una comparativa visual de estos valores.

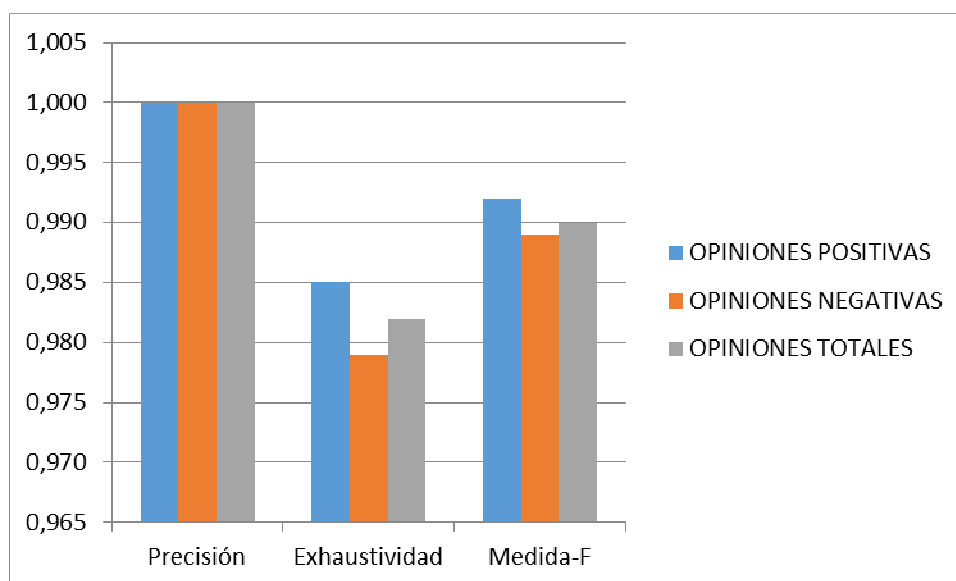


Figura 4.2. Comparativa visual de identificación de características en el dominio cinematográfico.

4.4.5 Resultados para la clasificación de sentimientos de características

Para evaluar el proceso automático de cálculo de la polaridad de las características se ha construido manualmente un fichero de validación. Este fichero contiene la clasificación de cada una de las características extraídas de las opiniones de los usuarios en una de las tres categorías siguientes: positiva (*POS*), negativa (*NEG*) o neutra (*NEU*). Este fichero se denomina “*label.txt*” y es utilizado como entrada del proceso de validación. En la tabla 4.4, es posible apreciar un extracto de la estructura de este fichero. En esta tabla se muestra que en la opinión contenida en el fichero “*neg1.txt*” la actriz “*Melissa Sagemiller*” tiene una evaluación negativa (“*NEG*”), pero la música de la película (“*music*”) es buena (“*POS*”). Además, el nombre del fichero de opinión (“*neg1.txt*”) indica que la polaridad global de esta opinión es negativa.

Tabla 4.4. Extracto del fichero de entrada para validación en el dominio cinematográfico.

Fichero de Opinión	Característica Identificada	Polaridad Manual
neg1.txt	genre	POS
neg1.txt	American Beauty	NEU
neg1.txt	Lost Highway	NEG
neg1.txt	runtime	NEG
neg1.txt	The Crow	NEU
neg1.txt	fantasy	NEG
neg1.txt	The Crow: Salvation	NEU
neg1.txt	Wes Bentley	NEG
neg1.txt	actors	POS
neg1.txt	Memento	NEG
neg1.txt	music	POS

neg1.txt	Melissa Sagemiller	NEG
neg1.txt	movie	NEG
neg1.txt	kids	NEU
neg1.txt	Blair Witch 2	NEU
neg1.txt	The Others	NEU
neg1.txt	thrilling	NEG
neg1.txt	Stir Of Echoes	NEU
neg1.txt	A Nightmare Of Elm Street 3	NEU
neg1.txt	horror	NEG
neg2.txt	Halloween H20	NEG
neg2.txt	Jamie Lee Curtis	POS
neg2.txt	movie	NEG
neg2.txt	Stan Winston	POS
neg2.txt	The Happy Bastard	NEG
neg2.txt	Baldwin	NEG
neg2.txt	action	NEG
neg2.txt	pink	NEU
neg2.txt	Donald Sutherland	NEG
neg2.txt	Y2k	NEG
neg2.txt	Schnazzy	POS

Una vez que el sistema ha llevado a cabo el proceso de determinar la polaridad de las características de cada opinión, se compararán las anotaciones manuales del fichero de validación de entrada “*label.txt*” con los resultados de salida producidos el sistema. El objetivo de esta comparación es obtener el número de características correctamente clasificadas por parte del sistema de minería de opiniones desarrollado.

En los experimentos realizados para esta fase, se han evaluado cuatro métodos diferentes para calcular la polaridad sentimental de las características identificadas dentro de las opiniones de los usuarios. Las diferencias entre estos métodos consisten únicamente en la forma en la que se recuperan las palabras adyacentes a cada característica para calcular la polaridad de ésta. Siendo necesario calcular la polaridad positiva, negativa o neutra de los nombres, adjetivos y verbos situados cerca de la expresión lingüística que representa a una característica dada en la opinión de los usuarios. Los cuatro métodos implementados son “*N_GRAM Before*”, “*N_GRAM After*”, “*N_GRAM Around*” y “*All Phrase*” (la descripción del funcionamiento de cada uno de estos métodos se encuentra en sección 2.6 de este documento). Los tres primeros métodos requieren del parámetro del sistema *N_GRAM*.

Para evaluar los métodos citados en el párrafo anterior, con el objetivo de calcular su precisión en el cálculo de la polaridad de características, es preciso determinar su eficiencia en el proceso de clasificación de sentimientos. Se ha experimentado con diferentes valores para el parámetro del sistema *N_GRAM* (entre 2 y 6) para descubrir la mejor configuración. Del mismo modo, también es necesario evaluar qué valores es recomendable asignar a los parámetros del sistema Z_1 , Z_2 y Z_3 . Z_i es un parámetro del

sistema que representa la importancia de la ocurrencia de una característica dada en una parte del texto (Z_1 para las ocurrencias en la parte inicial de la opinión, Z_2 para las ocurrencias en la parte media y Z_3 para las ocurrencias al final de la opinión). Se ha experimentado con diferentes valores para los parámetros Z_1 , Z_2 y Z_3 con el objeto de encontrar la configuración del sistema más eficiente para llevar a cabo la clasificación de sentimientos. En concreto, se han definido tres posibles valores: “*High*”, “*Medium*” y “*Low*”, que han sido mapeados a los valores numéricos 100, 50 y 25.

Resumiendo, es necesario conducir el experimento para que se puedan testar todas las posibilidades para los parámetros Z_1 , Z_2 y Z_3 y N_GRAM con los cuatro métodos de clasificación de sentimientos: “*N_GRAM Before*”, “*N_GRAM After*”, “*N_GRAM Around*” y “*All Phrase*”. Esto significa probar un total de 27 casos de estudio y 135 valores diferentes para todas las posibles combinaciones. Los resultados para el cálculo de la precisión en el cálculo de la polaridad de las características identificadas con estos cuatro métodos de clasificación de sentimientos se muestran a continuación (ver tablas 4.5, 4.6, 4.7 y 4.8). En las tablas de resultados, las abreviaciones “*H*”, “*M*” y “*L*” representan los valores “*High*”, “*Medium*” y “*Low*”, respectivamente, para los parámetros Z_1 , Z_2 y Z_3 , mientras que la abreviación “*NG=x*” indica que el valor x ha sido asignado al parámetro del sistema N_GRAM .

Las tablas 4.5, 4.6, 4.7 y 4.8 muestran los resultados obtenidos para la clasificación de sentimientos de las características identificadas por el sistema utilizando cada uno de los métodos mencionados anteriormente. En cada una de las tablas, se destacan en negrita los mejores resultados obtenidos. De esta forma será fácil identificar cuál es la mejor configuración de parámetros de entrada para cada uno de los métodos de clasificación de sentimientos. Además, en la figura 4.3 se presenta un gráfico que muestra una comparativa visual entre los distintos métodos de cálculo de la polaridad de las características identificadas.

Tabla 4.5. Valor de la precisión en el cálculo de la polaridad de las características con el método *N_GRAM Before*.

Casos de estudio				Precisión en el cálculo de la Polaridad de las Características					
Num.	Z ₁	Z ₂	Z ₃	NG=2	NG=3	NG=4	NG=5	NG=6	Media
1	H	H	H	73.05	71.11	72.51	75.82	78.43	74,18
2	H	H	M	73.05	71.11	72.51	75.82	78.43	74,18
3	H	H	L	73.05	71.11	72.51	75.82	78.43	74,18
4	H	M	H	73.05	71.11	72.51	75.82	78.43	74,18
5	H	M	M	73.05	71.11	72.51	75.82	78.43	74,18
6	H	M	L	73.05	71.11	72.51	75.82	78.43	74,18
7	H	L	H	73.05	71.11	72.51	75.82	78.43	74,18
8	H	L	M	73.05	71.11	72.51	75.82	78.43	74,18
9	H	L	L	73.49	71.11	72.51	75.82	78.43	74,27
10	M	H	H	73.05	71.11	72.51	75.82	78.43	74,18
11	M	H	M	73.05	71.11	72.51	75.82	78.43	74,18
12	M	H	L	73.05	71.11	72.51	75.82	78.43	74,18
13	M	M	H	73.05	71.11	72.51	75.82	78.43	74,18
14	M	M	M	73.05	71.11	72.51	75.82	78.43	74,18
15	M	M	L	73.05	71.11	72.51	75.82	78.43	74,18
16	M	L	H	73.05	71.11	72.51	75.82	78.43	74,18
17	M	L	M	73.05	71.11	72.51	75.82	78.43	74,18
18	M	L	L	73.05	71.11	72.51	75.82	78.43	74,18
19	L	H	H	72.83	70.91	72.51	75.82	78.43	74,10
20	L	H	M	72.83	70.91	72.51	75.82	78.43	74,10
21	L	H	L	72.83	70.91	72.51	75.82	78.43	74,10
22	L	M	H	73.05	71.11	72.51	75.82	78.43	74,18
23	L	M	M	73.05	71.11	72.51	75.82	78.43	74,18
24	L	M	L	73.05	71.11	72.51	75.82	78.43	74,18
25	L	L	H	73.05	71.11	72.51	75.82	78.43	74,18
26	L	L	M	73.05	71.11	72.51	75.82	78.43	74,18
27	L	L	L	73.05	71.11	72.51	75.82	78.43	74,18
Media				73.04	71.09	72.51	75.82	78.43	74,18

Tabla 4.6. Valor de la precisión en el cálculo de la polaridad de las características con el método *N_GRAM After*.

Casos de estudio				Precisión en el cálculo de la Polaridad de las Características					
Num.	Z ₁	Z ₂	Z ₃	NG=2	NG=3	NG=4	NG=5	NG=6	Media
1	H	H	H	60.31	59.56	58.75	59.82	60.94	59,88
2	H	H	M	68.31	67.05	66.47	67.82	68.94	67,72
3	H	H	L	67.59	66.33	66.47	66.98	68.37	67,15
4	H	M	H	68.31	67.56	66.75	67.82	68.94	67,88
5	H	M	M	68.31	67.05	66.75	67.82	68.94	67,77
6	H	M	L	68.31	67.05	66.47	67.82	68.94	67,72
7	H	L	H	68.31	67.56	66.75	67.82	68.94	67,88
8	H	L	M	68.31	67.05	66.75	67.82	68.94	67,77
9	H	L	L	68.31	67.05	66.75	67.82	68.94	67,77
10	M	H	H	68.31	67.56	66.85	67.92	69.45	68,02
11	M	H	M	68.31	67.56	66.47	67.82	68.94	67,82
12	M	H	L	67.59	66.33	66.47	66.98	68.37	67,15
13	M	M	H	68.31	67.56	66.85	68.33	69.45	68,10
14	M	M	M	68.31	67.56	66.75	67.82	68.94	67,88
15	M	M	L	68.31	67.05	66.47	67.82	68.94	67,72
16	M	L	H	68.31	67.56	67.26	68.33	69.45	68,18
17	M	L	M	68.31	67.56	66.75	67.82	68.94	67,88
18	M	L	L	68.31	67.05	66.75	67.82	68.94	67,77
19	L	H	H	67.98	67.16	66.85	68.01	69.13	67,83
20	L	H	M	68.39	67.16	66.58	67.73	68.85	67,74
21	L	H	L	67.39	66.84	66.07	66.38	68.18	66,97
22	L	M	H	68.58	67.16	66.85	67.92	69.32	67,97
23	L	M	M	68.31	67.56	66.85	67.92	69.45	68,02
24	L	M	L	68.31	67.56	66.47	67.82	68.94	67,82
25	L	L	H	68.58	67.16	66.85	67.92	69.32	67,97
26	L	L	M	68.31	67.56	66.85	68.33	69.45	68,10
27	L	L	L	68.31	67.56	67.77	67.82	68.94	68,08
Media				67.94	66.96	66.43	67.48	68.70	67,50

Tabla 4.7. Valor de la precisión en el cálculo de la polaridad de las características con el método *N_GRAM Around*.

Casos de estudio				Precisión en el cálculo de la Polaridad de las Características					
Num.	Z ₁	Z ₂	Z ₃	NG=2	NG=3	NG=4	NG=5	NG=6	Media
1	H	H	H	75.10	74.59	72.68	72.14	73.93	73,69
2	H	H	M	75.10	73.73	72.68	72.14	73.93	73,52
3	H	H	L	75.10	74.18	73.12	72.14	73.93	73,69
4	H	M	H	75.40	75.04	72.68	72.14	73.93	73,84
5	H	M	M	75.10	74.48	72.68	72.14	73.93	73,67
6	H	M	L	75.54	74.18	72.68	72.14	73.93	73,69
7	H	L	H	75.40	75.85	73.30	72.76	73.93	74,25
8	H	L	M	75.85	74.48	72.68	72.14	73.93	73,82
9	H	L	L	75.54	74.48	72.68	72.14	73.93	73,75
10	M	H	H	75.10	74.59	72.68	72.14	73.93	73,69
11	M	H	M	75.10	74.29	73.12	72.14	73.93	73,72
12	M	H	L	75.10	73.73	72.56	72.14	73.93	73,49
13	M	M	H	75.40	74.59	72.68	72.14	73.93	73,75
14	M	M	M	75.10	74.59	72.68	72.14	73.93	73,69
15	M	M	L	75.10	73.73	72.68	72.14	73.93	73,52
16	M	L	H	75.40	75.40	73.30	72.76	73.93	74,16
17	M	L	M	75.40	75.04	72.68	72.14	73.93	73,84
18	M	L	L	75.10	74.48	72.68	72.14	73.93	73,67
19	L	H	H	75.10	74.59	72.68	71.69	73.93	73,60
20	L	H	M	75.10	74.29	72.56	72.14	73.93	73,60
21	L	H	L	75.10	74.29	72.56	72.14	73.93	73,60
22	L	M	H	75.40	74.59	72.68	72.14	73.93	73,75
23	L	M	M	75.10	74.59	72.68	72.14	73.93	73,69
24	L	M	L	75.10	74.29	73.12	72.14	73.93	73,72
25	L	L	H	75.40	75.40	73.30	72.76	73.93	74,16
26	L	L	M	75.40	74.59	72.68	72.14	73.93	73,75
27	L	L	L	75.10	74.59	72.68	72.14	73.93	73,69
Media				75.25	74.54	72.78	72.19	73.93	73,74

Tabla 4.8. Valor de la precisión en el cálculo de la polaridad de las características con el método *All Phrase*.

Casos de estudio				Precisión en el cálculo de la Polaridad de las Características
Num.	Z ₁	Z ₂	Z ₃	
1	H	H	H	69.72
2	H	H	M	69.72
3	H	H	L	70.17
4	H	M	H	69.72
5	H	M	M	70.17
6	H	M	L	70.17
7	H	L	H	70.17
8	H	L	M	70.17
9	H	L	L	70.17
10	M	H	H	69.42
11	M	H	M	69.42
12	M	H	L	69.42
13	M	M	H	69.72
14	M	M	M	69.72
15	M	M	L	69.72
16	M	L	H	69.72
17	M	L	M	69.72
18	M	L	L	69.72
19	L	H	H	68.10
20	L	H	M	68.54
21	L	H	L	68.54
22	L	M	H	68.52
23	L	M	M	69.42
24	L	M	L	69.42
25	L	L	H	69.27
26	L	L	M	69.72
27	L	L	L	69.72
Media				69.56

Los resultados del método “*N_GRAM Before*” para el cálculo de la polaridad de características se muestran en la tabla 4.5. La mejor media de precisión, con un valor de 78.43%, se obtiene con *N_GRAM* igual a 6. Esto significa que el sistema obtiene los mejores resultados cuando tiene en cuenta el valor sentimental en la base de datos *SWN* de las 6 palabras que preceden a la expresión lingüística que representa a la característica en la opinión del usuario. Además, para este método los peores resultados se dan con *N_GRAM* igual a 3, obteniendo una precisión del 71.11%. Por otro lado, la tasa de éxito promedio no varía con las 27 configuraciones diferentes de los parámetros *Z*₁, *Z*₂ y *Z*₃, obteniendo en la mayoría de las ocasiones los mismos resultados. Por último, la media en la tasa de éxito de todas las combinaciones de parámetros para este método en el cálculo de polaridades de características es del 74.18%.

La tabla 4.6 muestra los resultados obtenidos en el cálculo de la polaridad de las características usando el método “*N_GRAM After*”. Se puede observar que los resultados obtenidos con este método son peores que los obtenidos con el método

“*N_GRAM Before*”. De nuevo, la mejor media de tasa de éxito en el proceso de identificación de polaridades se obtiene cuando el parámetro *N_GRAM* es igual a 6. Sin embargo, en este caso la precisión máxima es del 68.9%, lo que supone un descenso bastante acusado frente a los resultados obtenidos con el método anterior. Además, analizando todos los resultados obtenidos es posible comprobar que no hay diferencias significativas en la aplicación de este método independientemente de los valores de los distintos parámetros (*N_GRAM*, Z_1 , Z_2 y Z_3). La media total en este caso es del 67.80%.

Los resultados obtenidos para el cálculo de la polaridad de características con el método “*N_GRAM Around*” se muestran en la tabla 4.7. En esta tabla es posible observar que los resultados obtenidos con este método son significativamente mejores que los obtenidos con el método “*N_GRAM After*”, pero ligeramente peores que aquellos obtenidos con el método “*N_GRAM Before*”. Más concretamente, la mejor tasa de acierto media para el proceso de identificación de las polaridades de las características identificadas en las opiniones de los usuarios se obtiene con *N_GRAM* igual a 2, con una precisión máxima del 75.54%. Esto refleja que se consiguen los mejores resultados cuando se tienen en cuenta, para el cálculo de la polaridad de las características, las 2 palabras que preceden a la expresión lingüística de cada característica en la opinión del usuario junto con las 2 palabras que van inmediatamente después. La media total, agregando los resultados utilizando las distintas configuraciones de los parámetros *N_GRAM*, Z_1 , Z_2 y Z_3 , es del 73.74%, sin apreciar diferencias significativas por variaciones en los parámetros Z_1 , Z_2 y Z_3 .

La tabla 4.8 muestra los resultados obtenidos con el método “*All Phrase*”. Como se ha destacado previamente, este método difiere de los anteriores en que no es necesaria la configuración del parámetro del sistema *N_GRAM* para llevar a cabo el proceso de identificación de la polaridad de las características. En este método se tienen en cuenta todas las palabras que se encuentran en la misma frase que la expresión lingüística que representa a la característica que se desea clasificar sentimentalmente. Los resultados obtenidos por este método son peores que los obtenidos con los métodos de clasificación “*N_GRAM Before*” y “*N_GRAM Around*”, pero son mejores que los obtenidos con el método “*N_GRAM After*”. En particular, la tasa de éxito media para el cálculo de la polaridad de características es del 69.56%. Como ocurre con los métodos anteriores, la tasa media de acierto no difiere demasiado con las 27 diferentes combinaciones para Z_1 , Z_2 y Z_3 .

Por último, la figura 4.3 muestra gráficamente una comparativa visual entre los distintos métodos de cálculo de la polaridad de las características identificadas. La gráfica

emplea un color distinto para los valores medios de precisión obtenidos por cada uno de los métodos de cálculo de polaridad de características estudiados en el experimento, organizados en función del valor del parámetro N_GRAM . En el gráfico se puede ver claramente que los mejores resultados medios se obtienen con el método de clasificación “ N_GRAM Before” cuando el valor del parámetro N_GRAM es igual a 6. También es posible destacar el hecho de que el método “ N_GRAM Around” es el método más regular, ya que es el método que obtiene mejores resultados con las configuraciones de N_GRAM igual a 2, 3 y 4. Por último, se muestra claramente que el método “ N_GRAM After” es el que peores resultados obtiene para cualquier valor de N_GRAM .

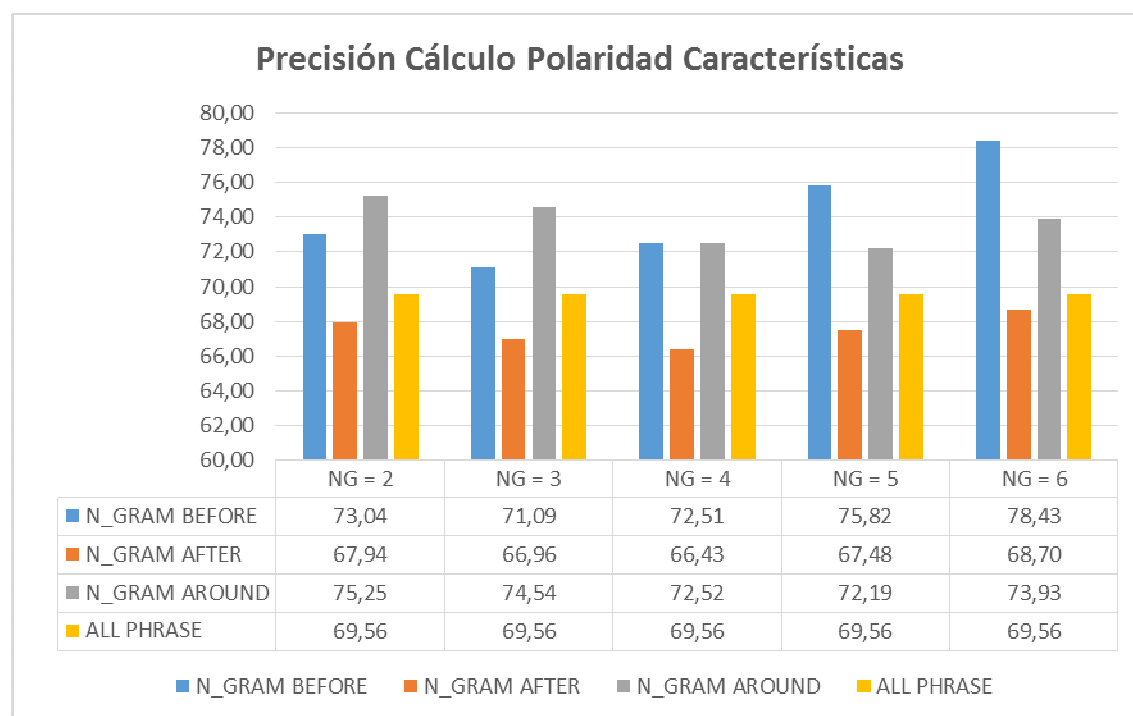


Figura 4.3. Comparativa visual de precisión en el cálculo de la polaridad de las características en el dominio cinematográfico.

4.4.6 Resultados para la clasificación de sentimientos de opiniones

Para evaluar el proceso de clasificación de sentimientos de las opiniones vertidas por los usuarios es necesario etiquetar previamente cada uno de los documentos del corpus empleado como positivo o negativo. Es decir, es preciso etiquetar manualmente la polaridad global de cada opinión de usuario. Sin embargo, el corpus utilizado (Pang & Lee, 2004) viene organizado en dos carpetas: una para los documentos con opiniones

positivas y otra para las negativas. Por tanto, el pre-requisito para poder acometer la tarea de validación ya se encuentra satisfecho.

El proceso de evaluación consistirá en calcular la polaridad global de cada opinión de usuario a través del sistema automático de análisis de sentimientos de opiniones para posteriormente comparar estos resultados con la clasificación de sentimientos llevada a cabo por expertos.

Se han llevado a cabo se han evaluado cuatro métodos diferentes para llevar a cabo el análisis de sentimientos de las opiniones de los usuarios. Además, también se estudia la incidencia de los parámetros Z_1 , Z_2 y Z_3 y el parámetro del sistema N_GRAM . Es necesario conducir el experimento para que sea posible testar todas las posibilidades para los parámetros Z_1 , Z_2 y Z_3 y N_GRAM con los cuatro métodos de clasificación de sentimientos: “ N_GRAM Before”, “ N_GRAM After”, “ N_GRAM Around” y “ All Phrase”. Esto significa probar un total de 27 casos de estudio y 135 valores diferentes para todas las posibles combinaciones.

Los resultados para el análisis de sentimientos de las opiniones de usuarios con estos cuatro métodos se presentan en las tablas 4.9, 4.10, 4.11 y 4.12. En ellas las abreviaciones “ H ”, “ M ” y “ L ” representan los valores “ $High$ ”, “ $Medium$ ” y “ Low ” para los parámetros Z_1 , Z_2 y Z_3 , mientras que la abreviación “ $NG = x$ ” indica que el valor ‘ x ’ ha sido asignado al parámetro del sistema N_GRAM .

Los resultados para el método “ N_GRAM Before” en el análisis de sentimientos de opiniones se muestran en la tabla 4.9. Como se puede observar, empleando este método la mejor tasa de éxito puntual que se consigue es del 89.6%. Este valor se obtiene para distintos valores del parámetro N_GRAM ($NG=2$, $NG=3$, $NG=4$, $NG=5$ o $NG=6$), siendo el valor de N_GRAM igual a 3 el que alcanza un mejor valor promedio para los 27 casos estudiados, con un 74.67% de precisión.

En la figura 4.4 aparece reflejada la influencia de los parámetros del sistema Z_1 , Z_2 y Z_3 en el análisis de sentimientos de opiniones usando el método “ N_GRAM Before”. En el eje de abscisas de la gráfica se encuentran las 27 combinaciones de valores H , M y L para los parámetros Z_1 , Z_2 y Z_3 . En el eje de ordenadas se representa la precisión en el cálculo sentimental obtenida con cada uno de los valores para el parámetro N_GRAM (del 2 al 6). Como se puede apreciar, los valores más altos de precisión se logran con la configuración $Z_1 = H$, $Z_2 = L$ y $Z_3 = H$ para cualquier valor del parámetro N_GRAM . Con esta configuración de parámetros se alcanza casi un 90% de precisión en el análisis de sentimientos, en concreto un 89.60%. Este resultado se puede catalogar como muy

positivo teniendo en cuenta la variabilidad y heterogeneidad de las opiniones de los usuarios en el dominio en el que nos encontramos. Esto refleja que, tal como se había adelantado en capítulos anteriores, las características situadas al principio y al final de la opinión del usuario tienen más peso a la hora de evaluar la opinión global del usuario. Se obtiene la misma precisión en el análisis de sentimientos cuando $Z_1 = H$, $Z_2 = L$ y $Z_3 = M$, y $Z_1 = H$, $Z_2 = L$ y $Z_3 = L$, y el parámetro N_GRAM es igual a 4 o 5. Se puede concluir que, de acuerdo con los resultados experimentales, en la metodología propuesta las características situadas al inicio de la opinión del usuario son las que deben de tener más peso para obtener los mejores resultados en el análisis de sentimientos con el método “ N_GRAM Before”.

Tabla 4.9. Valor de la precisión en el análisis de sentimientos de opiniones con el método *N_GRAM Before*.

Casos de estudio				Análisis de sentimientos					
Num.	Z ₁	Z ₂	Z ₃	NG=2	NG=3	NG=4	NG=5	NG=6	Media
1	H	H	H	67.20	78.40	67.20	67.20	56.00	67,20
2	H	H	M	56.00	67.20	67.20	78.40	56.00	64,96
3	H	H	L	56.00	67.20	67.20	67.20	44.80	60,48
4	H	M	H	78.40	78.40	78.40	78.40	67.20	76,16
5	H	M	M	78.40	67.20	78.40	78.40	78.40	76,16
6	H	M	L	78.40	67.20	67.20	78.40	78.40	73,92
7	H	L	H	89.60	89.60	89.60	89.60	89.60	89,60
8	H	L	M	78.40	89.60	89.60	89.60	78.40	85,12
9	H	L	L	78.40	78.40	89.60	89.60	78.40	82,88
10	M	H	H	67.20	67.20	67.20	56.00	56.00	62,72
11	M	H	M	56.00	67.20	67.20	56.00	56.00	60,48
12	M	H	L	44.80	78.40	56.00	44.80	44.80	53,76
13	M	M	H	56.00	78.40	67.20	56.00	56.00	62,72
14	M	M	M	67.20	78.40	67.20	67.20	78.40	71,68
15	M	M	L	56.00	67.20	67.20	78.40	56.00	64,96
16	M	L	H	56.00	78.40	78.40	78.40	56.00	69,44
17	M	L	M	78.40	78.40	78.40	78.40	67.20	76,16
18	M	L	L	78.40	67.20	78.40	78.40	78.40	76,16
19	L	H	H	56.00	78.40	67.20	67.20	56.00	64,96
20	L	H	M	67.20	78.40	56.00	67.20	67.20	67,20
21	L	H	L	67.20	78.40	44.80	56.00	67.20	62,72
22	L	M	H	56.00	67.20	67.20	56.00	56.00	60,48
23	L	M	M	67.20	67.20	67.20	56.00	56.00	62,72
24	L	M	L	56.00	67.20	67.20	56.00	56.00	60,48
25	L	L	H	56.00	78.40	67.20	44.80	44.80	58,24
26	L	L	M	56.00	78.40	67.20	56.00	56.00	62,72
27	L	L	L	67.20	78.40	67.20	67.20	56.00	67,20
Media				65.54	74.67	70.10	68.03	62.64	68,20

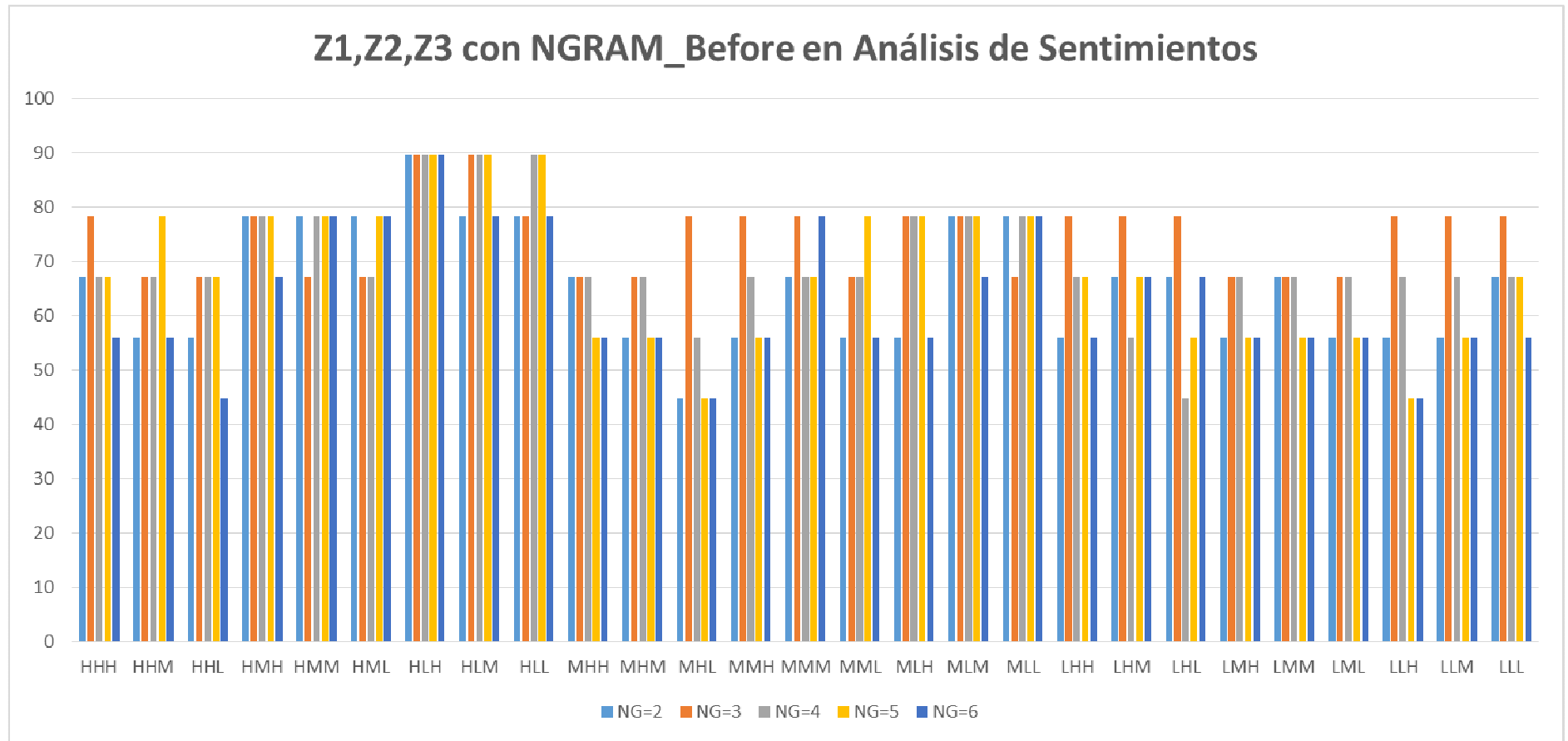


Figura 4.4. Efecto de los parámetros Z_1 , Z_2 y Z_3 en el análisis de sentimientos de opiniones con el método *N_GRAM Before*.

La tabla 4.10 muestra los resultados obtenidos en cuanto a tasa de acierto para el método “*N_GRAM After*” en el análisis de sentimientos de opiniones de usuarios sobre películas. Los resultados obtenidos con este método de clasificación de sentimientos son insuficientes haciendo patente la ineficacia de este método para la clasificación de sentimientos de opiniones en inglés. En concreto, los mejores resultados con este método, en promedio, se obtienen con *N_GRAM* igual a 2, y la tasa de acierto queda por debajo del 50% (46.04%). Estos resultados eran de esperar y ponen de manifiesto que en el idioma inglés las palabras más significativas a la hora de estudiar la polaridad de una característica son las inmediatamente anteriores, como reflejan los buenos resultados obtenidos con el método “*N_GRAM Before*”. Por ejemplo, si un usuario expresa una frase del tipo: “*Today I saw a great performance of Brad Pitt at the Metropolis cinema next to my house*”, es posible observar que para el cálculo de la polaridad de las características ‘*performance*’ o ‘*Brad Pitt*’ debería tenerse en cuenta el adjetivo calificativo ‘*great*’ que, con el método “*N_GRAM After*” sería obviado.

En la figura 4.5, se aprecia que este método obtiene picos de rendimiento para algunas configuraciones de Z_1 , Z_2 y Z_3 , llegando al 56% de precisión cuando el valor de *N_GRAM* es igual a dos. Debido a la aleatoriedad de los resultados obtenidos, no es posible destacar ninguna configuración de Z_1 , Z_2 y Z_3 como la óptima. En cualquier caso, queda empíricamente demostrado que este método no es adecuado para la clasificación de sentimientos de opiniones de usuarios en idioma inglés.

Tabla 4.10. Valor de la precisión en el análisis de sentimientos de opiniones con el método *N_GRAM After*.

Casos de estudio				Análisis de sentimientos					
Num.	Z ₁	Z ₂	Z ₃	NG=2	NG=3	NG=4	NG=5	NG=6	Media
1	H	H	H	33.60	22.40	33.60	33.60	33.60	31,36
2	H	H	M	44.80	33.60	33.60	33.60	33.60	35,84
3	H	H	L	56.00	44.80	44.80	44.80	44.80	47,04
4	H	M	H	44.80	22.40	22.40	33.60	33.60	31,36
5	H	M	M	44.80	33.60	22.40	22.40	33.60	31,36
6	H	M	L	56.00	33.60	33.60	33.60	44.80	40,32
7	H	L	H	44.80	22.40	22.40	22.40	33.60	29,12
8	H	L	M	56.00	22.40	22.40	33.60	44.80	35,84
9	H	L	L	56.00	44.80	44.80	44.80	56.00	49,28
10	M	H	H	33.60	22.40	33.60	33.60	33.60	31,36
11	M	H	M	56.00	22.40	33.60	33.60	33.60	35,84
12	M	H	L	56.00	44.80	33.60	33.60	33.60	40,32
13	M	M	H	44.80	22.40	33.60	33.60	33.60	33,60
14	M	M	M	33.60	22.40	33.60	33.60	33.60	31,36
15	M	M	L	44.80	33.60	33.60	33.60	33.60	35,84
16	M	L	H	44.80	22.40	22.40	33.60	33.60	31,36
17	M	L	M	44.80	22.40	22.40	33.60	33.60	31,36
18	M	L	L	44.80	33.60	22.40	22.40	33.60	31,36
19	L	H	H	44.80	22.40	33.60	33.60	33.60	33,60
20	L	H	M	44.80	22.40	33.60	33.60	33.60	33,60
21	L	H	L	56.00	44.80	33.60	33.60	33.60	40,32
22	L	M	H	44.80	22.40	33.60	33.60	33.60	33,60
23	L	M	M	33.60	22.40	33.60	33.60	33.60	31,36
24	L	M	L	56.00	22.40	33.60	33.60	33.60	35,84
25	L	L	H	44.80	22.40	33.60	33.60	33.60	33,60
26	L	L	M	44.80	22.40	33.60	33.60	33.60	33,60
27	L	L	L	33.60	22.40	33.60	33.60	33.60	31,36
Media				46.04	27.79	31.53	33.19	35.67	34,84

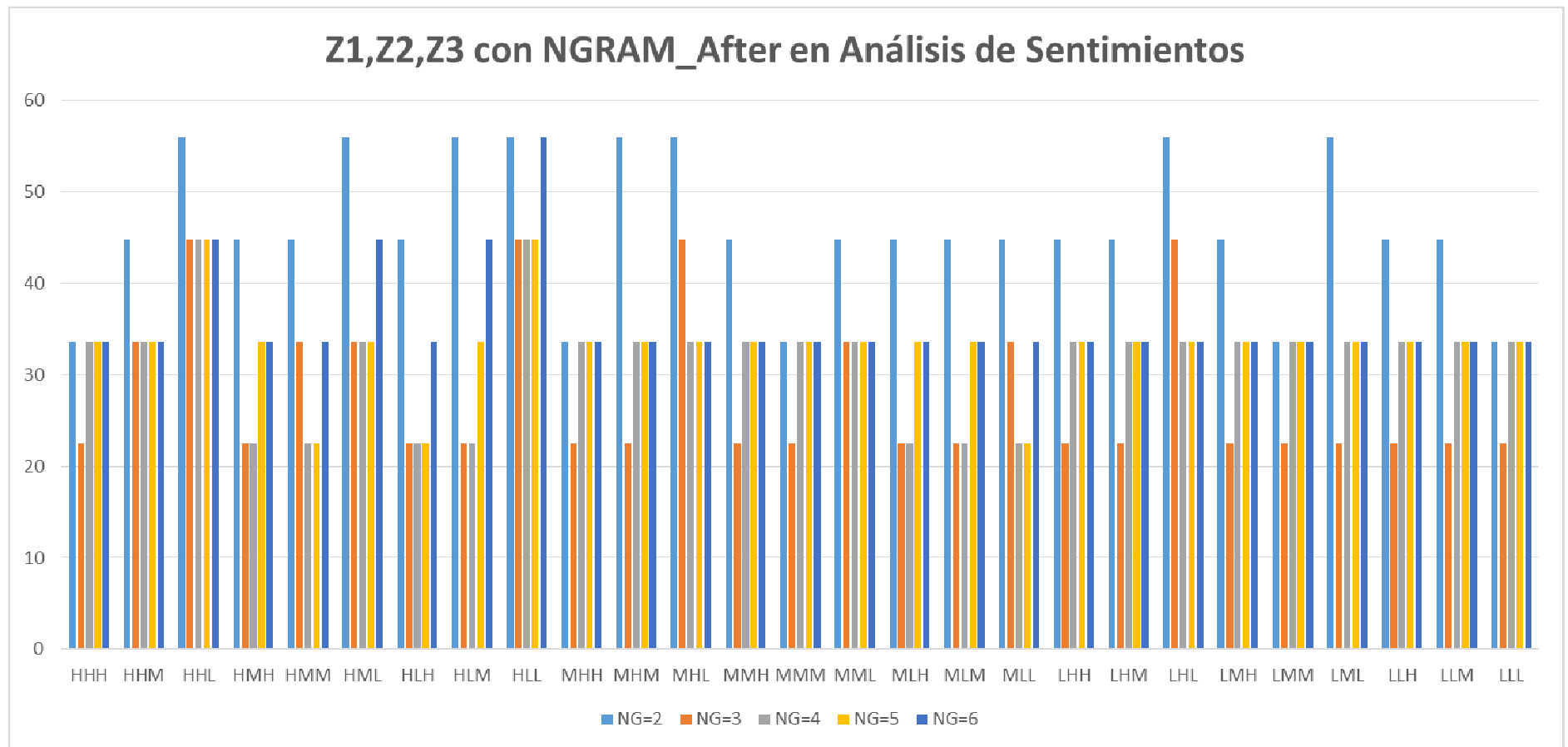


Figura 4.5. Efecto de los parámetros Z_1 , Z_2 y Z_3 en el análisis de sentimientos de opiniones con el método N_GRAM After.

En la tabla 4.11, se muestran los resultados para la clasificación de sentimientos de opiniones obtenidos con el método “*N_GRAM Around*”. Los resultados alcanzados con este método para el análisis de sentimientos son alentadores, aunque sensiblemente inferiores a los obtenidos con el método “*N_GRAM Before*”. Los valores máximos de precisión, que llegan al 78.40% de acierto, se obtienen con varias configuraciones de parámetros del sistema. Los experimentos revelan que el valor de *N_GRAM* igual a 3 es el que mejores resultados obtiene en promedio, alcanzando un 72.18%. En general, se trata de un método muy regular en cuanto a precisión, ya que de todas las configuraciones probadas la mayoría produce resultados positivos. Los peores resultados con este método se dan con el valor de *N_GRAM* igual a 6. Esto pone de manifiesto que para este método es recomendable tener en cuenta únicamente las palabras más cercanas a la expresión lingüística de la característica.

La figura 4.6 refleja la influencia de los parámetros del sistema Z_1 , Z_2 y Z_3 en la precisión del análisis de sentimientos de opiniones. Es posible apreciar que se obtienen los mejores resultados cuando el valor de $Z_1 = H$. La configuración óptima para estos parámetros se da con $Z_1 = H$, $Z_2 = L$ y $Z_3 = L$, con una precisión del 78.40% para la clasificación de sentimientos de opiniones, independientemente del valor asignado al parámetro *N_GRAM*. Por tanto, se puede concluir que el método “*N_GRAM Around*” es efectivo para el análisis de sentimientos de las opiniones de usuarios en lenguaje inglés en el dominio de las películas cinematográficas.

Tabla 4.11. Valor de la precisión en el análisis de sentimientos de opiniones con el método *N_GRAM Around*.

Casos de estudio				Análisis de sentimientos					
Num.	Z ₁	Z ₂	Z ₃	NG=2	NG=3	NG=4	NG=5	NG=6	Media
1	H	H	H	56.00	78.40	78.40	56.00	44.80	62,72
2	H	H	M	78.40	78.40	78.40	56.00	44.80	67,20
3	H	H	L	67.20	78.40	78.40	56.00	67.20	69,44
4	H	M	H	67.20	78.40	78.40	56.00	56.00	67,20
5	H	M	M	78.40	78.40	78.40	67.20	67.20	73,92
6	H	M	L	67.20	78.40	78.40	67.20	67.20	71,68
7	H	L	H	67.20	78.40	78.40	56.00	67.20	69,44
8	H	L	M	78.40	78.40	78.40	67.20	67.20	73,92
9	H	L	L	78.40	78.40	78.40	78.40	78.40	78,40
10	M	H	H	56.00	78.40	56.00	44.80	33.60	53,76
11	M	H	M	56.00	67.20	56.00	44.80	44.80	53,76
12	M	H	L	67.20	67.20	67.20	44.80	44.80	58,24
13	M	M	H	67.20	56.00	56.00	44.80	33.60	51,52
14	M	M	M	56.00	78.40	78.40	56.00	44.80	62,72
15	M	M	L	78.40	78.40	78.40	56.00	44.80	67,20
16	M	L	H	67.20	67.20	56.00	44.80	33.60	53,76
17	M	L	M	67.20	78.40	78.40	56.00	56.00	67,20
18	M	L	L	78.40	78.40	78.40	67.20	67.20	73,92
19	L	H	H	56.00	56.00	33.60	44.80	33.60	44,80
20	L	H	M	56.00	67.20	44.80	44.80	44.80	51,52
21	L	H	L	67.20	67.20	44.80	44.80	44.80	53,76
22	L	M	H	67.20	56.00	44.80	33.60	33.60	47,04
23	L	M	M	56.00	78.40	56.00	44.80	33.60	53,76
24	L	M	L	56.00	67.20	56.00	44.80	44.80	53,76
25	L	L	H	67.20	67.20	56.00	33.60	33.60	51,52
26	L	L	M	67.20	56.00	56.00	44.80	33.60	51,52
27	L	L	L	56.00	78.40	78.40	56.00	44.80	62,72
Media				65.96	72.18	65.96	52.27	48.53	60,98

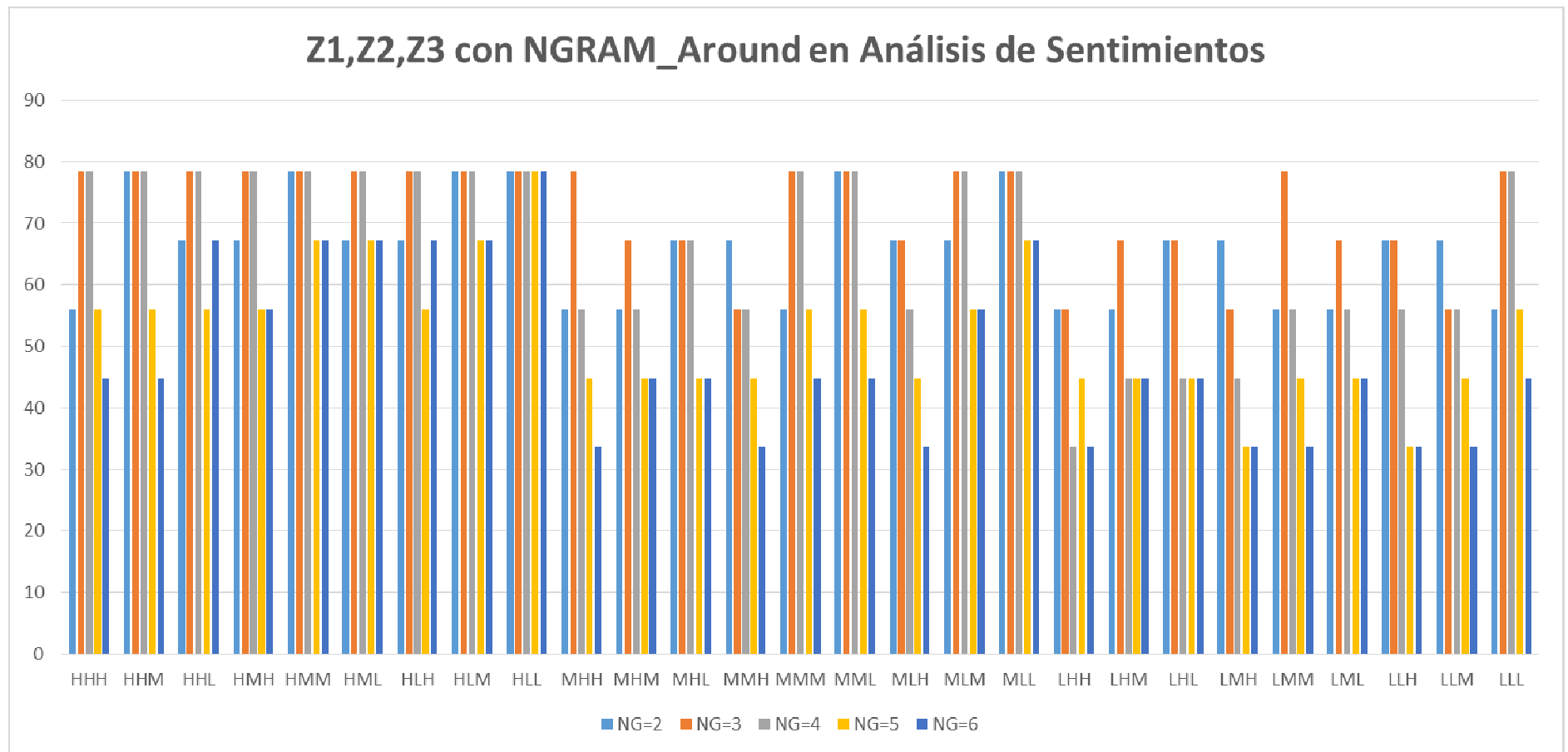


Figura 4.6. Efecto de los parámetros Z_1, Z_2 y Z_3 en el análisis de sentimientos de opiniones con el método N_GRAM_Around .

Por último, la tabla 4.12 muestra los resultados obtenidos para la clasificación de sentimientos de opiniones usando el método “*All Phrase*”. Como ya se ha mencionado anteriormente, este método no precisa del parámetro del sistema *N_GRAM*, ya que se analiza por completo cada frase que contenga una o varias características. Como se puede apreciar a partir de los datos presentes en la tabla, los resultados son bastante discretos, obteniendo un 53.10% de valor promedio en el cálculo de la precisión de la clasificación de sentimientos. Mejora los resultados del método “*N_GRAM After*”, pero queda muy por debajo de los obtenidos con “*N_GRAM Before*” y “*N_GRAM Around*”. El valor máximo de precisión obtenido con este método es del 67.20% para algunas configuraciones de parámetros del sistema, que representa un máximo muy inferior al obtenido con los métodos antes citados. Esto deja patente que, en general, no es conveniente analizar toda la frase que contiene la expresión lingüística de la característica identificada a la hora de calcular la polaridad de la opinión global. Se concluye, teniendo en cuenta los resultados de los experimentos, que las palabras más cercanas a la expresión lingüística de la característica son las que más deben considerarse a la hora de clasificarla sentimentalmente. Existen, sin embargo, matices a considerar que se analizan a continuación.

Tabla 4.12. Valor de la precisión en el análisis de sentimientos de opiniones con el método *All Phrase*.

Casos de estudio				Análisis de sentimientos
Num.	Z ₁	Z ₂	Z ₃	
1	H	H	H	56.00
2	H	H	M	67.20
3	H	H	L	67.20
4	H	M	H	67.20
5	H	M	M	67.20
6	H	M	L	67.20
7	H	L	H	56.00
8	H	L	M	67.20
9	H	L	L	67.20
10	M	H	H	33.60
11	M	H	M	44.80
12	M	H	L	44.80
13	M	M	H	44.80
14	M	M	M	56.00
15	M	M	L	67.20
16	M	L	H	44.80
17	M	L	M	67.20
18	M	L	L	67.20
19	L	H	H	33.60
20	L	H	M	44.80
21	L	H	L	44.80
22	L	M	H	33.60
23	L	M	M	33.60
24	L	M	L	44.80
25	L	L	H	44.80
26	L	L	M	44.80
27	L	L	L	56.00
Media				53.10

La figura 4.7 muestra la incidencia de los parámetros Z_1 , Z_2 y Z_3 en la precisión del análisis de sentimientos de opiniones con el método “*All Phrase*”. En la figura se observa que los mejores resultados se obtienen cuando $Z_1 = H$. Los resultados de este método para algunas configuraciones concretas son prometedores, consiguiendo 10 configuraciones diferentes para Z_1 , Z_2 y Z_3 con valores de precisión cercanos al 70%. De ahí que se pueda afirmar que, aunque el valor promedio agregando todas las configuraciones es pobre, existen configuraciones concretas que obtienen resultados positivos.

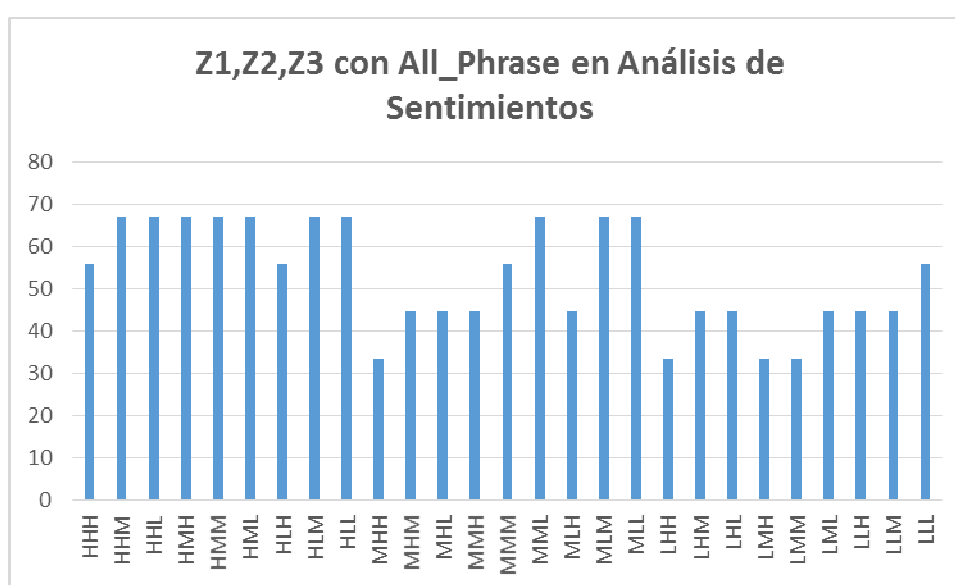


Figura 4.7. Efecto de los parámetros Z_1 , Z_2 y Z_3 en el análisis de sentimientos de opiniones con el método *All Phrase*.

La figura 4.8 presenta gráficamente un resumen de todos los resultados de los experimentos que se han llevado a cabo. En concreto, se muestra la precisión media obtenida con cada uno de los métodos estudiados (“*N_GRAM Before*”, “*N_GRAM After*”, “*N_GRAM Around*” y “*All Phrase*”) en el análisis de sentimientos de opiniones de usuarios en el dominio de las películas cinematográficas e idioma inglés. Los métodos “*N_GRAM Before*” y “*N_GRAM Around*” destacan como los mejores métodos de clasificación de sentimientos de opiniones de entre los estudiados. Además, estos métodos obtienen sus mejores resultados promedios de precisión con el valor de N_GRAM igual a 3. Por último, el método “*N_GRAM After*” puede ser descartado para este dominio e idioma, mientras que el método “*All Phrase*” no puede considerarse

como válido en resultados promedios, aunque sí como alternativa en algunas configuraciones de parámetros del sistema muy concretas, tal y como se ha destacado anteriormente.

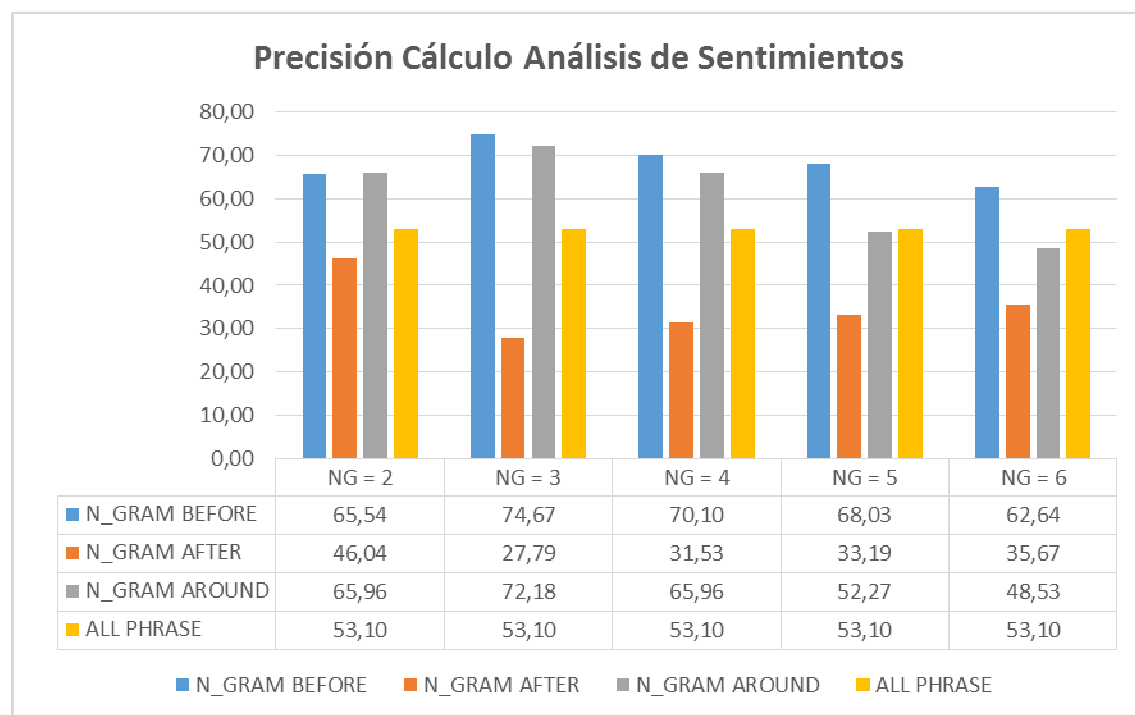


Figura 4.8. Comparativa visual de precisión en el cálculo del análisis de sentimientos de opiniones.

Además, la figura 4.8 muestra que el método *N_GRAM_Before* obtiene resultados muy superiores para sus valores de precisión en el proceso de análisis de sentimientos con respecto al método *N_GRAM_After*. La razón de esta diferencia tan significativa radica en el hecho de que este experimento haya sido realizado utilizando un corpus en inglés. En este idioma las palabras de sentimiento asociadas a una característica identificada por el sistema se sitúan, en la mayor parte de los casos, en las palabras precedentes a la expresión lingüística que representa a la característica. Esto implica que los mejores resultados para el cálculo de la polaridad de las características identificadas por el sistema se obtengan utilizando el método *N_GRAM_Before*. Por tanto, como la metodología propuesta en esta tesis doctoral realiza el análisis de sentimientos de las opiniones de los usuarios teniendo en cuenta la polaridad asociada a cada característica en dichas opiniones, el resultado del análisis de sentimientos es una consecuencia directa de ello.

4.4.7 Conclusión

Tal y como se puede apreciar de los resultados obtenidos en los experimentos llevados a cabo para el dominio cinematográfico, la clave para el correcto funcionamiento de la metodología propuesta se encuentra en la primera parte del proceso. En concreto, la identificación correcta de las características dentro de las opiniones de los usuarios y el cálculo correcto de su polaridad resultan fundamentales ya que éstas determinarán la polaridad global de la opinión del usuario guiando el proceso de análisis de sentimientos.

Atendiendo a los malos resultados tanto del método "*N_GRAM After*" como del método "*All Phrase*", es razonable concluir que éstos no son apropiados para el cálculo del análisis de sentimientos de opiniones de usuarios en el idioma inglés. Más concretamente, empleando el método "*N_GRAM After*" los resultados de la clasificación de sentimientos de documentos como positivos, negativos y neutros son muy pobres, con un promedio del 34.84% de precisión.

De los cuatro métodos analizados, el método "*N_GRAM Before*" es el que logra los mejores resultados tanto para el proceso de cálculo de polaridades de características como en el análisis de sentimientos de las opiniones de los usuarios en inglés, obteniendo unos valores máximos de precisión de 78.43% y 89.60%, respectivamente. Estos resultados se obtienen con $Z_1 = H$, $Z_2 = L$ y $Z_3 = H$, que son los mejores valores para estos parámetros de acuerdo con la tasa promedio de acierto para el análisis de sentimientos de las opiniones de los usuarios. Es decir, los mejores resultados en el análisis de sentimientos de opiniones de usuarios se obtienen cuando se les asignan los pesos más altos a las características colocadas al principio y al final de la opinión de un usuario. Este hecho demuestra empíricamente la afirmación que se realizó en el Capítulo II con respecto a la forma de calcular el peso asignado a cada característica dentro de la opinión de un individuo.

Por último, cabe destacar que todos los métodos estudiados (es decir, "*N_GRAM Before*", "*N_GRAM After*", "*N_GRAM Around*" y "*All Phrase*") logran, en general, mejores resultados promedio en el proceso de identificación de polaridades de características que en el análisis de sentimientos de todo el documento (polaridad global a nivel de documento). Este hecho refleja que existen ocasiones en las que, por ejemplo, un usuario valora positivamente un conjunto de características dentro de una opinión pero su opinión global ha sido clasificada por un experto como negativa. Por tanto, en este caso el sistema de análisis de sentimientos propuesto clasificaría la opinión como positiva, ya que la polaridad global de la opinión del documento se obtiene como la

suma vectorial ponderada de cada una de las características identificadas en ese documento.

Con la información recabada por las medidas de evaluación que se han empleado, es posible extraer las siguientes conclusiones para el dominio de las películas:

- Se trata de un dominio que usa un lenguaje informal donde, por lo general, las características a identificar están bastante acotadas y las opiniones de los usuarios son muy variadas, obteniendo casos de opiniones muy específicas y otras muy genéricas. Esto incide directamente en el cálculo del análisis de sentimientos de las opiniones, ya que, en ocasiones, una opinión de usuario puede contener muchas características clasificadas correctamente con una polaridad determinada y, sin embargo, la polaridad global de la opinión del usuario ser contraria a esa polaridad. Este hecho dificulta el proceso de análisis de sentimientos de las opiniones de los usuarios.
- Con unas precisiones máximas del 78,43% en el cálculo de polaridades de características y un 89,60% en el proceso de análisis de sentimientos de opiniones, se considera que los objetivos estipulados al llevar a cabo este experimento han sido plenamente satisfechos. En cualquier caso, también queda patente que la metodología propuesta es mejorable.

4.5 Validación en el dominio de las finanzas

4.5.1 Introducción

El dominio de las finanzas engloba aquellas actividades relacionadas con el intercambio de distintos bienes de capital entre individuos, empresas, o estados. Estas actividades, generalmente, conllevan incertidumbre y el riesgo (Pareja, 2003). Las finanzas son una de las ramas de la economía que se dedica al estudio de la obtención de capital para la inversión en bienes productivos y al estudio de las decisiones de inversión de los ahorradores (Restrepo & Medina, 2014).

En ese marco se estudia la obtención y gestión por parte de una compañía, un individuo, o del propio estado, de los fondos que necesita para cumplir sus objetivos (Doove et al., 2014) y de los criterios con los que dispone de sus activos. Es decir, se analiza lo

relativo a la obtención y gestión del dinero, así como de otros valores o activos (títulos, bonos, etc.).

Según Bodie y Merton, las finanzas *"estudian la manera en que los recursos escasos se asignan a través del tiempo"*. Esta área trata sobre la oportunidad, las condiciones y formas de conseguir el capital, de su uso y de los beneficios que obtienen los inversores a través de sus operaciones.

Según la revista *Información Financiera*⁴⁶, *"la información financiera seguirá siendo el mejor apoyo para la toma de decisiones de los negocios"*. Cuanto mejor sea la calidad de la información, mayor probabilidad de éxito se tendrá en la toma de las decisiones. Pero para lograr dicha calidad, esta deberá contar con datos e información que satisfagan las necesidades de los diferentes usuarios (Kim et al., 2014).

Dentro del área de las finanzas existen tres áreas generales:

- **Finanzas Corporativas.** Se centra en la forma en la que los directivos pueden crear valor con las decisiones financieras, básicamente, las decisiones de inversión y financiamiento. Utilizan conocimientos de otras disciplinas con las que básicamente están relacionadas: la economía y la contabilidad.
- **Finanzas Personales.** Se centra en los gastos cotidianos que posee una familia normal: créditos, hipotecas, domiciliaciones, alimentación, ropa, calzado, etc. También se contempla en este tema: el ahorro, las inversiones, los préstamos intrafamiliares, etc. El objetivo es alcanzar un control de los ingresos y de los gastos más eficaz a lo largo de un periodo de tiempo determinado.
- **Finanzas Públicas.** Es un campo de la economía que gestiona todos aquellos gastos que realiza el estado, con el fin de lograr la satisfacción de las necesidades de la sociedad. Las finanzas públicas están estrechamente vinculadas a las cuestiones de distribución del ingreso y la igualdad social.

Debido a la existencia de una multitud de información que debe ser gestionada de la mejor manera posible para facilitar la toma de decisiones a un gran número de personas y empresas, el área de finanzas corporativas puede ser el sector más beneficiado del sistema de análisis de sentimientos propuesto en esta tesis doctoral.

Actualmente, existen varios trabajos sobre minería de opiniones que han centrado sus investigaciones en el ámbito de las finanzas (Hai et al., 2014; Kim et al., 2014; Ruiz-

⁴⁶ Revista Online. Consultada en septiembre de 2014: <http://www.informacionfinanciera.es/financiera-informacion-financiera---perspectivas.html>.

Martínez et al., 2012; Smailovic et al., 2014). Por ejemplo, el trabajo presentado en (Ruiz-Martínez et al., 2012) lleva a cabo el análisis de sentimientos de noticias financieras basado en el uso de ontologías. En este trabajo, se propone una metodología semántica para la anotación de polaridades en noticias financieras, y muestra un algoritmo para la extracción de opiniones en el ámbito financiero. Otro trabajo sobre minería de opiniones en el dominio de las finanzas es el trabajo de (Smailovic et al., 2014) donde se estudian las relaciones existentes entre las opiniones vertidas por los usuarios en *Twitter* sobre compañías y sus productos, y los cambios de los precios en bolsa de estos productos. En (Kim et al. 2014) se propone un modelo inteligente de inversión basado en minería de opiniones en el dominio de las finanzas para predecir subidas y bajadas de índices bursátiles.

Se ha elegido este dominio para validar la metodología propuesta en esta tesis doctoral por su complejidad y su gran importancia en la sociedad actual. La crisis económica mundial a la que nos enfrentamos nos hace estar consumiendo este tipo de noticias y opiniones diariamente a través de la prensa económica, la radio o la televisión.

Existe una cantidad enorme de sitios Web dedicados exclusivamente a recoger las noticias financieras actuales y, como se ha comentado anteriormente, el análisis de toda esta información es un trabajo muy costoso. El sistema propuesto en esta tesis doctoral puede ayudar a analizar el conocimiento intrínseco en las opiniones de los usuarios en el dominio financiero. Este conocimiento podría ayudar en la toma de decisiones que los inversores y demás expertos en el dominio financiero toman en un momento determinado.

4.5.2 Corpus

Para llevar a cabo la validación del sistema en el dominio de las finanzas se ha construido un corpus de opiniones compuesto de 200 documentos. Estos documentos contienen opiniones de usuarios sobre noticias financieras en inglés. A veces un documento contiene la opinión de un usuario con respecto a una noticia financiera determinada pero, en otras ocasiones, se vierten opiniones sobre varios temas de actualidad en el dominio de las finanzas dentro de la misma opinión.

El corpus del experimento contiene 30.472 palabras distribuidas a lo largo de estos 200 ficheros de opinión. La mitad de los ficheros contienen opiniones que han sido etiquetadas como positivas y la otra mitad opiniones negativas. Esta clasificación

obedece al criterio de clasificación de un experto en el dominio de las finanzas. Por otro lado, al igual que se ha hecho con el dominio de las películas cinematográficas, se necesita conocer la polaridad de cada una de las características identificadas dentro de las opiniones con el objetivo de validar los procesos de identificación y cálculo de polaridades de características dentro de la metodología propuesta en esta tesis.

El corpus utilizado para este nuevo experimento es de creación propia, debido a que se quieren analizar opiniones sobre economía y finanzas actuales (a diferencia del dominio de las películas, donde el factor temporal no es tan importante). Para ello, se han utilizado las opiniones publicadas en los sitios Web que aparecen en la tabla 4.13. Se han tenido que llevar a cabo dos acciones, por un lado el etiquetado manual de las opiniones de los usuarios, para lo que se han construido dos directorios *POS* y *NEG* donde se han colocado cada uno de los ficheros de opinión en función de su polaridad y, por otro lado, el etiquetado manual de cada una de las características identificadas en los ficheros de opinión, para lo que se ha construido un nuevo fichero con la clasificación de polaridad de cada característica identificada en cada opinión.

Este proceso de etiquetado manual por parte de un experto en el dominio de las finanzas posibilita tener un marco de referencia sobre el que trabajar. Estos resultados de referencia se emplearán para evaluar los resultados obtenidos con el sistema automático de análisis de sentimientos desarrollado en esta tesis.

Tabla 4.13. Sitios Web del dominio financiero consultados para la elaboración del corpus.

Sitios Web con información financiera
http://www.economist.com/sections/business-finance
http://www.bloombergview.com/topics/wall-street
http://www.bloombergview.com/topics/european-economy
http://www.bloombergview.com/topics/economics
http://www.reuters.com/finance/economy
http://www.reuters.com/finance/markets
http://www.bbc.com/capital

4.5.3 Ontología del dominio financiero

El dominio financiero se ha convertido en un dominio intensivo de conocimiento donde trabajan un enorme número de empresas y compañías. Las opiniones realizadas por los usuarios en este dominio tienen un gran impacto económico en nuestra sociedad (Restrepo & Medina, 2014). En consecuencia, hay una necesidad de estrategias más precisas y potentes para el almacenamiento de los datos y el conocimiento en el ámbito financiero (Pérez et al., 2002).

Las tecnologías semánticas están alcanzando cierto grado de madurez (Bernal et al., 2014) y proporcionan una base coherente y fiable para la manipulación y visualización de los datos financieros (Castells et al., 2004). En los últimos años se han desarrollado varias ontologías relacionadas con las finanzas. La ontología *BORO (Business Object Reference Ontology)* (Partridge & Stefanova, 2001) se usó como el marco adecuado para facilitar, entre otras cosas, la interoperabilidad semántica de los sistemas operativos de las empresas. Por otro lado, la ontología *TOVE (Toronto Virtual Enterprise)* (Fox et al., 1998) desarrollada por el laboratorio de Integración Empresarial de la Universidad de Toronto describe una empresa como el conjunto de los procesos que la componen. Un ejemplo adicional de ontología financiera es la ontología desarrollada por el consorcio *DIP (Data Information and Process Integration)* (Losada et al., 2005) que se centra principalmente en la descripción de servicios Web semánticos. Por último, el *XBRL Ontology Specification Group*, ha desarrollado un conjunto de ontologías para la descripción de los datos financieros y económicos en lenguaje *RDF* para compartir e intercambiar datos. Estas ontologías se están convirtiendo en un estándar abierto de comunicación electrónica sobre negocios entre las empresas, los bancos y los reguladores (XBRL International, 2009).

En la validación de la metodología propuesta en este trabajo para el dominio de las finanzas se ha desarrollado una ontología financiera partiendo de la base de las ontologías anteriormente mencionadas. La ontología ha sido creada desde cero y se ha definido en lenguaje *OWL 2* para el idioma inglés. Esta ontología tiene como objetivo proporcionar un vocabulario controlado con el que describir semánticamente las relaciones entre conceptos dentro del dominio de las finanzas. La figura 4.9 muestra un extracto de la ontología construida. Los tres conceptos principales de esta ontología son los siguientes:

- La clase **mercado financiero** representa al mecanismo que permite a la gente comprar y vender fácilmente activos financieros (acciones, materias primas o divisas), entre otros. Los principales mercados de valores (como *Nueva York Stock Exchange*, *NASDAQ* o *La Bolsa de Londres*) han sido modelados en la ontología como subclases de la clase *Stock_market*.
- La clase **intermediario financiero** representa las entidades que normalmente invierten en los mercados financieros. Como bancos, compañías de seguros, corredores y asesores financieros.
- La clase **activo** representa cualquier cosa con valor en la que un intermediario financiero puede invertir (índices bursátiles, materias primas, empresas, divisas, etc.). Así, por ejemplo, empresas como Apple, Telefónica o Microsoft

pertenecen al concepto *Company*, y monedas como *US Dólar* o el *euro* se incluyen como instancias del concepto de *Currency*.

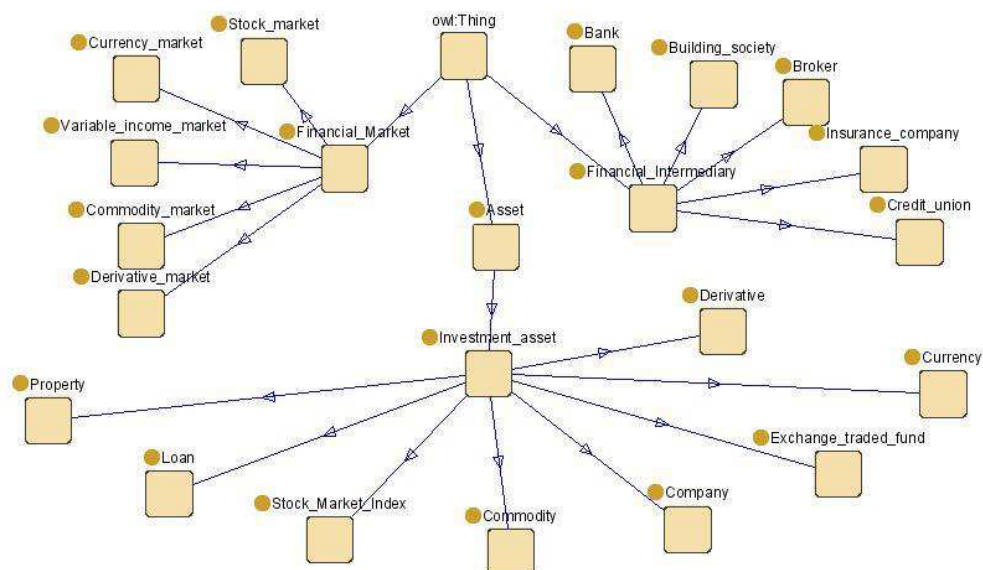


Figura 4.9. Un extracto de la ontología financiera.

4.5.4 Resultados para la búsqueda de características

Para evaluar el proceso de búsqueda de características en el dominio financiero se han utilizado las mismas métricas que para validar el dominio de las películas. Se hace así por consistencia, ya que el objetivo final es comparar los resultados obtenidos en cada dominio y sacar las conclusiones pertinentes. Estas métricas de validación son las de precisión, exhaustividad y medida-F.

Las tablas 4.14, 4.15 y 4.16 muestran los resultados experimentales obtenidos para el proceso de identificación o búsqueda de características. La primera fila en cada tabla muestra el número de opiniones utilizadas, la segunda fila el número de características que el sistema identificó, y la tercera fila el número de características reconocidas por un experto humano al leer las opiniones de los usuarios. A estas últimas las llamaremos ‘características relevantes’.

Tabla 4.14. Resultados experimentales para la identificación de características en opiniones positivas en el dominio financiero.

Opiniones Positivas	
Numero de Opiniones	100
Características identificadas por el sistema	1456
Características relevantes	1730

Precisión	1
Exhaustividad	0.841
Medida - F	0.913

Tabla 4.15. Resultados experimentales para la identificación de características en opiniones negativas en el dominio financiero.

Opiniones Negativas	
Numero de Opiniones	100
Características identificadas por el sistema	1397
Características relevantes	1692
Precisión	1
Exhaustividad	0.825
Medida - F	0.904

Tabla 4.16. Resultados experimentales totales para la identificación de características en el dominio financiero.

Opiniones Totales	
Numero de Opiniones	200
Características identificadas por el sistema	2853
Características relevantes	3422
Precisión	1
Exhaustividad	0.833
Medida - F	0.908

Como se puede apreciar en estas tablas, en términos generales el sistema obtiene buenos resultados para el proceso de identificación de características. A continuación, se detallan los resultados de precisión, exhaustividad y medida-F según han sido definidos anteriormente.

Al igual que en el dominio de las películas cinematográficas, la precisión alcanzada en el dominio financiero tanto para el conjunto de opiniones de usuarios etiquetadas como positivas como las etiquetadas como negativas (y, por tanto, también para la unión de ambas) es igual a 1 dado que todas las características identificadas por el sistema son una clase, relación o instancia de la ontología del dominio y, por tanto, relevantes.

En cuanto a la exhaustividad de las opiniones positivas, se obtiene un *0.841*. Algo parecido sucede con las opiniones etiquetadas como negativas, donde se alcanza una exhaustividad de *0.825*. Por lo tanto, la medida-F es de *0.913* y *0.904* para las opiniones positivas y negativas respectivamente. Esto indica que el método de identificación de características utilizado es bueno, ya que, al igual que para la exhaustividad, los valores para la medida-F varían entre 0 y 1. Por último, juntando todo el corpus del experimento

y calculando las métricas en cuestión, se obtiene una exhaustividad de 0.833 y una medida-F de 0.908 . Estos resultados confirman que usar una ontología del dominio en el proceso de identificación de características obtiene buenos resultados.

En la figura 4.10 se muestra, a través de una comparativa visual de estos valores, que la precisión para cualquier subconjunto del corpus utilizado y el corpus en su totalidad es siempre uno, ya que cualquier característica que el sistema identifica es relevante. La mayor exhaustividad en el proceso de identificación de características se consigue para el subconjunto de opiniones del corpus formado por las opiniones positivas, siendo esta ligeramente superior a la exhaustividad conseguida para opiniones negativas y para la totalidad del corpus. Este mismo resultado se repite para la medida-F donde, de nuevo, se obtienen los mejores resultados para las opiniones positivas. Los resultados tienen diferencias poco significativas como para concluir nada con respecto a que los mejores resultados se obtengan con las opiniones positivas.

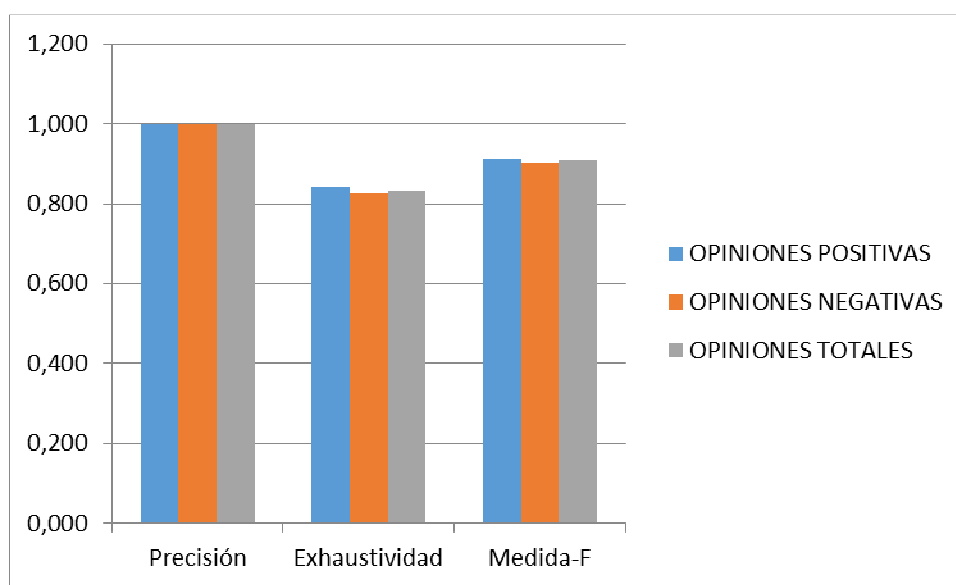


Figura 4.10. Comparativa visual de identificación de características en el dominio de las finanzas.

4.5.5 Resultados para la clasificación de sentimientos de características

Para evaluar el proceso automático de cálculo de la polaridad de las características se ha construido manualmente un fichero de validación. Este fichero contiene la clasificación de sentimiento realizada por un experto del dominio para cada una de las características

extraídas de las opiniones. Este fichero es utilizado como entrada del proceso de validación. La tabla 4.17 muestra un extracto de la estructura de este fichero. En la tabla se muestra que en la opinión “neg2.txt” la característica “Germany” tiene una evaluación positiva (“POS”), pero que otra característica como “China” tiene una polaridad negativa (“NEG”). Además, el nombre del fichero dónde está contenida la opinión del usuario (“neg2.txt”) indica que la polaridad global de esta opinión es negativa.

Tabla 4.17. Extracto del fichero de entrada para la validación en el dominio financiero.

Fichero de Opinión	Característica Identificada	Polaridad Manual
neg1.txt	Ibex 35	NEG
neg1.txt	corrections	NEG
neg1.txt	points	NEG
neg1.txt	Telefonica	NEG
neg1.txt	Convertible bonds	NEU
neg1.txt	euros	NEG
neg1.txt	investors	NEU
neg1.txt	President	NEG
neg1.txt	Banco Santander	POS
neg2.txt	inflation	POS
neg2.txt	Germany	POS
neg2.txt	China	NEG
neg2.txt	sales	NEG
neg2.txt	investors	NEG
neg2.txt	sale	NEG
neg2.txt	Americans	NEG
neg2.txt	unemployment	NEU
neg2.txt	Wall Street	NEG
neg2.txt	European market	NEG
neg2.txt	New York Stock	POS

Al igual que se hizo en el experimento anterior, una vez que el sistema ha llevado a cabo el proceso de cálculo de la polaridad de las características, se compararán las anotaciones manuales del fichero de validación de entrada con los resultados de salida producidos por el sistema. El objetivo de esta comparación es obtener el número de características correctamente clasificadas por parte del sistema de minería de opiniones desarrollado.

En los experimentos llevados a cabo para esta fase se han evaluado cuatro métodos diferentes para calcular la polaridad de las características identificadas dentro de las opiniones de los usuarios. Las diferencias entre estos métodos consisten únicamente en la forma en la se recuperan las palabras adyacentes a cada característica para calcular la polaridad de ésta. Será necesario calcular la polaridad de las palabras situadas cerca de la expresión lingüística que representa a una característica dentro de una opinión. Como

se ha explicado anteriormente, los cuatro métodos implementados son “*N_GRAM Before*”, “*N_GRAM After*”, “*N_GRAM Around*” y “*All Phrase*”.

Los nuevos experimentos en el dominio de las finanzas se llevan a cabo en las mismas condiciones que los experimentos realizados en el dominio de las películas cinematográficas. El objetivo es llevar a cabo una comparativa final entre los resultados obtenidos en ambos dominios. Por tanto, se estudian las mismas combinaciones de valores asignados a los parámetros del sistema *N_GRAM*, Z_1 , Z_2 y Z_3 para los experimentos llevados a cabo en el dominio financiero con los cuatro métodos de clasificación de sentimientos citados anteriormente.

Las tablas 4.18, 4.19, 4.20 y 4.21 muestran los resultados obtenidos para la clasificación de sentimientos de las características identificadas por el sistema. Se han utilizado cada uno de los métodos de clasificación de sentimientos mencionados. Además, en la figura 4.11 se presenta un gráfico que muestra una comparativa visual entre los distintos métodos de cálculo de la polaridad de las características identificadas para el dominio financiero.

Tabla 4.18. Valor de la precisión en el cálculo de la polaridad de las características con el método *N_GRAM Before*.

Casos de estudio				Precisión en el cálculo de la Polaridad de las Características					
Num.	Z ₁	Z ₂	Z ₃	NG=2	NG=3	NG=4	NG=5	NG=6	Media
1	H	H	H	46.83	60.24	59.80	59.51	58.89	57.05
2	H	H	M	52.36	60.24	59.80	59.51	58.89	58.16
3	H	H	L	52.36	60.24	59.80	59.51	57.14	57.81
4	H	M	H	52.36	60.24	59.80	59.51	58.89	58.16
5	H	M	M	52.36	60.24	59.80	59.51	58.89	58.16
6	H	M	L	52.36	60.24	59.80	59.51	58.89	58.16
7	H	L	H	56.33	60.24	61.13	61.04	60.30	59.81
8	H	L	M	52.36	66.92	59.80	61.04	58.89	59.80
9	H	L	L	52.36	60.24	59.80	59.51	58.89	58.16
10	M	H	H	52.36	60.24	59.80	57.45	58.89	57.75
11	M	H	M	52.36	60.24	59.80	57.45	58.89	57.75
12	M	H	L	52.36	60.24	59.80	59.51	58.89	58.16
13	M	M	H	52.36	60.24	59.80	59.51	58.89	58.16
14	M	M	M	46.83	60.24	59.80	59.51	59.13	57.10
15	M	M	L	52.36	60.24	59.80	59.51	58.89	58.16
16	M	L	H	52.36	63.85	58.10	59.51	58.89	58.54
17	M	L	M	52.36	60.24	59.80	57.45	58.89	57.75
18	M	L	L	52.33	60.24	59.80	59.51	58.89	58.15
19	L	H	H	52.33	60.24	59.80	59.51	58.89	58.15
20	L	H	M	52.36	60.24	59.80	59.51	57.14	57.81
21	L	H	L	52.33	60.24	59.80	59.51	57.14	57.80
22	L	M	H	52.36	60.24	59.80	59.51	57.14	57.81
23	L	M	M	52.36	60.24	59.80	59.51	58.89	58.16
24	L	M	L	52.36	60.24	59.80	59.51	58.89	58.16
25	L	L	H	52.36	63.85	58.10	57.45	59.13	58.18
26	L	L	M	52.36	63.85	58.10	57.45	58.89	58.13
27	L	L	L	46.83	60.24	59.80	59.51	58.89	57.05
Media				51.89	60.89	59.66	59.24	58.70	58.08

Tabla 4.19. Valor de la precisión en el cálculo de la polaridad de las características con el método *N_GRAM After*.

Casos de estudio				Precisión en el cálculo de la Polaridad de las Características					
Num.	Z ₁	Z ₂	Z ₃	NG=2	NG=3	NG=4	NG=5	NG=6	Media
1	H	H	H	55.87	60.44	57.69	57.55	59.95	58.30
2	H	H	M	55.87	59.02	57.55	54.44	52.61	55.90
3	H	H	L	55.87	59.02	57.55	54.44	52.61	55.90
4	H	M	H	55.87	60.44	57.69	57.55	59.95	58.30
5	H	M	M	55.87	60.44	57.69	57.55	59.95	58.30
6	H	M	L	56.03	60.44	57.69	57.55	59.95	58.33
7	H	L	H	56.03	60.44	57.69	57.55	59.95	58.33
8	H	L	M	55.87	60.44	57.69	57.55	60.14	58.34
9	H	L	L	55.87	60.44	57.69	58.22	59.95	58.43
10	M	H	H	55.87	61.58	58.49	58.22	59.95	58.82
11	M	H	M	55.87	60.44	57.69	57.55	59.95	58.30
12	M	H	L	53.05	60.44	57.69	57.55	59.95	57.74
13	M	M	H	53.05	60.44	57.69	57.55	52.61	56.27
14	M	M	M	55.87	60.44	57.69	57.55	52.61	56.83
15	M	M	L	55.87	59.02	57.69	57.55	52.61	56.55
16	M	L	H	55.87	60.44	57.69	57.55	52.61	56.83
17	M	L	M	55.87	60.44	57.69	54.44	59.95	57.68
18	M	L	L	55.87	60.44	57.69	54.44	59.95	57.68
19	L	H	H	55.87	60.44	57.69	54.44	59.95	57.68
20	L	H	M	55.87	60.44	57.69	54.44	59.95	57.68
21	L	H	L	56.03	60.44	57.55	57.55	59.95	58.30
22	L	M	H	56.03	60.44	57.55	57.55	59.95	58.30
23	L	M	M	55.87	60.44	57.55	57.55	59.95	58.27
24	L	M	L	53.05	59.02	57.69	57.55	59.95	57.45
25	L	L	H	53.05	59.02	57.69	57.55	52.61	55.98
26	L	L	M	53.05	59.02	57.69	57.55	52.61	55.98
27	L	L	L	55.87	60.44	57.69	57.55	59.95	58.30
Media				55.37	60.17	57.69	56.91	57.78	57.58

Tabla 4.20. Valor de la precisión en el cálculo de la polaridad de las características con el método *N_GRAM Around*.

Casos de estudio				Precisión en el cálculo de la Polaridad de las Características					
Num.	Z ₁	Z ₂	Z ₃	NG=2	NG=3	NG=4	NG=5	NG=6	Media
1	H	H	H	59.29	61.72	58.83	60.55	60.37	60,15
2	H	H	M	59.29	59.04	58.83	60.55	59.66	59,47
3	H	H	L	59.29	59.04	55.40	58.96	59.66	58,47
4	H	M	H	59.29	61.72	55.40	58.96	60.37	59,15
5	H	M	M	59.29	61.72	58.83	60.55	60.37	60,15
6	H	M	L	59.29	61.72	58.83	60.55	60.37	60,15
7	H	L	H	59.29	61.72	58.83	60.55	60.37	60,15
8	H	L	M	59.29	61.25	58.83	60.55	60.37	60,06
9	H	L	L	59.29	62.25	59.47	61.72	61.28	60,80
10	M	H	H	60.09	62.72	59.47	61.72	61.28	61,06
11	M	H	M	59.29	61.72	59.47	60.55	60.37	60,28
12	M	H	L	59.29	61.72	59.47	60.55	60.37	60,28
13	M	M	H	59.29	61.72	58.83	60.55	60.37	60,15
14	M	M	M	59.29	61.72	58.83	60.55	60.37	60,15
15	M	M	L	57.23	61.72	58.83	60.55	60.37	59,74
16	M	L	H	57.23	61.72	58.83	60.55	60.37	59,74
17	M	L	M	57.23	61.72	58.83	60.55	60.37	59,74
18	M	L	L	59.29	61.72	58.83	60.55	60.37	60,15
19	L	H	H	59.29	62.25	58.83	58.96	60.37	59,94
20	L	H	M	59.29	62.22	58.83	58.96	59.66	59,79
21	L	H	L	59.29	61.72	55.40	58.96	59.66	59,01
22	L	M	H	59.29	61.72	55.40	58.96	59.66	59,01
23	L	M	M	59.29	61.72	55.40	60.55	59.66	59,32
24	L	M	L	57.23	59.04	55.40	60.55	60.37	58,52
25	L	L	H	57.23	59.04	58.83	60.55	60.37	59,20
26	L	L	M	59.29	61.72	58.83	60.55	60.37	60,15
27	L	L	L	59.29	61.72	58.83	60.55	60.37	60,15
Media				58.94	61.40	58.16	60.28	60.28	59,81

Tabla 4.21. Valor de la precisión en el cálculo de la polaridad de las características con el método *All_Phrase*.

Casos de estudio				Precisión en el cálculo de la Polaridad de las Características
Num.	Z ₁	Z ₂	Z ₃	
1	H	H	H	61.04
2	H	H	M	61.04
3	H	H	L	61.04
4	H	M	H	61.04
5	H	M	M	61.04
6	H	M	L	61.04
7	H	L	H	62.22
8	H	L	M	62.22
9	H	L	L	62.22
10	M	H	H	61.04
11	M	H	M	61.04
12	M	H	L	61.04
13	M	M	H	61.04
14	M	M	M	61.04
15	M	M	L	60.19
16	M	L	H	61.04
17	M	L	M	61.04
18	M	L	L	61.04
19	L	H	H	61.04
20	L	H	M	61.04
21	L	H	L	61.04
22	L	M	H	61.04
23	L	M	M	61.04
24	L	M	L	60.19
25	L	L	H	60.19
26	L	L	M	60.19
27	L	L	L	61.04
Media				61.05

Los resultados del método “*N_GRAM Before*” para el cálculo de la polaridad de características se muestran en la tabla 4.18. La mejor media de precisión, con un valor del 60.89%, se obtiene con *N_GRAM* igual a 3. Esto significa que el sistema obtiene los mejores resultados cuando tiene en cuenta los valores de sentimiento de las 3 palabras que preceden a la expresión lingüística que representa a la característica en la opinión del usuario. Además, para este método los peores resultados se alcanzan con *N_GRAM* igual a 2, obteniendo una precisión media del 51.89%. La precisión máxima obtenida con este método es de 66.92% cuando *N_GRAM* es igual a 3 y $Z_1 = H$, $Z_2 = L$ y $Z_3 = M$. Por otro lado, la tasa de éxito promedio no varía significativamente con las 27 configuraciones de los parámetros Z_1 , Z_2 y Z_3 . El mejor resultado medio para las diferentes configuraciones de los parámetros Z_1 , Z_2 y Z_3 en este método se obtiene para $Z_1 = H$, $Z_2 = L$ y $Z_3 = H$, y es de un 59.81%. Por último, la media en la tasa de éxito de todas las combinaciones de parámetros para este método en el cálculo de polaridades de características es del 58.08%.

Al final de esta sección se profundiza en el análisis de estos resultados mediante un gráfico clarificador. Por el momento se destacan los resultados más relevantes obtenidos al aplicar cada uno de los métodos de clasificación de sentimientos citados anteriormente.

La tabla 4.19 muestra los resultados adquiridos en el cálculo de la polaridad de las características utilizando el método “*N_GRAM After*” en el dominio financiero. Los resultados son bastante igualados con respecto a los logrados con el método “*N_GRAM Before*” en el dominio financiero. Esto no ocurría en el dominio de las películas cinematográficas, donde los resultados eran más dispares. En la siguiente sección, de comparativa de los resultados alcanzados en ambos dominios, se explican estas diferencias. La mejor tasa media de éxito en el proceso de identificación de polaridades se obtiene con *N_GRAM* igual a 3, siendo esta del 60.17%. Además, se obtiene una precisión máxima del 61.58%, que es un resultado inferior a la precisión máxima obtenida con el método anterior. Esta precisión máxima se alcanza cuando *N_GRAM* es igual a 3 y $Z_1 = M$, $Z_2 = H$ y $Z_3 = H$. Por último, la tasa media de éxito de todas las combinaciones de parámetros para este método en el cálculo de polaridades de características es del 57.58%.

La tabla 4.20 muestra los resultados obtenidos para el cálculo de polaridades de características con el método “*N_GRAM Around*” en el dominio financiero. Esta tabla muestra que, en términos generales, los resultados cosechados con este método son bastante semejantes a los obtenidos con los métodos “*N_GRAM After*” y “*N_GRAM Before*”. La mejor tasa de acierto para el proceso de identificación de polaridades de características se obtiene con *N_GRAM* igual a 3 con una precisión máxima del 62.72% y una precisión media del 61.40%. Esto significa que los mejores resultados se alcanzan teniendo en cuenta las 3 palabras que preceden a cada característica en la opinión del usuario junto con las 3 palabras que van inmediatamente después. En términos generales, este método obtiene mejores resultados que los métodos “*N_GRAM After*” y “*N_GRAM Before*” para cualquier valor de *N_GRAM*. El mejor resultado medio para las diferentes configuraciones de los parámetros Z_1 , Z_2 y Z_3 con este método se obtiene cuando $Z_1 = H$, $Z_2 = L$ y $Z_3 = L$, y es de un 60,80%. Por último, la media en la tasa de éxito de todas las combinaciones de parámetros para este método es del 59,81%, lo cual supone un valor superior a los obtenidos para los métodos “*N_GRAM After*” (57.58%) y “*N_GRAM Before*” (58.08%).

La tabla 4.21 muestra los resultados obtenidos por el sistema para la clasificación de características cuando se utiliza el método “*All Phrase*” en el dominio financiero. Por

término medio, los resultados alcanzados con este método son mejores que los obtenidos con los métodos de clasificación “*N_GRAM Before*”, “*N_GRAM After*” y “*N_GRAM Around*”. La tasa media de éxito para el cálculo de la polaridad de características es del 61.05%, lo cual supone el mejor valor de todos los obtenidos anteriormente en este dominio.

Al igual que hicimos para el dominio cinematográfico, la figura 4.11 muestra una comparativa visual entre los distintos métodos de cálculo de la polaridad de las características identificadas en el dominio financiero. La gráfica muestra que los mejores resultados medios se obtienen con el método de clasificación “*All_Phrase*”, cuyos resultados son superiores a todos los métodos excepto para *N_GRAM* igual a 3, donde el mejor método es “*N_GRAM_Around*”. Los peores resultados se obtienen cuando el parámetro *N_GRAM* es igual a 2.

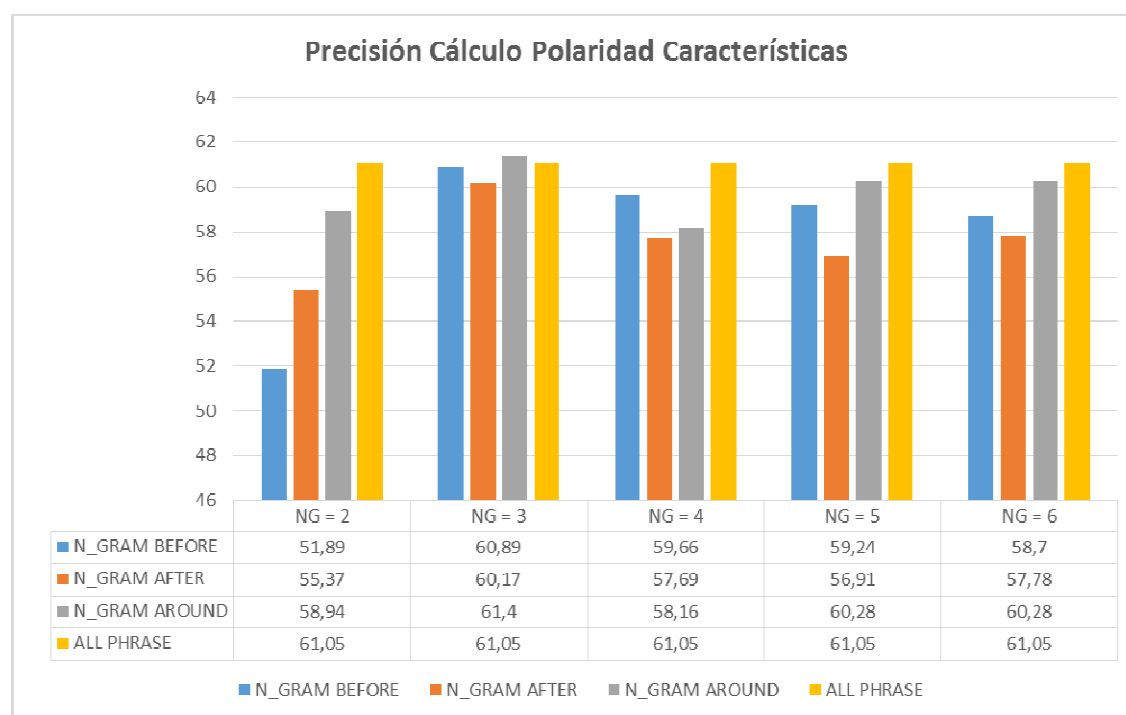


Figura 4.11. Comparativa visual de precisión en el cálculo de la polaridad de las características en el dominio financiero.

4.5.6 Resultados para la clasificación de sentimientos de opiniones

Al igual que sucedió para el dominio de las películas cinematográficas, para evaluar el proceso de clasificación de sentimientos de las opiniones dadas por los usuarios en el dominio financiero, se calcula la polaridad global de cada opinión a través del sistema automático de análisis de sentimientos y, después, se comparan estos resultados con la clasificación de sentimientos llevada a cabo por los expertos de dominio.

Se han evaluado los cuatro métodos vistos anteriormente para llevar a cabo el análisis de sentimientos de las opiniones de los usuarios: “*N_GRAM Before*”, “*N_GRAM After*”, “*N_GRAM Around*” y “*All Phrase*”. Además, se ha experimentado con los diferentes valores de los parámetros Z_1 , Z_2 y Z_3 y el parámetro del sistema *N_GRAM*.

Las tablas 4.22, 4.23, 4.24 y 4.25 muestran los resultados obtenidos para la clasificación de sentimientos de las opiniones de los usuarios en el dominio financiero.

Los resultados para el método “*N_GRAM Before*” en el análisis de sentimientos de opiniones se muestran en la tabla 4.22. La mejor tasa de éxito para el proceso de análisis de sentimientos que se obtiene es del 80,05%. Este valor se obtiene para los valores del parámetro *N_GRAM* igual a 3, 5 y 6. El mejor valor promedio de precisión se obtiene cuando el parámetro *N_GRAM* es igual a 6, siendo este de un 74.55%.

La figura 4.12 refleja la influencia de los parámetros del sistema Z_1 , Z_2 y Z_3 en el análisis de sentimientos de opiniones en el dominio financiero usando el método “*N_GRAM Before*”. Se aprecia que los valores más altos de precisión se obtienen con diversas configuraciones de valores de Z_1 , Z_2 y Z_3 cuando los valores del parámetro *N_GRAM* son 5 y 6. Con estas configuraciones de parámetros se obtienen valores del 80.05% de precisión en el análisis de sentimientos.

Este resultado pone de manifiesto que la hipótesis de que las características situadas al principio y al final de las opiniones tienen más peso a la hora de evaluar la opinión global del usuario no se cumple de manera tan estricta para el dominio financiero, como si ocurría para el dominio cinematográfico.

El motivo de que se produzca este hecho es la distinta naturaleza de los corpus empleados en ambos experimentos:

- En el dominio cinematográfico los usuarios no tenían límite para expresar sus opiniones y se dispersaban en la parte media de las mismas, dando más énfasis al principio y al final de la opinión.

- En el dominio financiero se ha empleado un corpus que recoge columnas de opinión de periodistas en el ámbito financiero. Esto influye en la manera en la que opinan este tipo de usuarios semi-expertos, ya que tienen un espacio limitado para expresar su opinión. Por lo tanto, utilizan frases cortas y construidas de modo directo porque son las más eficaces para transmitir una idea. Esto redundo en una estructura de opinión muy concreta donde las características situadas en la parte media de la misma tienen la misma relevancia que las situadas al inicio y al final de la frase. Por este motivo, la influencia de la configuración de los parámetros Z_1 , Z_2 y Z_3 en el dominio financiero es menor que el dominio cinematográfico.

Tabla 4.22. Valor de la precisión en el análisis de sentimientos de opiniones con el método *N_GRAM Before*.

Casos de estudio				Análisis de sentimientos						
Num.	Z ₁	Z ₂	Z ₃	NG=2	NG=3	NG=4	NG=5	NG=6	Media	
1	H	H	H	70.23	70.23	60.12	80.00	80.00	72,12	
2	H	H	M	70.23	70.23	60.12	70.23	70.23	68,21	
3	H	H	L	70.23	70.23	60.12	70.23	70.23	68,21	
4	H	M	H	70.23	70.23	60.12	70.23	70.23	68,21	
5	H	M	M	70.23	70.23	60.12	70.23	70.23	68,21	
6	H	M	L	70.23	70.23	60.12	70.23	60.12	66,19	
7	H	L	H	70.23	70.23	60.12	70.23	70.23	68,21	
8	H	L	M	70.23	70.23	60.12	50.00	60.12	62,14	
9	H	L	L	70.23	70.23	60.12	50.00	60.12	62,14	
10	M	H	H	70.23	70.23	70.23	80.00	80.00	74,14	
11	M	H	M	70.23	70.23	70.23	70.23	80.00	72,18	
12	M	H	L	70.23	70.23	70.23	80.00	80.00	74,14	
13	M	M	H	70.23	70.23	70.23	80.00	80.00	74,14	
14	M	M	M	70.23	70.23	60.12	80.00	80.00	72,12	
15	M	M	L	70.23	70.23	60.12	70.23	70.23	68,21	
16	M	L	H	70.23	70.23	60.12	70.23	70.23	68,21	
17	M	L	M	70.23	70.23	60.12	70.23	70.23	68,21	
18	M	L	L	70.23	70.23	60.12	70.23	70.23	68,21	
19	L	H	H	70.23	70.23	70.23	80.05	80.05	74,16	
20	L	H	M	70.23	70.23	70.23	80.05	80.05	74,16	
21	L	H	L	70.23	80.05	70.23	80.05	80.05	76,12	
22	L	M	H	60.12	70.23	70.23	80.05	80.05	72,14	
23	L	M	M	70.23	70.23	70.23	80.05	80.05	74,16	
24	L	M	L	70.23	70.23	70.23	80.05	80.05	74,16	
25	L	L	H	60.12	70.23	70.23	80.05	80.05	72,14	
26	L	L	M	70.23	70.23	70.23	80.05	80.05	74,16	
27	L	L	L	70.23	70.23	60.12	80.05	80.05	72,14	
Media				69,48	70,59	64,61	73,81	74,55	70,61	

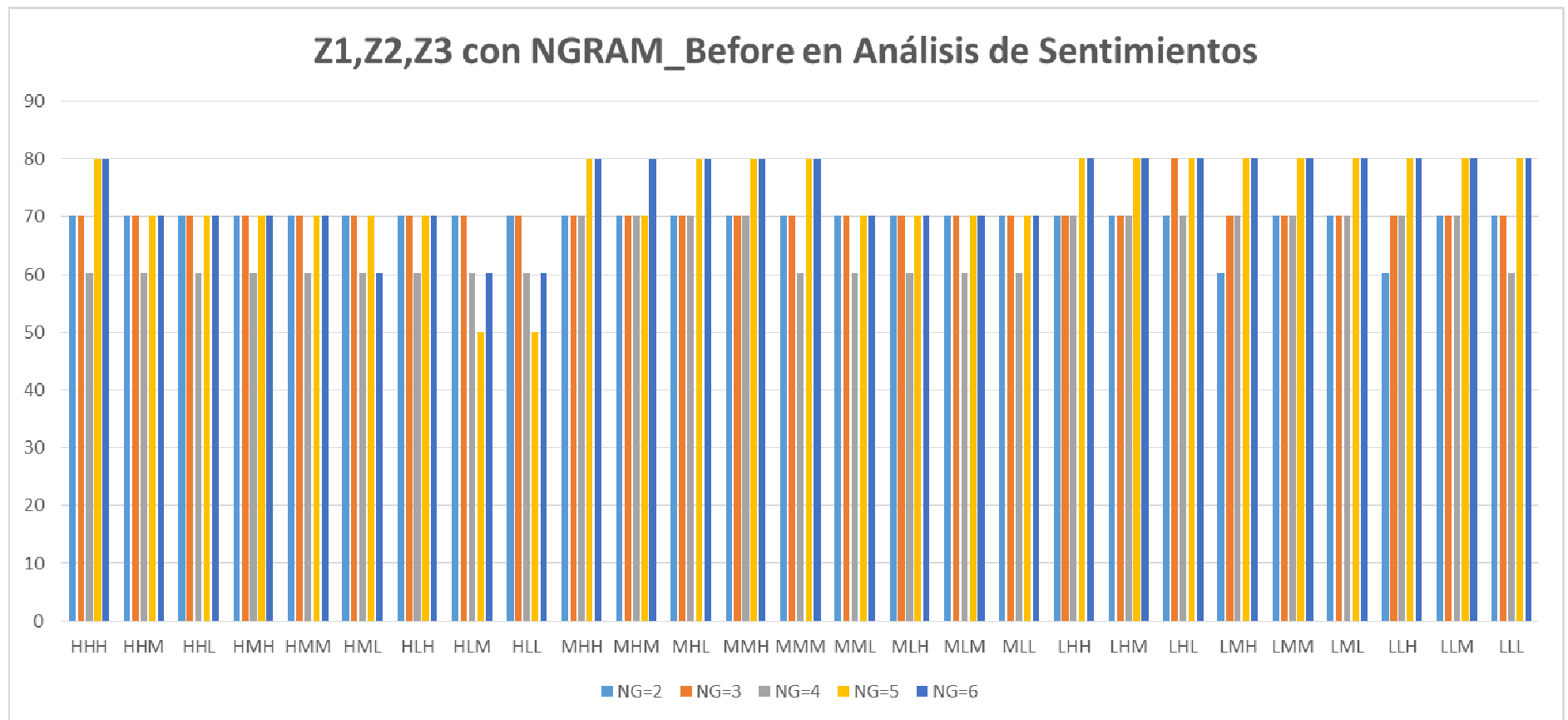


Figura 4.12. Efecto de los parámetros Z_1 , Z_2 y Z_3 en el análisis de sentimientos de opiniones con el método *N_GRAM Before*.

En la tabla 4.23 se muestran los resultados obtenidos en el análisis de sentimientos de opiniones de usuarios en el dominio financiero utilizando el método “*N_GRAM After*”. Destacan los buenos resultados obtenidos con este método para este dominio ya que en el dominio cinematográfico se obtuvieron malos resultados al aplicarlo. Se ha obtenido un 70.85% de precisión media tomando en cuenta todas las posibles configuraciones de los parámetros *N_GRAM* y Z_1 , Z_2 y Z_3 , lo cual supone un resultado ligeramente superior a “*N_GRAM Before*” (70.61%). Además, existen varias configuraciones de parámetros que obtienen picos de precisión que alcanzan el 80.05%, concretamente 58 veces, lo cual supone un muy buen resultado. Para analizar las causas que dan lugar a estos resultados, tenemos que analizar las opiniones de los usuarios utilizadas para evaluar este dominio que, como sabemos, son bastante diferentes de las opiniones empleadas en el dominio cinematográfico. Al analizar las opiniones de los usuarios en el dominio financiero se ha visto que expresiones del tipo “*The Ibex 35 has fallen back and has four consecutive corrections*”, o “*Telefónica has lost 21.97% after Siterday announced that it will issue convertible bonds amounting to 1,500 million euros*”, o “*Banco Sabadell has announced a quarterly net profit slightly worse than expected: 81,2M €*”, son muy frecuentes. Partiendo de este hecho se ve que la carga emocional en estas frases para las características identificadas *Ibex 35*, *Telefónica* o *Banco Sabadell* está después de la expresión lingüística que identifica a la característica, con lo cual el método de clasificación de sentimientos “*N_GRAM After*” funciona muy bien en estos casos. Este patrón se repite con frecuencia en el dominio financiero, donde, por ejemplo, se tiende a nombrar a una empresa antes de dar una valoración sobre la misma. Esto deriva en los buenos resultados que obtiene este método de clasificación de sentimientos en este dominio.

En la figura 4.13 se muestra que este método obtiene picos de precisión para bastantes configuraciones de los parámetros Z_1 , Z_2 y Z_3 , llegando al 80.05% de precisión máxima en diversos valores de *N_GRAM*. Debido a la diversidad de los resultados obtenidos, no se puede destacar ninguna configuración de Z_1 , Z_2 y Z_3 como la óptima. Tampoco se puede concluir que la hipótesis de que las características situadas al principio y al final de la opinión del usuario tienen más peso a la hora de evaluar la opinión global del mismo, ya que los resultados no corroboran eso para este dominio. Las causas que explican este hecho son las mismas que las mencionadas anteriormente para el método “*N_GRAM Before*”.

Tabla 4.23. Valor de la precisión en el análisis de sentimientos de opiniones con el método *N_GRAM After*.

Casos de estudio				Análisis de sentimientos					
Num.	Z ₁	Z ₂	Z ₃	NG=2	NG=3	NG=4	NG=5	NG=6	Media
1	H	H	H	60.12	70.23	80.05	70.23	80.05	72,14
2	H	H	M	60.12	60.12	80.05	70.23	80.05	70,11
3	H	H	L	60.12	60.12	70.23	70.23	80.05	68,15
4	H	M	H	60.12	60.12	60.12	60.12	70.23	62,14
5	H	M	M	60.12	60.12	60.12	60.12	60.12	60,12
6	H	M	L	60.12	60.12	60.12	60.12	60.12	60,12
7	H	L	H	60.12	60.12	60.12	60.12	60.12	60,12
8	H	L	M	60.12	60.12	60.12	60.12	60.12	60,12
9	H	L	L	60.12	60.12	60.12	60.12	60.12	60,12
10	M	H	H	80.05	70.23	80.05	80.05	80.05	78,09
11	M	H	M	80.05	70.23	80.05	80.05	80.05	78,09
12	M	H	L	60.12	70.23	80.05	70.23	80.05	72,14
13	M	M	H	80.05	70.23	80.05	80.05	80.05	78,09
14	M	M	M	60.12	70.23	80.05	70.23	80.05	72,14
15	M	M	L	60.12	60.12	80.05	70.23	80.05	70,11
16	M	L	H	80.05	70.23	70.23	70.23	70.23	72,19
17	M	L	M	60.12	60.12	60.12	60.12	70.23	62,14
18	M	L	L	60.12	60.12	60.12	60.12	60.12	60,12
19	L	H	H	80.05	70.23	80.05	80.05	80.05	78,09
20	L	H	M	80.05	70.23	80.05	80.05	80.05	78,09
21	L	H	L	80.05	70.23	80.05	80.05	80.05	78,09
22	L	M	H	80.05	70.23	80.05	80.05	80.05	78,09
23	L	M	M	80.05	70.23	80.05	80.05	80.05	78,09
24	L	M	L	80.05	70.23	80.05	80.05	80.05	78,09
25	L	L	H	80.05	70.23	80.05	80.05	80.05	78,09
26	L	L	M	80.05	70.23	80.05	80.05	80.05	78,09
27	L	L	L	60.12	70.23	80.05	70.23	80.05	72,14
Media				68,98	66,11	73,42	71,24	74,53	70,85

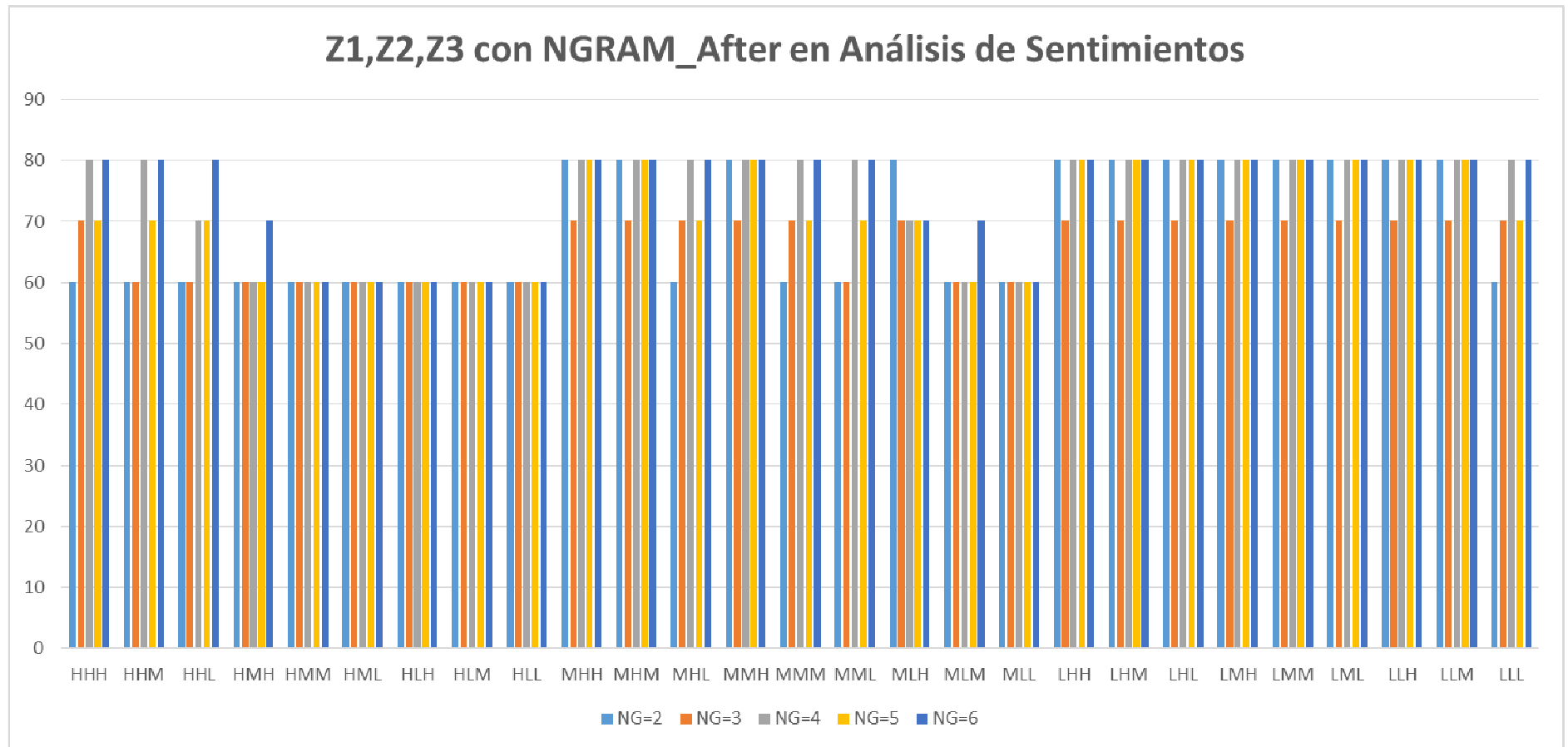


Figura 4.13. Efecto de los parámetros Z_1 , Z_2 y Z_3 en el análisis de sentimientos de opiniones con el método *N_GRAM After*.

En la tabla 4.24 se muestran los resultados obtenidos para la clasificación de sentimientos de opiniones utilizando el método “*N_GRAM Around*”. Este método obtiene un 72.44% de precisión media para todas las posibles configuraciones de los parámetros Z_1 , Z_2 y Z_3 y N_GRAM , lo cual mejora los resultados obtenidos para “*N_GRAM Before*” (70.61%) y “*N_GRAM After*” (70.85%). Se trata de un método que se comporta de manera favorable y fiable en el análisis de sentimientos de opiniones, ya que tanto para el dominio financiero como para el dominio cinematográfico ha obtenido buenos resultados en términos generales. Además, con este método se obtiene el pico de precisión más alto para el análisis de sentimientos de opiniones tanto en el dominio financiero como en el dominio cinematográfico. En concreto, se obtiene un pico de precisión del 90.03% para el valor de N_GRAM igual a 3 y los valores $Z_1 = M$, $Z_2 = H$ y $Z_3 = L$. También se obtiene un valor del 90% de precisión para esta misma configuración de Z_1 , Z_2 y Z_3 cuando N_GRAM es igual a 6. Además, el mejor valor promedio para los valores de N_GRAM se obtiene para N_GRAM igual a 6, con un valor del 77.15%.

La figura 4.14 muestra la influencia de los parámetros del sistema Z_1 , Z_2 y Z_3 en el análisis de sentimientos de opiniones utilizando este método. Al igual que en los métodos anteriores, debido a la diversidad de los resultados obtenidos, no se puede destacar ninguna configuración de Z_1 , Z_2 y Z_3 como la óptima. Tampoco se puede concluir que la afirmación de que las características situadas al principio y al final de la opinión del usuario tienen más peso a la hora de evaluar la opinión global del mismo, ya que los resultados no corroboran eso para este dominio. Las causas que explican este hecho son las mismas que las explicadas para los métodos anteriores. Se destacan los picos de precisión para $Z_1 = M$, $Z_2 = H$ y $Z_3 = L$ y N_GRAM igual a 3 y 6. Además, se ve que los peores resultados se obtienen para cualquier configuración de Z_1 , Z_2 y Z_3 cuando N_GRAM es igual a 2. Pero, en general, se obtienen muy buenos resultados promedio para cualquier valor de Z_1 , Z_2 y Z_3 cuando utilizamos este método de clasificación de sentimientos. Por tanto, se puede concluir que el método “*N_GRAM Around*” es un método óptimo para el análisis de sentimientos de las opiniones escritas en inglés para el dominio de las finanzas.

Tabla 4.24. Valor de la precisión en el análisis de sentimientos de opiniones con el método *N_GRAM Around*.

Casos de estudio				Análisis de sentimientos					
Num.	Z ₁	Z ₂	Z ₃	NG=2	NG=3	NG=4	NG=5	NG=6	Media
1	H	H	H	60.12	70.23	70.23	70.23	70.23	68.21
2	H	H	M	60.12	70.23	80.05	70.23	80.05	72.14
3	H	H	L	60.12	70.23	80.05	70.23	80.05	72.14
4	H	M	H	60.12	70.23	80.05	70.23	80.05	72.14
5	H	M	M	60.12	70.23	80.05	70.23	80.05	72.14
6	H	M	L	60.12	70.23	70.23	70.23	80.05	70.17
7	H	L	H	60.12	70.23	80.05	80.05	80.05	74.10
8	H	L	M	60.12	70.23	70.23	70.23	70.23	68.21
9	H	L	L	60.12	70.23	70.23	70.23	70.23	68.21
10	M	H	H	70.23	80.05	70.23	80.05	80.05	76.12
11	M	H	M	70.23	80.05	70.23	80.05	80.05	76.12
12	M	H	L	70.23	90.03	80.05	80.05	90.00	82.07
13	M	M	H	60.12	70.23	70.23	70.23	70.23	68.21
14	M	M	M	60.12	70.23	70.23	70.23	70.23	68.21
15	M	M	L	60.12	70.23	80.05	70.23	80.05	72.14
16	M	L	H	60.12	60.12	70.23	70.23	70.23	66.19
17	M	L	M	60.12	70.23	80.05	70.23	80.05	72.14
18	M	L	L	60.12	70.23	80.05	70.23	80.05	72.14
19	L	H	H	80.05	80.05	70.23	80.05	80.05	78.09
20	L	H	M	80.05	80.05	70.23	80.05	80.05	78.09
21	L	H	L	80.05	80.05	70.23	80.05	80.05	78.09
22	L	M	H	60.12	80.05	70.23	80.05	80.05	74.10
23	L	M	M	70.23	80.05	70.23	80.05	80.05	76.12
24	L	M	L	70.23	80.05	70.23	80.05	80.05	76.12
25	L	L	H	60.12	70.23	70.23	70.23	70.23	68.21
26	L	L	M	60.12	70.23	70.23	70.23	70.23	68.21
27	L	L	L	60.12	70.23	70.23	70.23	70.23	68.21
Media				64.21	73.50	73.50	73.87	77.15	72.44

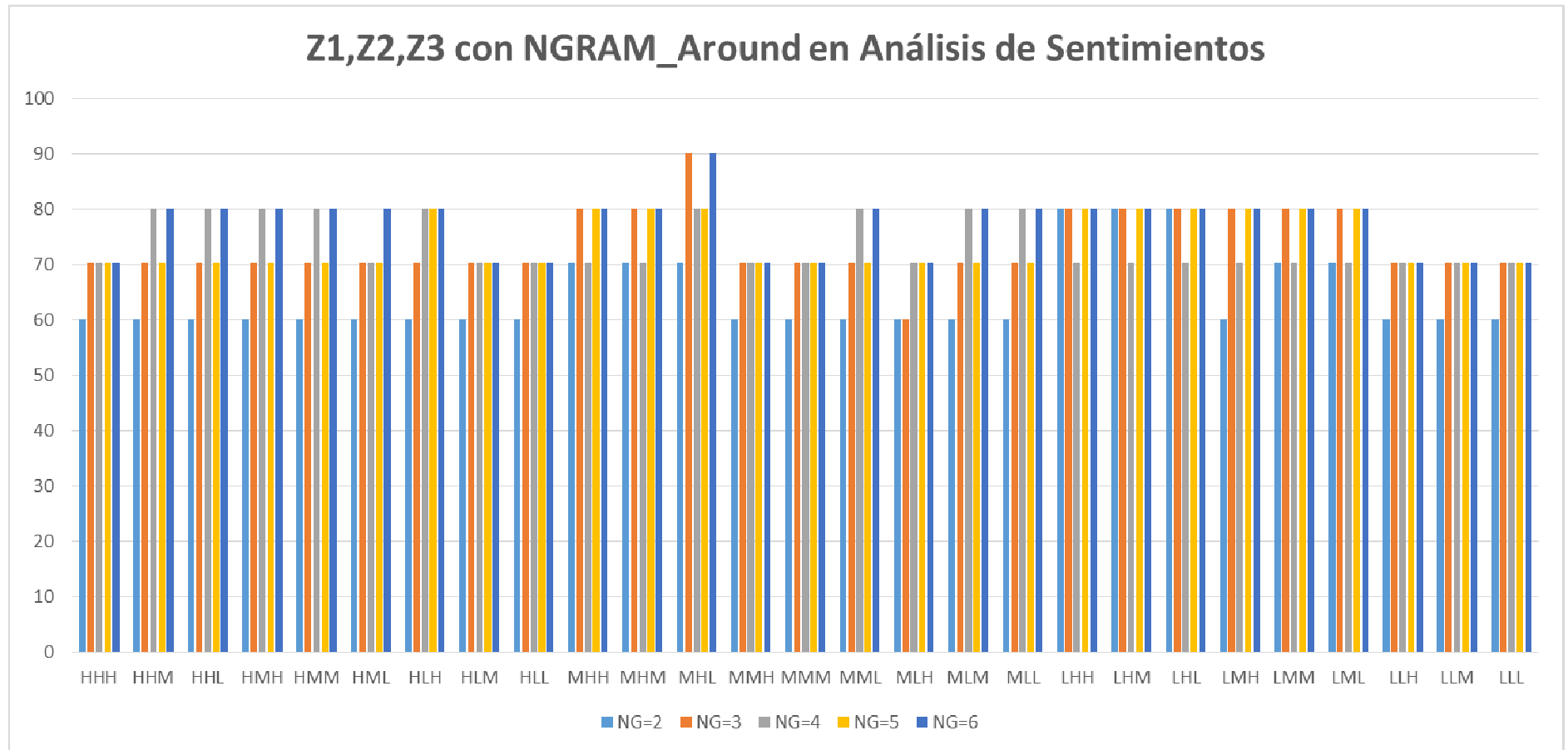


Figura 4.14. Efecto de los parámetros Z_1 , Z_2 y Z_3 en el análisis de sentimientos de opiniones con el método N_GRAM Around.

Por último, la tabla 4.25 muestra los resultados obtenidos para la clasificación de sentimientos de opiniones usando el método “*All Phrase*”. En este método no se utiliza el parámetro del sistema *N_GRAM* ya que se analiza cada frase en su totalidad. Los resultados obtenidos son prometedores, ya que se obtiene un 74.27% de valor promedio en el cálculo de la precisión de la clasificación de sentimientos, que supone el mejor valor promedio de todos los métodos estudiados. El valor máximo de precisión obtenido con este método es del 81.68%. Esto nos indica que, en el dominio financiero, es una buena idea analizar toda la frase que contiene la expresión lingüística de la característica identificada a la hora de calcular la polaridad de la opinión del usuario. Esto es así porque las opiniones de los usuarios contienen frases cortas, muy directas y prácticamente sin ambigüedad (columnas de opinión de periodistas expertos en el dominio financiero). En estas condiciones el método “*All Phrase*” se comporta muy bien, a diferencia del dominio cinematográfico donde la mayoría de las frases son largas, con subordinadas, y poco directas.

Tabla 4.25. Valor de la precisión en el análisis de sentimientos de opiniones con el método *All Phrase*.

Casos de estudio			Análisis de sentimientos
Num.	Z ₁	Z ₂	Z ₃
1	H	H	H
2	H	H	M
3	H	H	L
4	H	M	H
5	H	M	M
6	H	M	L
7	H	L	H
8	H	L	M
9	H	L	L
10	M	H	H
11	M	H	M
12	M	H	L
13	M	M	H
14	M	M	M
15	M	M	L
16	M	L	H
17	M	L	M
18	M	L	L
19	L	H	H
20	L	H	M
21	L	H	L
22	L	M	H
23	L	M	M
24	L	M	L
25	L	L	H
26	L	L	M
27	L	L	L
Media			74.27

La figura 4.15, muestra la incidencia de los parámetros Z_1 , Z_2 y Z_3 en la precisión del análisis de sentimientos de opiniones usando el método “*All Phrase*”. En la gráfica se observa claramente que el mejor resultado se obtiene cuando para $Z_1 = H$, $Z_2 = L$ y $Z_3 = H$, con lo cual, en este caso si se confirmaría que el mejor resultado se obtiene cuando se asigna más peso a las características identificadas al inicio y al final de la opinión de cada usuario. Sin embargo, 12 configuraciones diferentes para Z_1 , Z_2 y Z_3 obtienen el valor de precisión de 80.05%, con lo cual, el análisis hecho para estos parámetros en el dominio financiero con los métodos anteriores es válido también ahora, es decir, no existe una diferencia clara en las configuraciones de estos parámetros que afecte directamente al análisis de sentimientos debido a la clase de corpus que se ha empleado en el experimento.

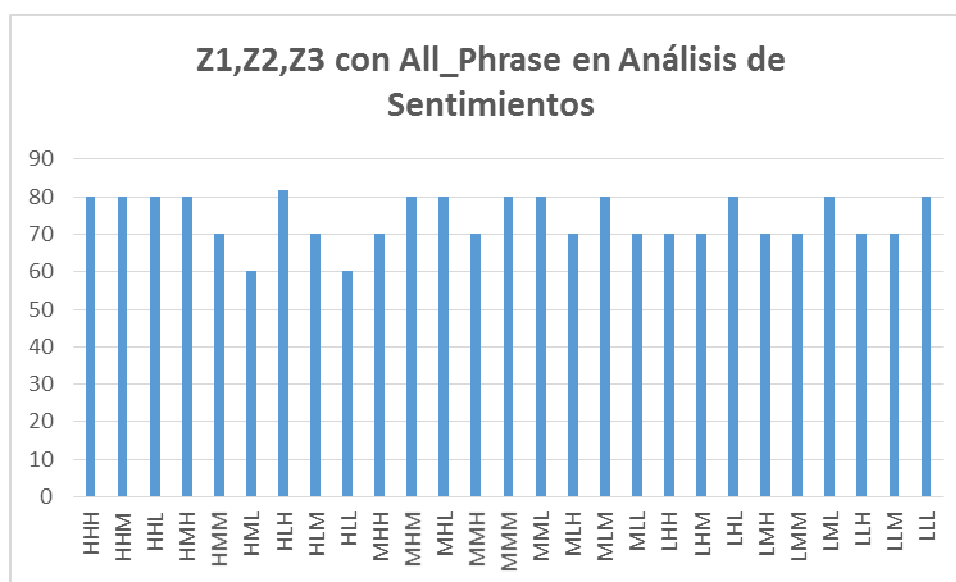


Figura 4.15. Efecto de los parámetros Z_1 , Z_2 y Z_3 en el análisis de sentimientos de opiniones con el método *All Phrase*.

La figura 4.16 presenta gráficamente un resumen de todos los resultados de los experimentos que se han llevado a cabo en el dominio de las finanzas. Esta figura muestra la precisión media obtenida con cada uno de los métodos de clasificación estudiados, esto es, “*N_GRAM Before*”, “*N_GRAM After*”, “*N_GRAM Around*” y “*All Phrase*”. Los resultados reflejan mucha igualdad en los valores promedios obtenidos con los diferentes métodos de clasificación de sentimientos. En términos generales, destacan levemente los métodos “*N_GRAM Around*” y “*All Phrase*” que se mueven en valores de precisión media cercanos al 75% para cualquier configuración de *N_GRAM*.

El mejor resultado promedio para este dominio e idioma, con un valor del 77.15%, se obtiene con el método “*N_GRAM Around*” con el valor de *N_GRAM* igual a 6.

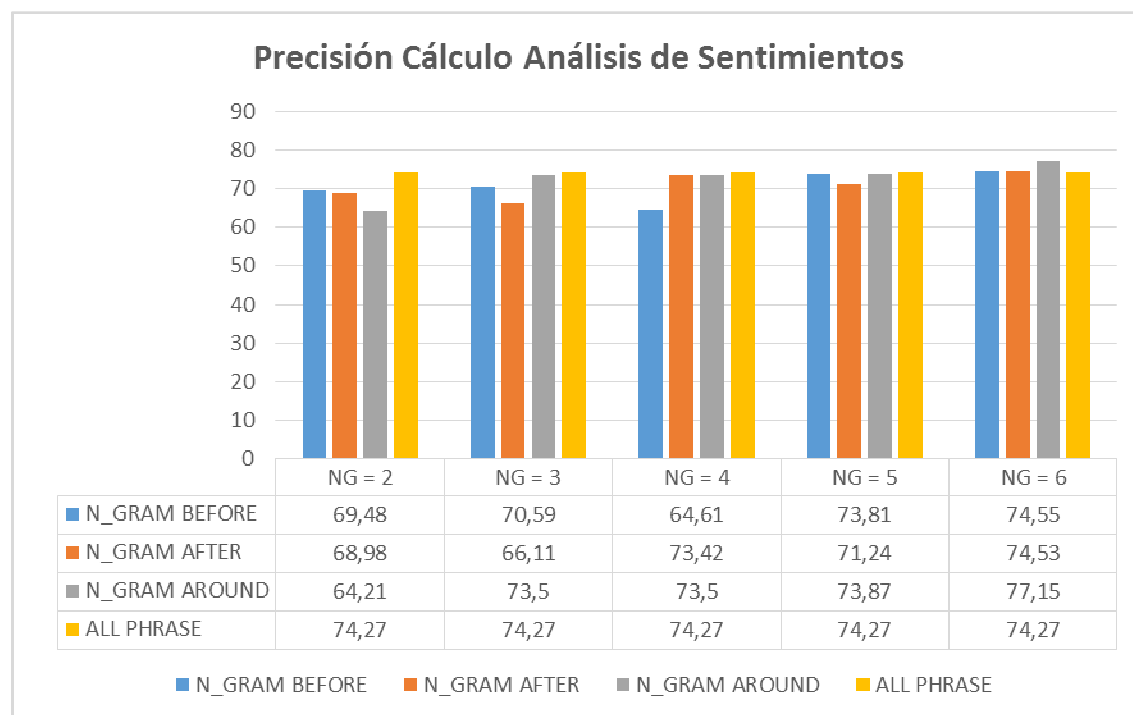


Figura 4.16. Comparativa visual de precisión en el cálculo del análisis de sentimientos de opiniones.

La razón de la igualdad de los resultados obtenidos para la clasificación de sentimientos en el dominio de las finanzas radica en la forma en la que hemos construido el corpus de los experimentos. Como se ha comentado anteriormente, el lenguaje utilizado en el corpus de las opiniones de los usuarios en el dominio financiero es muy diferente al empleado en las opiniones del dominio cinematográfico. Se han utilizado columnas de opinión de periodistas (usuarios especializados en el dominio) para crear el corpus de opiniones en el dominio financiero. Esto implica que las opiniones contengan un lenguaje culto, con frases cortas, carentes de ironía, directo y con pocas ambigüedades. Se trata de un lenguaje preciso a fin de expresar sólo lo que se pretende decir y no dar margen para interpretaciones erróneas. Los profesionales de la información deben evitar, siempre que sea posible, expresiones vagas y utilizar frases cortas. Las frases son construidas de modo directo porque son más eficaces para transmitir una idea. Se emplearán oraciones subordinadas de manera muy esporádica.

En el contexto explicado en el párrafo anterior, el supuesto de dar más importancia a las características identificadas al inicio y al final de la opinión de los usuarios deja de tener tanta relevancia en el análisis de sentimientos de las opiniones de los usuarios, ya que, en este tipo de opinión, las características centrales también son muy importantes a la

hora de valorar la clasificación de sentimientos global de la opinión del usuario. Se trata de opiniones con muchísima cantidad de información y que no divagan en su parte central, sino que se mantienen constantes y rígidas en cualquiera de sus partes. Por tanto, en este dominio los parámetros Z_1 , Z_2 y Z_3 , que dictaminan el peso de cada característica en función de las partes de la opinión en la que aparezca, pasan a tener un papel menos relevante en el proceso de clasificación de sentimientos de opiniones. Esto propicia que los resultados de los métodos de análisis de sentimientos analizados en esta tesis se igualen.

4.5.7 Conclusión

En los experimentos llevados a cabo en el dominio de la finanzas ha quedado probado que la metodología propuesta en esta tesis doctoral funciona bien en el proceso de análisis de sentimientos de opiniones escritas en inglés para este dominio. Se ha comprobado que, a pesar de que en el dominio cinematográfico se obtienen mejores resultados en la identificación y cálculo de la polaridad de características, en el financiero se obtienen mejores resultados promedios para la clasificación de sentimientos de opiniones para los cuatro métodos estudiados (“*N_GRAM Before*”, “*N_GRAM After*”, “*N_GRAM Around*” y “*All Phrase*”).

La razón que explica estos resultados radica en las singularidades de los corpus utilizados en los experimentos de cada dominio. Por un lado, el lenguaje coloquial y desenfadado de los usuarios que opinan sobre películas y, por otro, el lenguaje formal, directo y sin ambigüedad empleado en las opiniones sobre finanzas.

Otra característica del lenguaje financiero es que se emplean muchas cifras concretas y porcentajes numéricos donde se da por entendido el que determinada característica tenga una polaridad positiva o negativa. Por ejemplo, son comunes expresiones del tipo “*And is that the New York Stock Exchange corrected whenever the S&P 500 is about 2,000 points*”. En esta expresión se puede apreciar que hay una cifra numérica que sería interpretada de distinta manera por un economista que por un usuario no experto, ya que implica tener un conocimiento extra del dominio (no presente en el texto de la opinión) para saber asignar la polaridad correcta a la característica *New York Stock Exchange*. Además, es posible que distintos expertos no se pusiesen de acuerdo sobre si esta característica debería tener una polaridad positiva, negativa o neutra, ya que, como se ha comentado, se trata de un dominio que deja mucho margen para la interpretación. Todo

esto dificulta el proceso de cálculo de polaridades de características en el dominio de las finanzas.

Teniendo en cuenta lo anterior, los experimentos llevados a cabo en el dominio de las finanzas concluyen lo siguiente:

- Se trata de un dominio donde las características a identificar son muy dispersas y no están acotadas. Esto incide directamente en los procesos de identificación de características y en el cálculo de la polaridad de las mismas, ya que, en ocasiones, la clasificación de una característica como positiva o negativa es bastante subjetiva.
- Las dificultades de los procesos de identificación de características y cálculo de las polaridades de las mismas no tienen que implicar necesariamente unos resultados malos en el análisis de sentimientos de las opiniones de los usuarios. En este dominio es frecuente que existan ciertas características que se repiten en las opiniones de los usuarios que toman mayor protagonismo y dirigen la polaridad global de la opinión.
- Se han alcanzado los objetivos propuestos inicialmente para este experimento. Se han obtenido unas precisiones máximas del 66.92% de acierto en el cálculo de polaridades de características y un 90.03% de acierto en el proceso de análisis de sentimientos de opiniones. No se descarta el hecho de que es posible mejorar la metodología.

4.6 Conclusión global del proceso de validación

El dominio de las películas y el financiero tienen características muy distintas. Por un lado, el financiero es un dominio general, bastante ambiguo y sin definiciones claras de sus conceptos. Además, en el dominio financiero existen opiniones de los usuarios que no tienen una clasificación de sentimientos clara como positiva o negativa, y da mucho margen a la interpretación por parte lector de esa opinión. En cambio, en el dominio cinematográfico es bastante más sencillo clasificar manualmente las opiniones de los usuarios como positivas o negativas, dejando menos margen a la interpretación del lector. Los conceptos de este dominio están bastante más acotados que en el dominio financiero, ya que cuando un usuario opina sobre una película cinematográfica, generalmente alude en su opinión a la trama de la película, los actores que intervienen en ella y da su opinión general acerca de lo que le ha parecido la película. Esta opinión es inmutable en el tiempo y no está sujeta a factores externos que dependan del

momento de la opinión. Sin embargo, en el dominio financiero las opiniones de los usuarios están generalmente marcadas por condicionantes externos al dominio que hacen que la opinión tenga un componente temporal muy marcado. Por ejemplo, si un usuario está opinando sobre la situación financiera de una “*Compañía*” situada en el país “*España*”, es frecuente que aluda en su opinión a conceptos de la política de ese país, sus gobernantes e, incluso, haga alusiones a la idiosincrasia cultural de las gentes que habitan ese país. Es por esto que, por lo general, el conjunto de conceptos que aparecen en una opinión vertida sobre el dominio financiero suele ser mucho más disperso que el conjunto de conceptos que pueda aparecer en una opinión vertida sobre el dominio de las películas cinematográficas. Este hecho dificulta el proceso de identificación de características en el dominio financiero con respecto al dominio cinematográfico.

En las figuras que se presentan a continuación, se establece una comparación entre los valores obtenidos por uno y otro dominio en el proceso de identificación de características con respecto a las métricas de validación en los dominios cinematográfico y financiero. Nótese que se ha omitido la métrica de precisión debido a las características del sistema, ya que tal y como está implementado siempre obtiene una precisión de uno para todos los casos que queramos comparar.

La figura 4.17 muestra una comparativa de la exhaustividad entre ambos dominios para el proceso de identificación de características. En esta figura se representan en azul los resultados obtenidos para la exhaustividad en dominio cinematográfico y en rojo los resultados obtenidos en el dominio financiero. En ambos dominios el valor máximo de exhaustividad se obtiene para el subconjunto de los corpus de opiniones de usuarios etiquetadas como positivas. El margen con respecto a los otros corpus empleados en el experimento para cada dominio (el de las opiniones negativas y el corpus completo) es tan reducido que no se puede concluir que el proceso de identificación de características propuesto en la metodología se comporte mejor con las opiniones positivas que con las opiniones negativas u opiniones totales. Lo que sí se confirma experimentalmente es que el proceso de identificación de características en el dominio financiero es mucho más complejo de llevar a cabo que en el dominio de las películas cinematográficas. Es por esto que el proceso de identificación de características obtiene mejores resultados de exhaustividad en el dominio cinematográfico que en el dominio financiero.

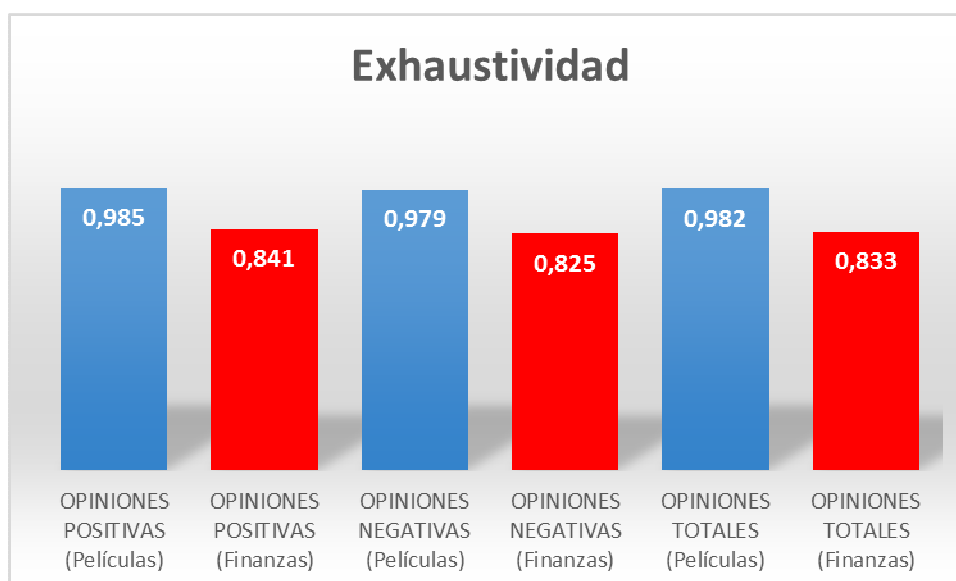


Figura 4.17. Comparativa visual de resultados de exhaustividad obtenidos en el proceso de identificación de características entre los dominios cinematográfico y financiero.

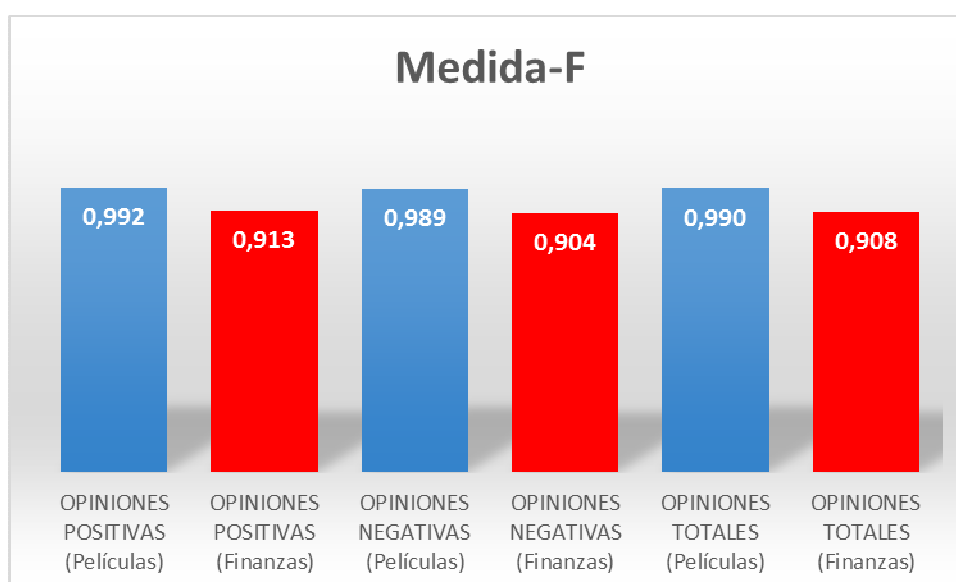


Figura 4.18. Comparativa visual de resultados de medida-F obtenidos en el proceso de identificación de características entre los dominios cinematográfico y financiero.

En la figura 4.18 se muestra la comparativa de los resultados obtenidos con respecto a la métrica de medida-F entre ambos dominios para el proceso de identificación de características. De nuevo, se representan en azul los resultados obtenidos para la medida-F en dominio cinematográfico y en rojo los resultados obtenidos en el dominio financiero. En la figura se aprecia que los resultados relativos a esta métrica están mucho más igualados que para la exhaustividad. Una vez más los resultados obtenidos en el dominio cinematográfico son sensiblemente mejores que los obtenidos para el

financiero. Este hecho vuelve a corroborar la hipótesis que se planteó al principio de este apartado: la identificación de características es más compleja de llevar a cabo en el dominio de las finanzas que en el de las películas atendiendo a las diferencias entre ambos.

En los experimentos realizados para evaluar las fases de cálculo de la polaridad de las características identificadas y de análisis de sentimientos de las opiniones de los usuarios en su conjunto se han empleado cuatro métodos: “*N_GRAM Before*”, “*N_GRAM After*”, “*N_GRAM Around*” y “*All Phrase*”. A continuación se muestran los resultados obtenidos con cada método en el proceso de cálculo de polaridades de las características identificadas por el sistema en los dominios cinematográfico y financiero. Las figuras 4.19, 4.20, 4.21 y 4.22 muestran gráficamente los resultados obtenidos empleando los citados métodos en el proceso de cálculo de polaridades de las características identificadas por el sistema en los dominios cinematográfico y financiero.

En la figura 4.19 se representan los resultados obtenidos para este proceso aplicando el método “*N_GRAM Before*”. Los resultados obtenidos en el dominio de las películas cinematográficas son mejores que los obtenidos en el dominio de las finanzas. Para todas las posibles configuraciones del valor del parámetro *N_GRAM*, el resultado en el dominio cinematográfico es superior al alcanzado en el dominio financiero. Es posible concluir, por tanto, que el método de cálculo de polaridades de características “*N_GRAM Before*” es más eficaz en el dominio cinematográfico que en el dominio financiero.

Las razones se encuentran asociadas a las propias particularidades del dominio financiero. Tal y como se ha comentado anteriormente, se trata de un dominio donde la clasificación de características no es trivial. Hay un componente subjetivo muy acusado. Incluso para los expertos del dominio no es sencillo clasificar una característica de este dominio como positiva, negativa o neutra dentro de una opinión concreta. Para poner de manifiesto este hecho, se propone la siguiente frase extraída de la opinión de un usuario en el dominio financiero: “*In China, the auto sales have improved but given the worst data since March*”. En esta frase, el sistema identificaría la característica “*China*”, ya que es instancia de la clase “*Country*” en la ontología del dominio financiero. Si un experto del dominio quisiese clasificar la polaridad de la característica “*China*”, la subjetividad de la frase complicaría la operación. El experto podría clasificar a la característica “*China*” como positiva, ya que las ventas han aumentado. Sin embargo, otro experto más pesimista podría clasificar esta característica como negativa, ya que en

la misma frase se dice que las ventas tienen los peores números desde Marzo. Ambos expertos tendrían su parte de razón, ya que el dato que se proporciona en la frase es un tanto ambiguo. Este tipo de frases se dan con bastante frecuencia en las opiniones sobre noticias financieras, de ahí que la clasificación de sentimientos en este dominio por parte del sistema automático propuesto en esta tesis doctoral no sea una tarea sencilla. Esto no ocurre de la misma manera para el dominio cinematográfico, donde, por ejemplo, frases de opiniones de usuarios como *“The actors are pretty good for the most part, although Wes Bentley just seemed to be playing the exact same character that he did in American Beauty, only in a new neighborhood”*, dejan de manifiesto de manera clara que la característica *“actors”* (identificada porque es el plural de la clase *“Actor”* en la ontología del dominio) tiene una polaridad marcadamente positiva, mientras que la característica *“Wes Bentley”* (identificada porque es una instancia de la clase *“Actor”* en la ontología del dominio) tiene una polaridad claramente negativa en esta frase. Es por esto que resulta natural que los resultados para la clasificación de características en el dominio de las películas cinematográficas sean mejores que en el dominio de las finanzas.

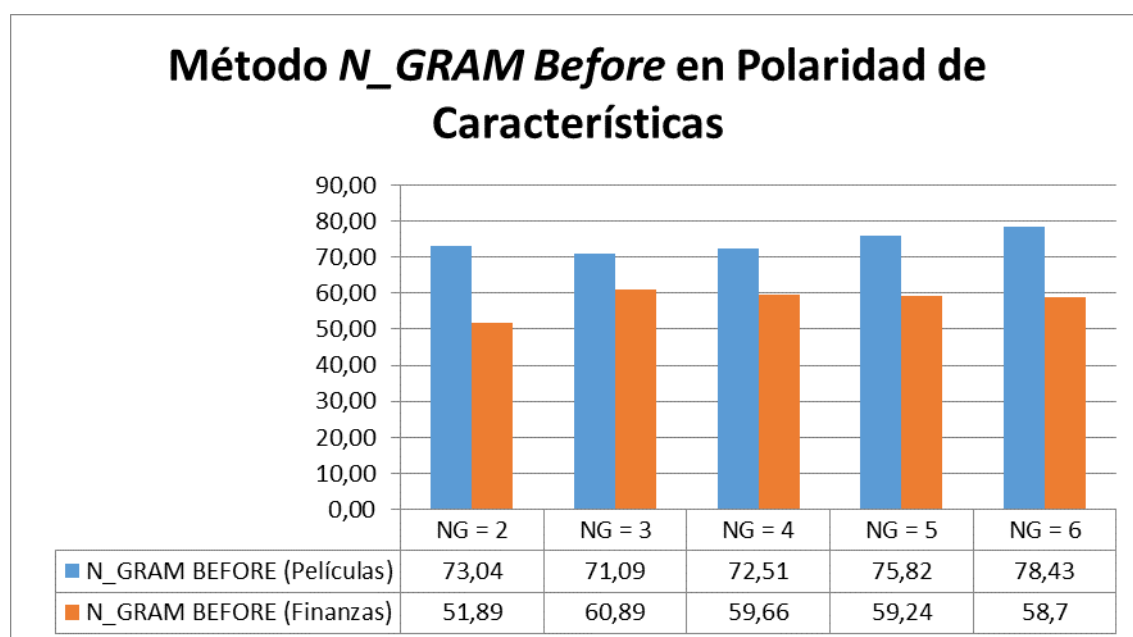


Figura 4.19. Comparativa visual de precisión en el cálculo de polaridad de características con el método *N_GRAM Before* entre distintos dominios.

La figura 4.20 muestra gráficamente los resultados obtenidos para el proceso de cálculo de polaridades de las características identificadas en los dominios cinematográfico y financiero con el método *“N_GRAM After”*. Los resultados que se obtienen con este

método de clasificación de características en el dominio de las películas cinematográficas son, por lo general, significativamente más negativos que los obtenidos con el resto de métodos de clasificación. En el dominio financiero no existe una diferencia tan significativa, pero también se obtienen peores resultados. En la figura 4.20 se puede observar que las diferencias en los resultados alcanzados con este método para un dominio y otro no son muy abultadas. En cualquier caso, se sigue repitiendo la circunstancia de que, independientemente del valor del parámetro N_GRAM , el dominio de las películas cinematográficas resulta en datos de precisión más positivos. Las razones que explican este hecho son de la misma naturaleza que las comentadas para el método “ N_GRAM_Before ”.

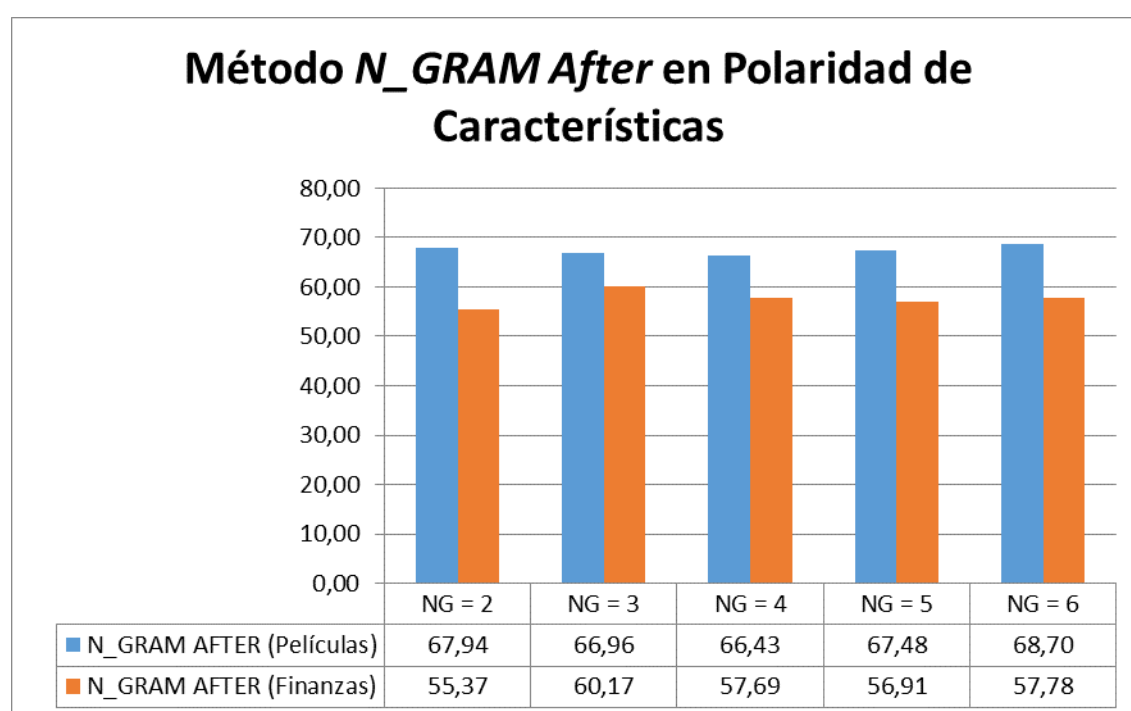


Figura 4.20. Comparativa visual de precisión en el cálculo de polaridad de características con el método N_GRAM After entre distintos dominios.

La figura 4.21 muestra gráficamente los resultados obtenidos para el proceso de cálculo de polaridades de características identificadas en los dominios cinematográfico y financiero con el método “ N_GRAM Around”. Los resultados muestran que este método de clasificación de sentimientos de características obtiene mejores resultados en el dominio de las películas cinematográficas que en el dominio de las finanzas. Una vez analizados los métodos de clasificación anteriores, por analogía, estos resultados son los esperados, ya que las condiciones son las mismas.

La figura 4.22 muestra gráficamente los resultados obtenidos para el proceso de cálculo de polaridades de características identificadas en los dominios cinematográfico y financiero con el método “*All Phrase*”. En este método se ve que, de nuevo, los mejores resultados se obtienen para el dominio cinematográfico, aunque las diferencias son bastante menores que en los métodos anteriores.

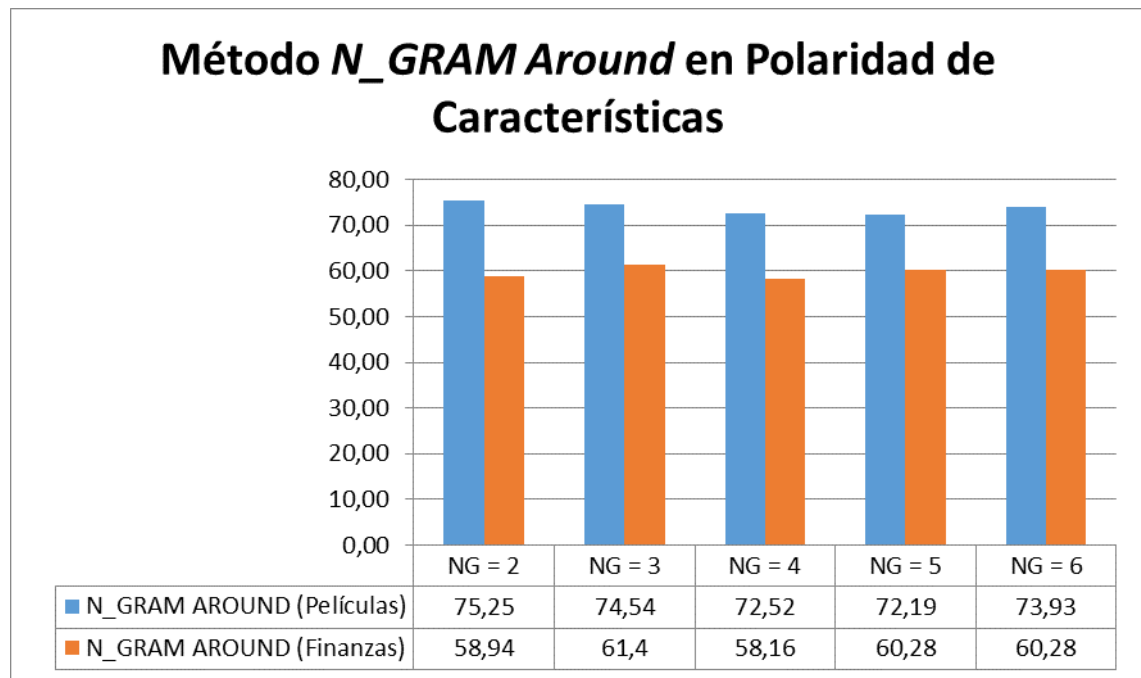


Figura 4.21. Comparativa visual de precisión en el cálculo de polaridad de características con el método *N_GRAM Around* entre distintos dominios.

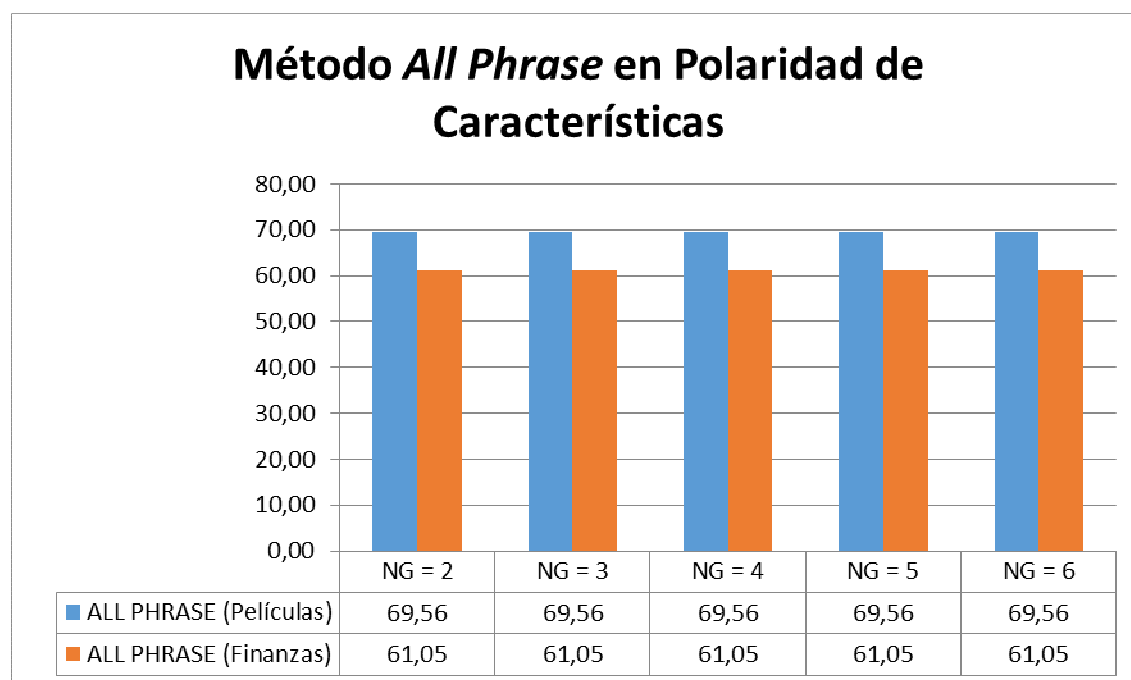


Figura 4.22. Comparativa visual de precisión en el cálculo de polaridad de características con el método *All Phrase* entre distintos dominios.

Para finalizar, el proceso de identificación de características que forma parte de la metodología propuesta en esta tesis doctoral obtiene mejores resultados en el dominio de las películas cinematográficas que en el dominio de las finanzas, independientemente del método de clasificación de sentimientos de características que se emplee. Las razones residen en la naturaleza del dominio financiero. Se trata de un dominio donde la clasificación de características tiene un componente subjetivo muy marcado, dificultando la clasificación de una característica de este dominio como positiva, negativa o neutra dentro de una opinión concreta.

Una vez que se han analizado las diferencias entre los dominios cinematográfico y financiero en el cálculo de la polaridad de características, queda pendiente estudiar el comportamiento de la metodología propuesta con respecto al análisis de sentimientos de opiniones de usuarios expresadas en lenguaje natural para estos dos dominios. En primer lugar, se presentan gráficos comparativos (ver figuras 4.23, 4.24, 4.25 y 4.26), uno para cada método de clasificación de sentimientos, con los resultados obtenidos en cada dominio y se comentan brevemente. A continuación, concluyendo la sección, se analizan en profundidad todas las evidencias obtenidas.

En la figura 4.23 se muestra una comparativa de la precisión obtenida en el proceso de análisis de sentimientos de opiniones de usuarios con el método “*N_GRAM Before*”. Los resultados obtenidos en este experimento establecen que el dominio financiero se

comporta mejor en tres de las cinco configuraciones estudiadas del parámetro N_GRAM . Concretamente, el sistema obtiene mejores resultados para el dominio financiero con las configuraciones de N_GRAM igual a 2, 5 y 6. En cualquier caso, las diferencias entre los resultados para uno y otro dominio no son significativas.

En la figura 4.24 se presenta visualmente la comparativa de la precisión obtenida en el proceso de análisis de sentimientos de opiniones de usuarios con el método “ N_GRAM After”. Los resultados recopilados indican que para este método de clasificación de sentimientos de opiniones, los mejores resultados se alcanzan siempre en el dominio financiero, con independencia del valor asignado al parámetro N_GRAM . Es de resaltar, la abultada diferencia entre los resultados para un dominio con respecto al otro. Para el dominio cinematográfico este método obtuvo resultados poco favorecedores pero, sin embargo, en el dominio financiero los resultados son bastante prometedores.

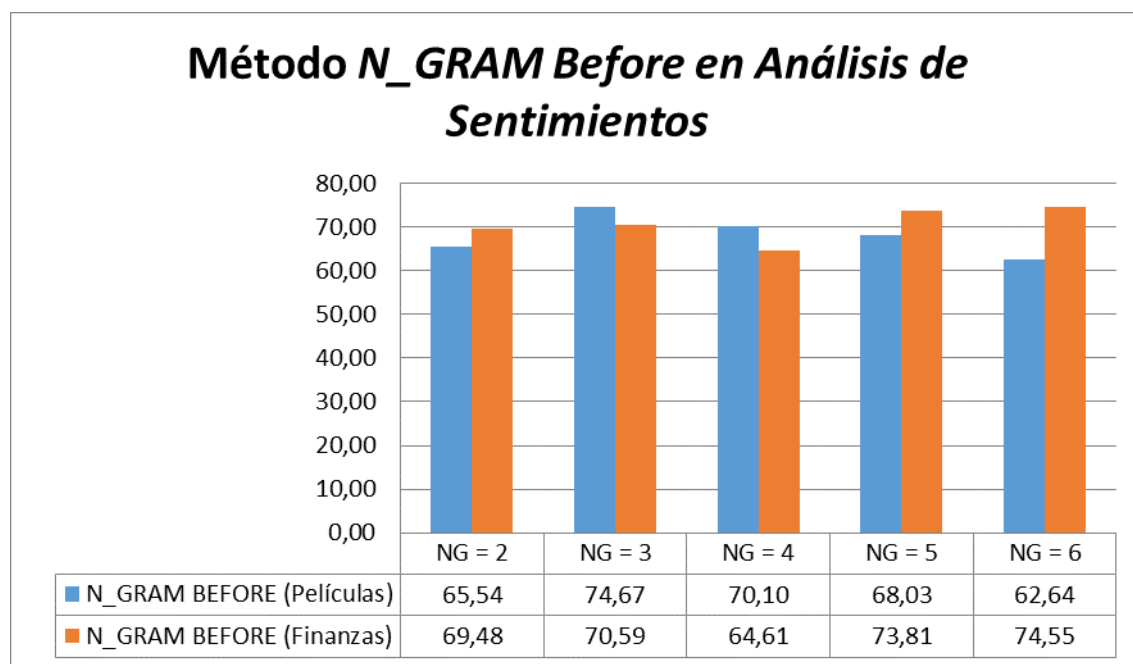


Figura 4.23. Comparativa visual de precisión en el cálculo del análisis de sentimientos de opiniones con el método N_GRAM Before entre distintos dominios.

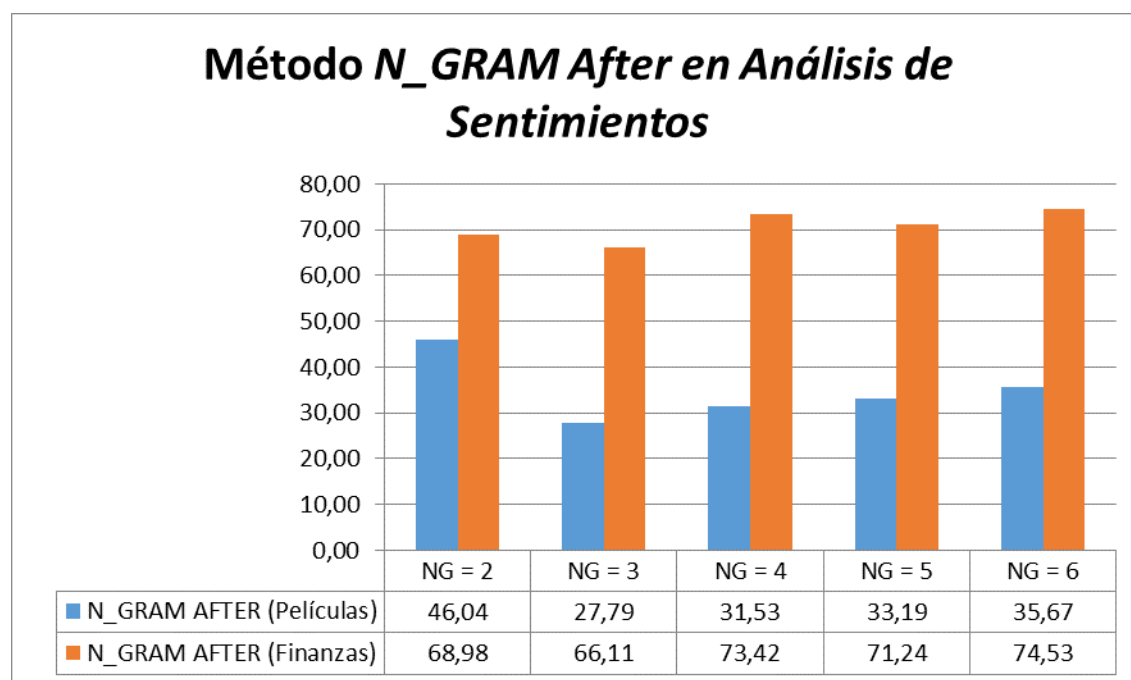


Figura 4.24. Comparativa visual de precisión en el cálculo del análisis de sentimientos de opiniones con el método *N_GRAM After* entre distintos dominios.

En la figura 4.25 se muestra una comparativa sobre la precisión obtenida en el proceso de análisis de sentimientos de opiniones de usuarios con el método “*N_GRAM Around*”. Los resultados obtenidos indican que este método de clasificación de sentimientos de opiniones se comporta mejor en el dominio financiero, con independencia del valor asignado al parámetro *N_GRAM*. En esta comparativa destacan las abultadas diferencias en los resultados entre ambos dominios para los valores 5 y 6 de *N_GRAM*. La razón principal de que se produzca este hecho radica en que en el dominio financiero las frases de las opiniones tienen, por lo general, mayor contenido emocional que en el dominio cinematográfico. Es decir, los usuarios emplean con frecuencia frases mucho más contundentes y cargadas de emoción en este dominio. Esto propicia que existan características en las opiniones cuya polaridad asignada quede muy marcada y, por lo tanto, tengan una mayor incidencia que las demás en el proceso de análisis de sentimientos. Cuando se emplea un mayor número de palabras de contenido emocional para calcular la polaridad de las características existe una mayor probabilidad de que se produzca este hecho dentro del ámbito del dominio financiero, de ahí los resultados obtenidos.

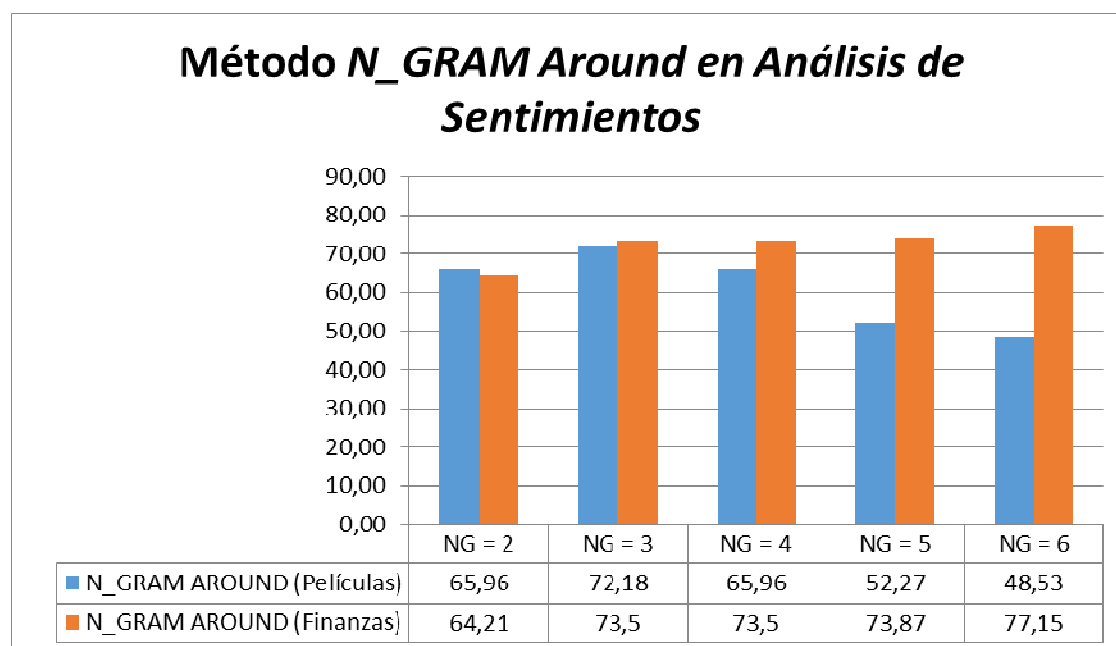


Figura 4.25. Comparativa visual de precisión en el cálculo del análisis de sentimientos de opiniones con el método *N_GRAM Around* entre distintos dominios.

En la figura 4.26 se presentan visualmente los datos de precisión obtenidos en el proceso de análisis de sentimientos de opiniones de usuarios con el método “*All Phrase*”. Nuevamente, atendiendo a los resultados, este método de clasificación de sentimientos de opiniones se comporta mejor en el dominio financiero que en el de películas cinematográficas, para cualquier configuración del parámetro *N_GRAM*.

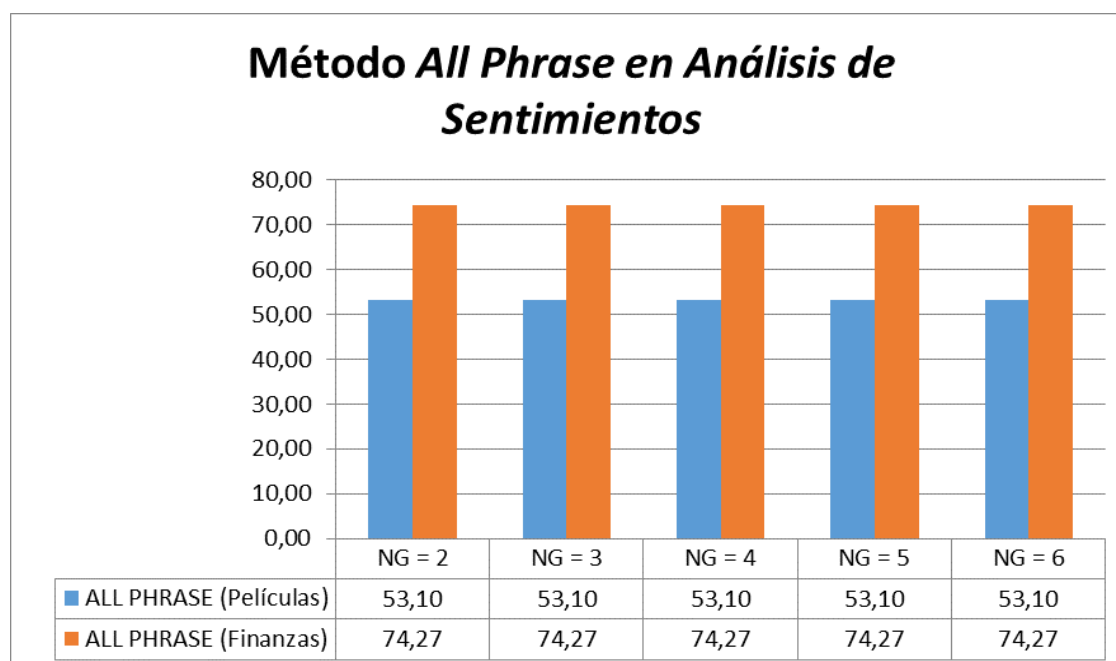


Figura 4.26. Comparativa visual de precisión en el cálculo del análisis de sentimientos de opiniones con el método *All Phrase* entre distintos dominios.

Las claves para comparar los resultados de análisis de sentimientos de opiniones en los dominios cinematográfico y financiero no son inmediatas de obtener. En principio todo hacía prever que, tras obtener para el dominio cinematográfico mejores resultados en el cálculo de la polaridad de características empleando cualquiera de los métodos de clasificación de sentimientos estudiados, el análisis de sentimientos de las opiniones seguiría esta misma tendencia, es decir, los mejores resultados se obtendrían siempre en este dominio. Esto era de esperar puesto que en la metodología de análisis de sentimientos de opiniones propuesta se hace uso de los vectores de polaridad resultantes para cada una de las características identificadas en cada opinión en el proceso de análisis del sentimiento de la opinión global del usuario. Por tanto, la pregunta que es necesario plantearse es la siguiente: si los resultados de la clasificación de características en el dominio de las finanzas son siempre peores que los obtenidos en el dominio cinematográfico, ¿por qué bastantes resultados en el análisis de sentimientos de opiniones son mejores en el dominio de las finanzas que en el dominio cinematográfico? La respuesta proviene del análisis del lenguaje empleado en las opiniones de los usuarios en uno y otro dominio. A continuación se describen las causas que dan lugar a los resultados obtenidos.

El primer motivo para que se produzca esta circunstancia obedece a la naturaleza de los corpus de opiniones que se han empleado en un dominio y otro para llevar a cabo los experimentos. Mientras que en el dominio cinematográfico se ha hecho uso de un corpus existente que había sido utilizado en otros trabajos de investigación, para el dominio financiero ha sido necesario construir un corpus de opiniones ad hoc utilizando opiniones de usuarios publicadas en los sitios Web que aparecen en la tabla 4.13. La diferente naturaleza de ambos corpus reside en que en el corpus de opiniones del dominio cinematográfico las opiniones son suministradas por usuarios no expertos, es decir, personas que ha ido al cine a ver una película y opina sobre ella. Mientras tanto, el corpus sobre opiniones en el dominio financiero que se ha construido para los experimentos contiene opiniones de usuarios semi-expertos dado que se trata, fundamentalmente, de columnas de opinión de periodistas especializados en el dominio financiero. Estos usuarios no se pueden catalogar como “expertos del dominio”, ya que no son economistas ni directores financieros, por ejemplo, pero sí que tienen un alto conocimiento sobre el dominio en el que están opinando. Por tanto, la forma en la que opinan un tipo de usuarios y otros difiere bastante. Por ejemplo, un periodista tiene un determinado espacio en su columna de opinión, mientras que un usuario “normal” puede expresarse libremente sin límite de espacio. La densidad de información en una opinión y otra varía bastante. Más adelante se presenta un ejemplo sencillo que refleja esta situación.

El segundo motivo identificado que tiene impacto en los resultados de análisis de sentimientos obtenidos, es la manera en la que la metodología definida en esta tesis doctoral lleva a cabo el cálculo de polaridades de características y la clasificación de sentimientos de las opiniones de los usuarios. Como ya se ha comentado previamente, en la metodología propuesta se asignan polaridades a las características identificadas en la opinión de los usuarios. Estas polaridades están ponderadas en función del número de ocurrencias de cada característica y la posición relativa de cada una de las ocurrencias. Tras un análisis exhaustivo de un gran número de opiniones en el dominio de las finanzas, se ha observado que, pese a que en general es más complicado el proceso de cálculo de polaridades de características (por los motivos expuestos en el apartado 4.5.5), hay ciertas características que toman una relevancia especial cuando los usuarios opinan. En concreto, de manera frecuente existen características que son ponderadas con mucha importancia (porque se nombran mucho en varias zonas de la opinión), y estas características “dirigen” la polaridad sentimental global de la opinión del usuario. Por tanto, se da la circunstancia de que existen muchas características de poca relevancia cuya polaridad no está clasificada con la misma precisión que en el dominio cinematográfico (de ahí que el dominio financiero obtenga peores resultados en el proceso de clasificación de sentimientos de características), pero, al mismo tiempo, hay otras características con mayor relevancia y a las que se les otorga mayor ponderación (porque se hace mucho énfasis sobre ellas por parte de los usuarios que opinan), que hacen que la polaridad global de la opinión del usuario se clasifique de manera correcta. Es por esto que, los resultados del análisis de sentimientos de opiniones en el dominio financiero, para determinadas configuraciones, son más positivos que en el dominio cinematográfico. Esto se ilustra mediante un ejemplo más adelante.

El tercer, y último, motivo que tiene impacto en los resultados de análisis de sentimientos obtenidos es el tipo de lenguaje empleado por los usuarios en cada dominio. En el dominio cinematográfico el lenguaje empleado por los usuarios es informal, mientras que en el dominio de las finanzas, por la forma en la que se ha construido el corpus, el lenguaje empleado es el periodístico. Por tanto, los resultados obtenidos en la clasificación de sentimientos de opiniones se ven afectados por las diferencias entre los lenguajes informal y periodístico (o formal). A continuación se analizan las diferencias fundamentales entre una forma de expresarse y otra de cara a entender los resultados en el análisis de sentimientos de opiniones en ambos dominios.

Según se explica en la *Corporación de Radio y Televisión Española*⁴⁷, las características principales del lenguaje periodístico son las siguientes:

- **Claridad.** La claridad del mensaje se consigue únicamente empleando un lenguaje que sea comprensible para todos, sin distinción de niveles de formación ni culturales. Debe estar construido mediante estructuras gramaticales simples (“sujeto, verbo, complemento”). El lenguaje periodístico en los medios audiovisuales puede ser imaginativo pero nunca complejo por lo que salvo excepciones, deben evitar las figuras retóricas y las estructuras gramaticales propias de la escritura literaria.
- **Precisión.** El lenguaje en los medios audiovisuales debe ser preciso a fin de expresar sólo lo que se pretende decir y no dar margen para interpretaciones erróneas. Se deben evitar, siempre que sea posible, expresiones vagas en materia de cantidad (varios, algunos, etc.).
- **Brevedad.** Las frases cortas y construidas de modo directo son las más eficaces para transmitir una idea. La construcción ocasional de oraciones subordinadas tiene como fin evitar la monotonía de un lenguaje demasiado lacónico.
- **Verbos de acción.** Los verbos que expresan acción atraen más la atención y, si están bien escogidos, expresan con más precisión lo que se quiere comunicar. Es conveniente evitar los verbos estáticos (ser, estar) si se puede expresar la misma idea con un verbo de movimiento.
- **Informar en positivo.** El mensaje resulta más claro si se construye en positivo. Siempre que sea posible, deben evitarse las frases negativas. Por idéntico motivo, se prefieren verbos en positivo como, por ejemplo, “rechazar” en lugar de “no admitir” o “prohibir” en lugar de “no autorizar”.

El lenguaje formal se emplea en situaciones serias, por ejemplo en una conferencia o cuando se está informando sobre una noticia importante. En el lado opuesto, el lenguaje informal o coloquial no sigue las reglas de la lengua con precisión. Se basa en la espontaneidad y suele caer en las reiteraciones y omisiones. Es frecuente el uso de algunas palabras en jerga en este tipo de lenguaje. Este lenguaje corresponde a situaciones más informales donde se emite una opinión sobre un tema que no es trascendental.

Todo esto deriva en que cuando los usuarios opinan sobre películas cinematográficas emplean el lenguaje de manera coloquial y no se ajustan al rigor de las reglas

⁴⁷ <http://manualdeestilo.rtve.es/el-lenguaje/6-1-caracteristicas-esenciales-del-lenguaje-periodistico/>, última consulta en Enero de 2015

gramaticales del lenguaje. Es frecuente que usen ironías y palabras de jerga callejera que no están presentes en bases de datos de sentimientos. Sin embargo, las opiniones de los usuarios en el dominio financiero (generalmente periodistas del sector) siguen con rigor las reglas gramaticales del lenguaje y las opiniones son muy densas en cuanto a la información que proporcionan. Hay mucha cohesión entre las palabras y dejan un mensaje claro.

De este último punto se deriva el hecho de que se obtengan buenos resultados en el análisis de sentimientos de las opiniones de los usuarios en el dominio financiero, ya que las frases son directas y con bastante contenido emocional, mientras que en el dominio cinematográfico hay más frases vagas que no aportan mucho contenido emocional. Por tanto, pese a que por las características de ambos dominios, es más fácil identificar y clasificar las características en el dominio cinematográfico, el dominio financiero obtiene resultados prometedores en la clasificación de sentimientos de las opiniones de los usuarios.

4.7 Resumen

En este capítulo, se han presentado los resultados de los experimentos que se han llevado a cabo para evaluar y validar la metodología propuesta en esta tesis doctoral. Se ha aplicado la metodología para la minería de opiniones basada en características guiada por ontologías en dos dominios de aplicación bien diferenciados: el cinematográfico y el financiero. Se han escogido dominios tan divergentes para comprobar el comportamiento de esta metodología en situaciones totalmente diferentes. Las métricas empleadas para comprobar la bonanza del sistema han sido las de precisión, exhaustividad y medida-F. Estas métricas son ampliamente utilizadas en la validación de sistemas de procesamiento del lenguaje natural y recuperación de información.

En primer lugar, se ha realizado un detallado análisis de los resultados obtenidos en el dominio cinematográfico. Para llevar a cabo los experimentos en este dominio se ha utilizado un corpus pre-existente y disponible en la Web que ha sido empleado en multitud de trabajos de investigación. Se trata de un corpus con opiniones de usuarios que utilizan un lenguaje informal. Los resultados obtenidos en este dominio han sido satisfactorios, tanto para la clasificación de sentimientos de características identificadas en las opiniones de los usuarios como en el cálculo de la polaridad global de las opiniones.

A continuación, se han comentado los experimentos realizados sobre el dominio financiero. En este caso, debido a la dificultad de encontrar un corpus apropiado disponible públicamente en la Web, ha sido preciso construir desde cero uno propio a emplear en estos experimentos. Se trata de un corpus con opiniones de usuarios semi-expertos del dominio que utilizan un lenguaje periodístico o formal. Los resultados obtenidos para este dominio también han sido bastante prometedores, ya que, a pesar de las características tan particulares de este dominio, el sistema desarrollado realiza el análisis de sentimientos de las opiniones de los usuarios de forma precisa en un alto porcentaje de las ocasiones.

Por último, se ha llevado a cabo una exhaustiva comparativa de los resultados obtenidos en cada dominio. Se han analizado cada una de las variables y parámetros que condicionan el funcionamiento de la metodología propuesta, y se han comentado las posibles causas de las diferencias de resultados obtenidas en un dominio y otro.

Capítulo V

Conclusiones y trabajo futuro

5.1 Introducción

En este capítulo se aportan las conclusiones que se han extraído de las tareas llevadas a cabo en la presente tesis doctoral: trabajo de investigación, especificación de la metodología de minería de opiniones, validación de la misma en dos dominios distintos y desarrollo de la aplicación de software. Además se abren posibles vías de investigación futura.

5.2 Conclusiones

La metodología que se ha creado responde a un innovador sistema de análisis de sentimientos basado en características que utiliza como base de conocimiento una ontología del dominio. Esta metodología propuesta es independiente del dominio y del idioma empleados. A partir de ella se identifican las características relevantes dentro de cada opinión. Durante el proceso de identificación se tiene en cuenta la estructura semántica de la ontología. Cada una de las frases que contengan expresiones lingüísticas que estén directamente relacionadas con términos de la ontología se tienen en cuenta para calcular la polaridad global de la opinión del usuario. La aplicación de esta metodología ha propiciado el desarrollo de un nuevo software para la clasificación de sentimientos de opiniones basado en cálculos vectoriales en el espacio R^3 .

Para la creación y el desarrollo de esta nueva metodología se ha estudiado detalladamente el estado del arte en cada una de la tecnologías que intervienen. Concretamente en tecnologías de la Web Semántica y las ontologías, procesamiento del lenguaje natural, espacios vectoriales y minería de opiniones. Dentro de esta metodología se ha diseñado una arquitectura que de soporte al sistema de minería de opiniones basada en características y guiada por ontologías. Esta arquitectura esta compuesta por los módulos que se han definido en la metodología: módulo de procesamiento de lenguaje natural, de identificación de características, de cálculo de polaridades de características y de análisis de sentimientos de opiniones.

La arquitectura requiere el uso de tres elementos externos: un corpus de opiniones, una ontología del dominio y de un lexicón de sentimientos. Opcionalmente la metodología contempla el uso de una base de datos de sinónimos dentro de la arquitectura.

El siguiente paso fue realizar el diseño y la implementación de una aplicación de software que permitiese validar la metodología en un dominio e idioma específicos. En

este punto se hizo patente la escasez de recursos lingüísticos de calidad en el idioma español, por lo que se decidió validar la metodología con corpus de opiniones en inglés. Se eligió este idioma por la gran variedad de recursos lingüísticos computacionales existentes para él, entre los que destacan el lexicon de sentimientos *SentiWordNet* y la base de datos de sinónimos *WordNet*.

Además se necesitaba construir una ontología del dominio. Tras una investigación inicial, se utilizó una ontología del dominio cinematográfico con información lingüística en inglés. Poblar esta ontología resultó fácil debido a la gran cantidad de datos que existen sobre películas en distintos sitios Web, entre los que destaca *IMDb*.

Encontrar un corpus suficientemente significativo con opiniones de usuarios en inglés sobre películas resultó sencillo, ya que existen algunos trabajos anteriores en minería de opiniones que han empleado este corpus y han publicado sus resultados. Por tanto, supuso una motivación extra emplear el mismo corpus de opiniones que habían empleado otros trabajos de minería de opiniones de cara a comparar los resultados obtenidos con el enfoque presentado en esta tesis respecto a los resultados obtenidos con otras soluciones.

Se ha elegido una implementación cliente/servidor para el desarrollo de la aplicación de software debido a las numerosas ventajas que aportan este tipo de aplicaciones con respecto a las aplicaciones tradicionales de escritorio. Una vez implementada la aplicación, obtenidos los recursos lingüísticos necesarios, construida la ontología del dominio y el corpus de opiniones, se llevaron a cabo los experimentos con el sistema de software aplicado en dominio de la películas.

El siguiente paso fue contrastar los resultados obtenidos en un dominio con unas características totalmente diferentes a las del dominio cinematográfico. Se eligió el dominio de las finanzas. Se desarrolló una ontología financiera sobre la base de ontologías existentes en la Web. Además, en este caso no se disponía de un corpus de opiniones, por tanto, se tuvo que crear uno propio consultando sitios Web de economía y finanzas. Las opiniones realizadas por los usuarios en este dominio tenían un carácter más rígido y formal que las realizadas en el dominio cinematográfico. Los resultados obtenidos sobre este nuevo dominio fueron distintos de los obtenidos para las opiniones sobre películas.

Los experimentos sirvieron, entre otras cosas, para llevar a cabo la validación de los módulos del sistema: de identificación de características, de cálculo de polaridades de características y de análisis de sentimientos de opiniones. Para ello se eligieron las

métricas de precisión, exhaustividad y medida-F, ampliamente utilizadas en el análisis de los sistemas de procesamiento de lenguaje natural y los sistemas de minería de opiniones.

La estrategia de validación diseñada consistió en identificar aquellos parámetros del sistema relevantes desde el punto de vista científico y las posibles configuraciones de los mismos. Debido al número de parámetros que maneja el sistema y el conjunto de posibles valores de los mismos, se analizó detenidamente la estrategia de pruebas debido al problema de explosión combinatoria que supone probar cada posible combinación de valores de entrada en el sistema. Se construyó una batería de pruebas significativa con aquellas configuraciones de valores de parámetros del sistema que aportasen valor al estudio.

El paso final de esta investigación fue analizar sistemáticamente el gran volumen de datos obtenidos como resultado de los experimentos. En este contexto, se llevó a cabo un exhaustivo estudio donde se analizaron las causas de la obtención de resultados diferentes al aplicar la metodología sobre los dominios cinematográfico y financiero.

Se analizó el volumen de datos generado por el sistema en el dominio de las películas cinematográficas con opiniones en inglés. Los resultados obtenidos fueron muy satisfactorios y esperanzadores. A continuación se validó la metodología en un dominio radicalmente distinto al de las películas cinematográficas de cara a extraer conclusiones más significativas sobre el valor del trabajo propuesto en esta tesis doctoral. El dominio elegido debía cumplir con las siguientes exigencias: (i) ser un dominio en el que se manejase un lenguaje muy distinto al de las opiniones sobre películas cinematográficas y (ii) ser un dominio de actualidad, con repercusión social y cuyo estudio aportase valor real.

El dominio elegido ha sido el de las finanzas. Se trata de un dominio de actualidad debido a la gran crisis económica a nivel mundial en la que vivimos y con un lenguaje muy distinto al empleado cuando se opina sobre películas. Tras llevar a cabo los nuevos experimentos se analizaron los resultados obtenidos en el dominio de las finanzas.

Se obtuvieron buenos resultados en el análisis de sentimientos de las opiniones en el dominio financiero. Se trata de un dominio con frases directas y generalmente cargadas de contenido emocional, mientras que en el dominio cinematográfico hay frases más vagas que no aportan mucho contenido de sentimiento. Los resultados mostraron que es más fácil identificar y clasificar correctamente la polaridad de las características en el dominio cinematográfico. El dominio financiero es un dominio general, bastante

ambiguo y sin definiciones claras de sus conceptos. Además, en el dominio financiero existen opiniones de los usuarios que no tienen una clasificación de sentimientos clara como positiva o negativa, y la interpretación depende del lector de esa opinión.

Por tanto, se comprobaron empíricamente las diferencias existentes entre un dominio y otro. Los resultados de los experimentos reflejan las diferencias entre el uso de un lenguaje informal frente al uso de un lenguaje más formal a la hora de emitir opiniones. Y cómo afecta el tipo de lenguaje a los procesos de identificación de características y análisis de sentimientos de las opiniones realizadas por los usuarios. Este contraste de datos experimentales obtenidos en dominios tan dispares proporcionó una información muy valiosa de cara al futuro, ya que abre nuevas líneas de investigación sobre las que trabajar.

Por todo lo expuesto anteriormente, se puede concluir que los objetivos que se establecieron inicialmente para el desarrollo de la metodología propuesta en esta tesis doctoral se han alcanzado con éxito.

En la figura 5.1 se muestra un esquema general detallado de los procesos que componen la metodología propuesta en esta tesis doctoral. Los procesos identificados se pueden englobar en las siguientes categorías: (i) procesamiento del lenguaje natural, (ii) identificación de características, (iii) cálculo de la polaridad de características, (iv) análisis de sentimientos de opiniones y (v) validación de resultados.

A continuación se exponen las conclusiones extraídas para cada una de las categorías de procesos identificadas a lo largo de este trabajo. Para ello, se toma como referencia la figura 5.1 que se comenta paso a paso.

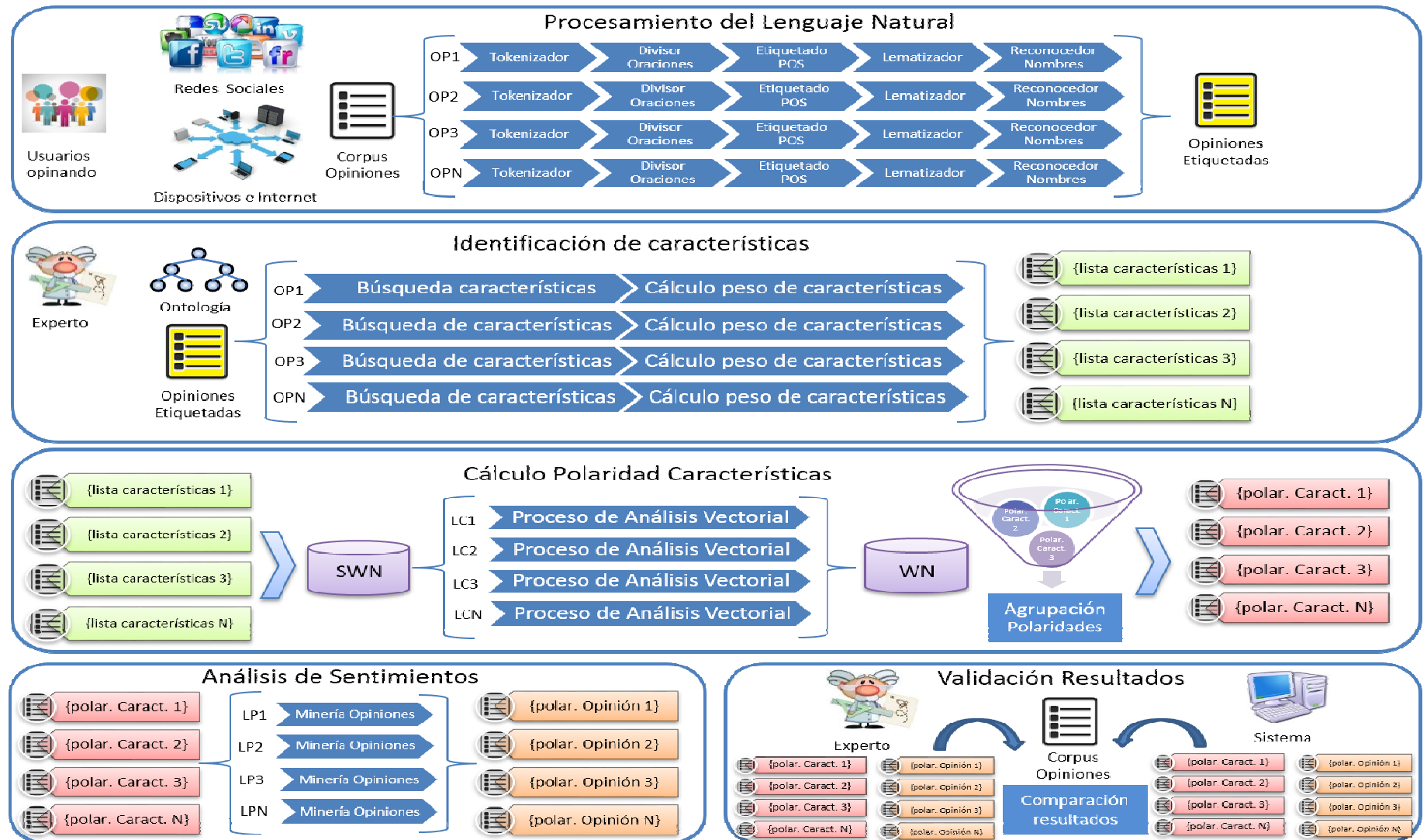


Figura 5.1. Esquema general detallado de los procesos que componen la metodología propuesta.

5.2.1 Procesamiento del lenguaje natural

La metodología propuesta en esta tesis doctoral se compone de un módulo de procesamiento del lenguaje natural. Los usuarios escriben sus opiniones en la Web, generalmente en redes sociales, foros o blogs. Los medios que utilizan para escribir estas opiniones son cada día más variados (smartphones, portátiles, tablets, ordenadores de sobremesa, etc.). El conjunto de todas estas opiniones es lo que compone el corpus de opiniones de los usuarios. Este corpus es la entrada para el conjunto de procesos que tienen lugar dentro del módulo de procesamiento de lenguaje natural. Cuando este módulo recibe el corpus de opiniones de los usuarios, automáticamente empieza a procesarlo y realiza un análisis morfológico de cada opinión. El resultado de este proceso es la obtención de un conjunto de opiniones etiquetadas. Todo este proceso se puede ver esquemáticamente en la figura 5.1.

En el sistema implementado en esta tesis doctoral se ha utilizado la herramienta de procesamiento de lenguaje natural *Stanford Log-linear Part-Of-Speech Tagger*. Se ha comprobado experimentalmente que se trata de una herramienta óptima para usar en proyectos que hagan uso intensivo del procesamiento de textos escritos en lenguaje natural. Son destacables los aspectos de rendimiento y facilidad de uso de esta herramienta en proyectos software basados en la arquitectura J2EE.

Una conclusión extraída con respecto al proceso de *PLN*, es que el tipo de lenguaje informal que emplean los usuarios en sus opiniones dificulta mucho el empleo de técnicas de *PLN* efectivas. Las herramientas que tenemos en la actualidad no contemplan irregularidades lingüísticas en la mayoría de los casos. Teniendo en cuenta este contexto, la investigación llevada a cabo en este trabajo ha estado motivada principalmente por la carencia de recursos, métodos y herramientas para un efectivo procesamiento de la información subjetiva.

5.2.2 Identificación de características

Como muestra la figura 5.1, una vez finalizado el análisis morfológico de las opiniones se lleva a cabo la de identificación de características para cada una de esas opiniones. El módulo utiliza una ontología del dominio construida con rigor y exactitud por parte de un experto. Por tanto, este módulo recibirá como entrada el conjunto de opiniones etiquetadas morfológicamente junto a la ontología del dominio para identificar las características relevantes dentro de las opiniones de los usuarios. Además, les asignará

un peso en función de su número de ocurrencias y de la posición relativa de cada ocurrencia. La salida de este módulo es un conjunto de características identificadas y ponderadas para cada opinión de usuario.

Una conclusión con respecto al proceso de identificación de características es que la ontología del dominio representa un papel fundamental dentro de la metodología. De su corrección, exactitud, completitud y potencialidad semántica dependen en gran medida los buenos resultados que aporte la solución propuesta en esta tesis. Para realizar los experimentos se ha contado con la ayuda de expertos del dominio que han llevado a cabo un trabajo exhaustivo a la hora de construir, modificar y poblar las ontologías del dominio utilizadas. Por tanto, los resultados del proceso de identificación de características han sido prometedores.

Otra de las conclusiones extraídas es la mayor dificultad de la identificación de características en el dominio financiero con respecto al cinematográfico. El conjunto de conceptos que aparecen en las opiniones del dominio financiero suele ser mucho más disperso y menos acotado que el que aparece en una opinión sobre películas. Este hecho dificulta el proceso de identificación de características en el dominio financiero, ya que se ha comprobado que es más complejo construir una ontología que funcione bien en este dominio que en el cinematográfico.

Cabe destacar la importancia que tiene dentro de la metodología propuesta en esta tesis el hecho de ponderar (asignar pesos) correctamente las características identificadas por el sistema. Este aspecto es clave, ya que esta ponderación tiene influencia directa en el proceso de análisis de sentimientos de las opiniones de los usuarios. Los resultados de los experimentos concluyen que la decisión de ponderar las características identificadas por el sistema en función de su número de ocurrencias y la posición relativa de cada una de ellas dentro las opiniones de los usuarios, produce resultados prometedores en la clasificación de sentimientos.

5.2.3 Análisis de sentimientos de características

El cálculo de la polaridad de las características identificadas se lleva a cabo dentro del módulo de cálculo de polaridades de características, tal y como aparece en la figura 5.1. Este módulo utiliza un proceso de análisis vectorial junto a la base de datos de sentimientos *SentiWordNet* para conseguir su objetivo. Una vez que calcula las polaridades de las características dentro de las opiniones, el módulo se ayuda de la base

de datos de sinónimos *WordNet* para agrupar aquellas polaridades de características equivalentes. A la salida de este módulo se obtiene una lista de polaridades de características para cada opinión de usuario.

La primera conclusión que se extrae de este apartado, fruto de las pruebas realizadas con el sistema y el análisis manual de textos de opiniones, es que el sistema propuesto funciona mejor cuando hay muchas palabras con alto contenido emocional cerca de la expresión lingüística que identifica a una característica dentro de la opinión del usuario. Ya que el sistema tiene mucha información para poder calcular correctamente la polaridad de la característica.

Otra conclusión es el hecho de que dependiendo del dominio en el que se trabaje se tendrá un grado de subjetividad más o menos alto a la hora de clasificar las características identificadas en las opiniones. En dominios donde la ambigüedad y el margen de interpretación en función del contexto son bajos, como por ejemplo el dominio de las películas, los resultados para la clasificación de sentimientos de características serán buenos. En otros donde las opiniones tengan mucho margen de interpretación, exista un componente de ambigüedad alto y las opiniones tengan un significado muy ligado al momento temporal en el que se emiten, los resultados para la clasificación de sentimientos de características serán peores. Este es el caso del dominio financiero.

Una película es algo inmutable en el tiempo, un usuario opina sobre ella y no hay factores externos (políticos o temporales) que influyan en esta opinión. Sin embargo, por lo general, la opinión de un usuario en el dominio financiero está marcada por los tintes políticos del momento (el gobierno actual, situaciones bélicas, casos de corrupción, etc.) y por situaciones externas a la opinión del usuario (por ejemplo, la crisis financiera mundial en el momento de emitir la opinión). Todo esto hace que sea más complicado clasificar algunas de las características identificadas por el sistema.

5.2.4 Análisis de sentimientos de opiniones

El proceso de clasificación de sentimientos de opiniones recibe una lista de polaridades de características por cada opinión de usuario, tal y como se aprecia en el esquema de la figura 5.1. En base a esta lista de polaridades, y utilizando cálculos con vectores en el espacio R^3 , este proceso realiza la clasificación de sentimientos de cada una de las opiniones de los usuarios.

La primera conclusión que se extrae está relacionada con el apartado de identificación de características: es fundamental que se ponderen con mayor peso aquellas características con mayor número de ocurrencias que aparecen en las partes más importantes de la opinión del usuario. Esto propicia que estas características tengan más peso que el resto en el proceso de clasificación de sentimientos de opiniones y que ‘dirijan’ la polaridad global de la opinión del usuario.

Es necesario que el sistema clasifique de manera correcta las características con mayor ponderación, ya que tiene efectos muy negativos en la clasificación de sentimientos una característica con mucho peso cuya polaridad ha sido mal calculada. Por lo tanto, se necesita que se clasifiquen bien aquellas características más importantes y que se les asigne una ponderación muy alta para realizar un análisis de sentimientos óptimo.

5.2.5 Validación de resultados

El proceso de validación de resultados requiere de la intervención de un experto para llevarse a cabo. El experto selecciona un subconjunto representativo del corpus de opiniones de los usuarios y realiza un análisis manual del mismo. Su labor consiste en identificar primero las características relevantes del dominio dentro de cada opinión y asignarles una polaridad (positiva, negativa o neutra). Además, también asigna una polaridad global a la opinión del usuario. Tras esto el sistema compara los resultados obtenidos por él mismo de manera automática con los resultados anotados por el experto. Este proceso se lleva a cabo para la identificación de características, el cálculo de polaridades de características y la clasificación de sentimientos de opiniones.

La conclusión que se obtiene de este apartado es que la validación una metodología de clasificación de sentimientos en un dominio concreto es un proceso laborioso, ya que hay que realizar un conjunto de tareas amplio. Estas tareas son:

- Preparación de un corpus de opiniones representativo dentro del dominio en el que se vaya a trabajar.
- Búsqueda de un experto del dominio en cuestión.
- Elaboración de una ontología del dominio por parte del experto, o búsqueda y modificación, si fuese necesario, de una existente.
- Instanciación correcta de la ontología del dominio que se vaya a usar.
- Elaboración manual por parte del experto de los artefactos pertinentes donde se recoge la clasificación de sentimientos de las características identificadas en

cada opinión de usuario y la clasificación de sentimientos global de cada opinión.

- Programación del software de validación que compara automáticamente los resultados de salida del sistema con los resultados aportados por el experto.
- Interpretación de los resultados de validación obtenidos y aporte de las conclusiones pertinentes.

Todas estas tareas han sido llevadas a cabo en el proceso de validación de resultados del presente trabajo de tesis doctoral dentro de los dominios financiero y cinematográfico.

5.3 Líneas futuras

A lo largo de este trabajo se citan varios temas que no han sido considerados como parte del desarrollo de esta tesis pero que proporcionan nuevas líneas de investigación futuras. En los siguientes apartados se exponen estas líneas con respecto a los asuntos tratados en esta investigación: procesamiento del lenguaje natural, identificación de características, análisis de sentimientos de características y análisis de sentimientos de opiniones. Por último, se propone una posible solución para el análisis de sentimientos que combina la lógica difusa con la minería de opiniones basada en características.

5.3.1 Procesamiento del lenguaje natural

El *PLN* es una parte esencial de la inteligencia artificial que investiga y formula mecanismos computacionalmente efectivos que facilitan la interrelación hombre-máquina. Estos mecanismos permiten una comunicación mucho más fluida y menos rígida que los lenguajes formales y sistemas de menús utilizados tradicionalmente.

Sin embargo existen algunas limitaciones en los los sistemas de *PLN* actuales que permiten abrir nuevas líneas de investigación en este campo. En la metodología propuesta en esta tesis doctoral se han detectado tres problemas principales:

- **Problemas de ambigüedad.** La ambigüedad del lenguaje humano se produce cuando se asocian varios sentidos o significados a una palabra (Ureña, 2001). El contexto tiene mucha importancia en la eliminación de las posibles ambigüedades, es decir, la misma palabra puede ser ambigua en un contexto y no serla en otro. Por ejemplo, si un sistema de *PLN* se encuentra una frase del

tipo: “*estuve toda la mañana esperándote en el banco*”, el sistema no tendrá mecanismos suficientes para saber si se trata de un banco para sentarse o de una institución financiera. Para resolver este conflicto, el sistema tendría que conocer la información del contexto, y esta puede estar presente o no. En concreto, en los experimentos llevados a cabo con el sistema de software implementado en esta tesis doctoral nos hemos encontrado en ocasiones con frases de contenido emocional como la siguiente: “*this film is tremendous!*”, y el sistema no ha sabido diferenciar si se trataba de una película tremendamente buena o tremendamente mala. Para abordar este problema existen diversas soluciones. En el trabajo de Ureña (2001) se estudian una serie de métodos de resolución de la ambigüedad léxica⁴⁸ para sistemas basados en PLN, para ello emplea recursos lingüísticos específicos (diccionarios y tesauros). Más recientemente, en el trabajo de (Fakhrahmad et al., 2014) se presenta un sistema experto basado en minería de datos para eliminar la ambigüedad léxica de las palabras. Este sistema experto utiliza un algoritmo para deducir el sentido correcto de las palabras basado en *forward chaining*, que es un método de razonamiento que parte de un conjunto de datos base y emplea reglas de inferencia para extraer más datos y alcanzar su objetivo.

- **Problemas de anáfora.** Este problema se produce cuando se referencia a un término o a una parte anterior del discurso u opinión. Los pronombres, los adverbios y los verbos son frecuentemente anáforas que evitan la repetición de las palabras ya dichas en un discurso. En concreto, en los experimentos llevados a cabo nos hemos encontrado frases del tipo: “*In this film Brad Pitt and Angelina Jolie are discrete. He makes a vulgar paper and she is misplaced*”. En este caso el sistema debería asociar el adjetivo *vulgar* al papel desempeñado por Brad Pitt en esa película y el adjetivo *misplaced* al papel desempeñado por Angelina Jolie. Para abordar este problema existen diversas soluciones como las presentadas en (Refoufi, 2014; Toledo-Gómez et al., 2014; Varela, 2014; Wang, 2014). Por ejemplo, en (Toledo-Gómez et al., 2014) se presenta una herramienta para llevar a cabo los procesos de identificación y resolución de la anáfora directa en textos escritos en español. Para ello se centra en el trabajo con los pronombres de este idioma y se trata de localizar el referente o antecedente en las frases anteriores (que es la palabra que el pronombre reemplaza).
- **Problemas de elipsis:** Este problema se produce cuando se presenta la supresión de una o más palabras de una frase que, desde un punto de vista gramatical, deberían estar presentes, pero sin las cuales se comprende perfectamente el

⁴⁸ En inglés, *Word Sense Disambiguation* – WSD.

sentido de la frase. Por ejemplo, en la frase “*Ana se va a Berlín, y yo, a Venecia*” hay una elipsis, ya que se han suprimido la palabras “*me voy*”. Existen diversas soluciones para abordar este problema, como las presentadas en (Malková, 2014; Miller & Hemforth, 2013; Giuliani et al., 2014). En el trabajo de Giuliani (2014) se presentan dos métodos para transformar el lenguaje escrito en oraciones gramaticalmente correctas. El primer método presenta un algoritmo para identificar las elipsis en los textos y sustituirlas por palabras apropiadas gramaticalmente, el segundo método presenta un algoritmo basado en similitud semántica de palabras utilizando la base de datos *WordNet* (Miller, 1995).

De lo anterior se concluye que existen líneas de investigación en el campo del *PLN* que pasan por el estudio y la experimentación de soluciones que aporten los formalismos lógicos necesarios para abordar los problemas expuestos en este apartado.

5.3.2 Identificación de características

El proceso de identificación de características descrito en este trabajo alcanza resultados muy prometedores debido a la potencia semántica que proporciona el uso de ontologías. Como se ha comprobado en los resultados empíricos obtenidos en los dominios financiero y cinematográfico, al usar la ontología como medio fundamental para la identificación de características, cada característica identificada en la opinión de un usuario es relevante. Ya que se trata de una clase, instancia, relación o atributo de la ontología del dominio. Si la ontología construida es completa y exhaustiva los resultados obtenidos son óptimos.

La línea de investigación propuesta dentro de este apartado pasa por mejorar la ponderación de cada una de las características identificadas por el sistema. Ahora mismo se calcula la ponderación asociada a cada característica según la fórmula expuesta en el apartado 2.5.3:

$$peso(f, opinion_i) = Z_1 * |O_1| + Z_2 * |O_2| + Z_3 * |O_3|$$

Se asigna mayor peso a aquellas características que se nombran más y aparecen en las partes ‘importantes’ de la opinión de un usuario. Este hecho repercute positivamente en la clasificación de sentimientos de las opiniones de los usuarios.

Sin embargo, en la metodología propuesta en esta tesis doctoral se trata de igual manera cualquier característica identificada en la opinión de un usuario. Se mide su número de

ocurrencias y dónde se ha producido cada una de ellas. Tras un análisis detallado de las opiniones de los usuarios en los dominios financiero y cinematográfico, se concluye que una posible línea de investigación podría ser establecer distintos niveles de características o categorías según el dominio en el que se esté trabajando.

Se puede idear un método de selección de características de primera, segunda, tercera categoría, etc. según el dominio en el que se vaya a trabajar. Posteriormente, las características con mayor categoría tendrán mayor peso en el análisis de sentimientos de las opiniones de los usuarios. Este aspecto complementa el número de ocurrencias de una características y la posición de las mismas.

Una posible solución para resolver este problema puede ser la identificación como características de primera categoría de aquellas asociadas a las clases de la ontología que tienen mayor número de relaciones con otras clases. Las características de segunda categoría podrían ser aquellas asociadas a clases con menor número de relaciones. Las características de tercera categoría podrían ser los atributos e instancias de las clases. Una cuarta categoría puede contener aquellas características derivadas de relaciones en la ontología del dominio.

Para comprobar la efectividad de este método se debería experimentar con los resultados que se obtendrían para distintos dominios probando con distintos niveles de categorías de características. También se deberían estudiar los resultados obtenidos al asignar distintos pesos a cada categoría de características. Trabajando sobre la fórmula anterior de cálculo del *peso* de las características identificadas y aplicando todo lo comentado, se obtendría una expresión como la que aparece en la fórmula 5.1.

$$\text{peso}(f, \text{opinión}_i) = \begin{cases} \text{Peso}_A * (Z_1 * |O_1| + Z_2 * |O_2| + Z_3 * |O_3|) & \text{si tieneCategoría}(f, A) = \text{verdadero} \\ \text{Peso}_B * (Z_1 * |O_1| + Z_2 * |O_2| + Z_3 * |O_3|) & \text{si tieneCategoría}(f, B) = \text{verdadero} \\ \text{Peso}_C * (Z_1 * |O_1| + Z_2 * |O_2| + Z_3 * |O_3|) & \text{si tieneCategoría}(f, C) = \text{verdadero} \\ \text{Peso}_D * (Z_1 * |O_1| + Z_2 * |O_2| + Z_3 * |O_3|) & \text{si tieneCategoría}(f, D) = \text{verdadero} \end{cases}$$

(5.1)

La fórmula 5.1 calcula el peso asociado a la característica f dentro de la opinión de usuario opinión_i . En este caso, la fórmula distingue entre cuatro niveles de características o categorías llamadas A , B , C y D . El número de categorías será un parámetro configurable del sistema. La función $\text{tieneCategoría}(f, \text{categoría})$ devolverá *verdadero* en el caso de que la característica f pertenezca a la *categoría* dada como parámetro. Por tanto, una vez que se sepa la categoría en la que está enmarcada la característica f , en función del conocimiento aportado por la ontología del dominio, se le

asignará uno de los cuatro pesos: $Peso_A$, $Peso_B$, $Peso_C$ o $Peso_D$, que serán parámetros configurables del sistema y representan el peso para las características de primera, segunda, tercera y cuarta categoría respectivamente.

Otra mejora sería que se pudiesen configurar los parámetros del sistema Z_i . En la metodología propuesta en este trabajo se dividen las opiniones en tres partes iguales (el inicio, la parte media y el final de la opinión). La idea es que el número de partes en las que se divida el texto sea configurable y contrastar los resultados obtenidos en función de este aspecto. La fórmula 5.2 muestra cómo quedaría la fórmula 5.1 si se añade la posibilidad de dividir las opiniones de los usuarios en N partes iguales.

$$peso(f, opinion_i) = \begin{cases} Peso_A * (\sum_{i=1}^N Z_i * |O_i|) & \text{si tieneCategoria}(f, A) = \text{verdadero} \\ Peso_B * (\sum_{i=1}^N Z_i * |O_i|) & \text{si tieneCategoria}(f, B) = \text{verdadero} \\ Peso_C * (\sum_{i=1}^N Z_i * |O_i|) & \text{si tieneCategoria}(f, C) = \text{verdadero} \\ Peso_D * (\sum_{i=1}^N Z_i * |O_i|) & \text{si tieneCategoria}(f, D) = \text{verdadero} \end{cases} \quad (5.2)$$

5.3.3 Análisis de sentimientos de características

La polaridad de una característica viene definida en función de su vector de polaridad tal y como se muestra en la siguiente fórmula expuesta en el apartado 2.6.1:

$$\vec{V}(f, opinion_x) = peso(f, opinion_x) * (SentiPos(f, opinion_x), SentiNeg(f, opinion_x), SentiNeu(f, opinion_x))$$

Para calcular la polaridad de una característica se utilizan los valores de sentimiento *SentiPos*, *SentiNeg* y *SentiNeu* de las palabras cercanas a ella en el texto de la opinión del usuario. Las palabras elegidas dependerán del método que se esté empleando para el cálculo de la polaridad de características. La salida del proceso de clasificación de sentimientos de características será un conjunto de características con sus vectores de polaridad asociados.

Una vez que se obtiene la lista de características identificadas en la opinión con sus vectores de polaridad asociados, se podrían agrupar las características por conceptos ontológicos. Estos conceptos serían clases o instancias de la ontología del dominio. Es decir, se podría calcular la proximidad entre cada característica identificada por el

sistema y los conceptos ontológicos presentes en el texto de la opinión del usuario. Esto se podría realizar mediante medidas de distancia semántica y similitud como las propuestas en los trabajos (Hernández-González et al., 2014; Rodríguez-García et al., 2014a; Rodríguez-García et al., 2014b).

Para ilustrar lo anterior se desarrolla un ejemplo. En la tabla 5.1 aparece la opinión de un usuario en el dominio cinematográfico. Una vez se aplican los procesos de identificación de características y cálculo de sus polaridades, se extrae una lista de la forma que se muestra en la tabla 5.2.

Tabla 5.1. Extracto de la opinión de un usuario en el dominio de las películas cinematográficas

“En esta película Harrison Ford está muy mal. La BSO es muy buena pero la duración de la película es muy larga. Con mi iPhone se ve bien, pero la calidad de la chapa es horrible y el sonido no está mal”.

Tabla 5.2. Lista de características identificadas con sus vectores de polaridad asociados

Característica: **película**, polaridad: $\vec{V}_1(\text{película}, \text{userop}_x) = (2, 2, 18)$
 Característica: **Harrison Ford**, polaridad: $\vec{V}_2(\text{Harrison Ford}, \text{userop}_x) = (2, 45, 3)$
 Característica: **BSO**, polaridad: $\vec{V}_3(\text{BSO}, \text{userop}_x) = (20, 11, 5)$
 Característica: **duración**, polaridad: $\vec{V}_4(\text{duración}, \text{userop}_x) = (11, 25, 3)$
 Característica: **iPhone**, polaridad: $\vec{V}_5(\text{iPhone}, \text{userop}_x) = (15, 9, 3)$
 Característica: **chapa**, polaridad: $\vec{V}_6(\text{chapa}, \text{userop}_x) = (4, 12, 3)$
 Característica: **sonido**, polaridad: $\vec{V}_7(\text{sonido}, \text{userop}_x) = (14, 9, 2)$

Las características *película* y *BSO* son identificadas por el sistema porque son clases de la ontología del dominio. Las características *Harrison Ford* y *iPhone* son identificadas porque son instancias en la ontología del dominio, en concreto de las clases *Película* y *Teléfono*. Las características *duración*, *chapa* y *sonido* son identificadas por el sistema porque son atributos de las clases *Película* y *Teléfono* respectivamente.

Con la lista de características y sus vectores de polaridad asociados se puede agrupar estas características en torno a un concepto ontológico. Es decir, se puede calcular el vector de polaridad asociado a una característica teniendo en cuenta las polaridades de las características que están relacionadas con ella dentro de la opinión del usuario. En vez de realizar un cálculo individual del vector de polaridad, se realiza un cálculo teniendo en cuenta las relaciones entre las características identificadas por el sistema. Para llevar

a cabo este cometido se necesita de la información semántica de la ontología del dominio.

Una opción sería el cálculo del vector de polaridad asociado a una característica como la suma de los vectores de polaridad siguientes: vector de polaridad de la propia característica, vectores de polaridad de cada una de las características relacionadas en la ontología del dominio con la característica y vectores de polaridad de las características que son atributos de clases relacionadas con la característica. Además, se trataría de una suma ponderada donde se podría asignar distintos pesos a cada componente vectorial.

Haciendo uso de la agrupación en torno a conceptos ontológicos (instancias o clases de la ontología) las polaridades finales de las características identificadas se muestran en la tabla 5.3. En esta tabla, las variables A , B y C representan el peso que tendrán en el vector de polaridad resultante de una característica los siguientes conceptos: el propio vector de polaridad, los vectores de polaridad de las características relacionadas y los vectores de polaridad de los atributos, respectivamente.

Tabla 5.3. Agrupación de polaridades de características identificadas en torno a conceptos ontológicos

$$\begin{aligned}\vec{V}_1(\textit{película}, op_x) &= A * \vec{V}_1(\textit{película}, op_x) + B * \vec{V}_2(\textit{Harrison Ford}, op_x) + \\ &\quad B * \vec{V}_3(\textit{BSO}, op_x) + C * \vec{V}_4(\textit{duración}, op_x) \\ \vec{V}_2(\textit{Harrison Ford}, op_x) &= A * \vec{V}_2(\textit{Harrison Ford}, op_x) + B * \\ &\quad \vec{V}_1(\textit{película}, op_x) \\ \vec{V}_3(\textit{BSO}, op_x) &= A * \vec{V}_3(\textit{BSO}, op_x) + B * \vec{V}_1(\textit{película}, op_x) \\ \vec{V}_5(\textit{IPhone}, op_x) &= A * \vec{V}_5(\textit{IPhone}, op_x) + C * \vec{V}_6(\textit{chapa}, op_x) + C * \\ &\quad \vec{V}_7(\textit{sonido}, op_x)\end{aligned}$$

5.3.4 Análisis de sentimientos de opiniones

Uno de los principales inconvenientes encontrados a la hora de trabajar en el análisis de sentimientos de opiniones de usuarios es la ausencia de recursos lingüísticos en la Web. A la hora de implementar la metodología propuesta en esta tesis doctoral, se eligió el idioma inglés porque no existe actualmente una base de datos de sentimientos para el

idioma español de la misma calidad y completitud que *SentiWordNet* para el idioma inglés.

Existen algunas aproximaciones, como la presentada en (Perez-Rosas et al. 2012) donde se construyen dos bases de datos léxicas (lexicones) de sentimientos en español. El primer lexicón, llamado *FullStrengthLexicon*, contiene un léxico de sentimientos en español que aprovecha las anotaciones manuales de sentimientos hechas para el lexicón *OpinionFinder* (Wiebe et al., 2005). El segundo lexicón, llamado *MediumStrengthLexicon*, contiene un léxico que aprovecha las anotaciones de sentimiento realizadas para *SentiWordNet* (Esuli & Sebastiani, 2006). El formato de ambos lexicones⁴⁹ se muestra en la tabla 5.4. y tiene la siguiente estructura: palabra en español, posición del *synset* en *WordNet 1.6*, anotación en inglés y anotación en español (opcional). La anotación en inglés es la anotación de sentimiento generada automáticamente usando los lexicones *OpinionFinder* y *SentiWordNet*. La anotación en español es una anotación de sentimiento realizada por dos hablantes nativos de español (esta anotación es opcional y está disponible solamente para las 100 primeras palabras de cada lexicón). Además, en ocasiones aparecen divergencias entre una anotación y otra, como se puede apreciar en la tabla 5.4 para la anotación de sentimiento de los términos *abundancia* o *agudo*. Por otro lado, a diferencia de *SentiWordNet*, estos lexicones carecen de valores sentimentales numéricos asociados a las palabras de contenido emocional, lo cual supone una importante restricción para su uso.

Tabla 5.4. Extracto del lexicón *FullStrengthLexicon* construido para el idioma español

abundancia	5119096	neg	pos
abundante	13887	pos	pos
adecuado	51373	pos	pos
admirable	1807605	pos	pos
adorar	1777817	pos	pos
agitacion	7513508	neg	neg
agotamiento	14016361	neg	neg
agradable	89051	pos	pos
agudo	1214430	neg	pos
alerta	4664778	pos	neg

Otra aproximación más reciente es la que se propone en (Cruz et al., 2014). En este trabajo se construyen cinco lexicones de sentimientos a nivel de lema para los idiomas inglés, español, catalán, vasco y gallego. Para cada lema se provee una estimación de polaridad de muy negativa (-1.0) a muy positiva (+1.0), y una desviación estándar

⁴⁹ Disponibles en:

http://lit.csci.unt.edu/index.php?P=research/downloads#SPANISH_SENT_LEXICONS, última consulta realizada en Abril de 2015.

relacionada con la estimación de la polaridad. Todos los lexicones aportados en este trabajo están escritos en formato *XML* y reciben el nombre de *ML-Senticon*⁵⁰. La tabla 5.5 muestra un extracto del lexicon *ML-Senticon* en idioma español.

Tabla 5.5. Extracto del lexicon *ML-Senticon* construido para el idioma español

```
<?xml version="1.0" encoding="UTF-8" ?>
<senticon lang="es">
  <layer level="1">
    <positive>
      <lemma pos="a" pol="0.708" std="0.149"> acertado </lemma>
      <lemma pos="a" pol="0.906" std="0.125"> admirable </lemma>
      <lemma pos="n" pol="0.45" std="0.331"> admiración </lemma>
      <lemma pos="v" pol="0.75" std="0.177"> admirar </lemma>
      <lemma pos="a" pol="0.375" std="0.0"> afectivo </lemma>
      <lemma pos="n" pol="0.321" std="0.112"> afecto </lemma>
```

Sin embargo, como se ha comentado anteriormente, ninguna de estas aproximaciones tiene la misma completitud, generalidad y exhaustividad que *SentiWordNet*, ya que este lexicon para el idioma inglés contiene muchas más palabras (y sentidos de cada palabra) que los lexicones construidos en los trabajos citados para el idioma español. En los lexicones construidos para el idioma español no se contemplan las distintas polaridades asociadas a los distintos sentidos de una misma palabra. En la tabla 5.6 se muestra una comparativa del número de términos que contiene cada uno de los lexicones vistos para el idioma español frente a *SentiWordNet* (en su versión 3.0). Se ha contado como un término nuevo cada uno de los sentidos de cada palabra presente en *SentiWordNet*, ya que cada sentido de una misma palabra lleva asociados valores de polaridad diferentes.

Tabla 5.6. Comparación de lexicones construidos para el idioma español frente a *SentiWordNet*.

Lexicón	Número de entradas
<i>FullStrengthLexicon (ES)</i>	1.347
<i>MediumStrengthLexicon (ES)</i>	2.496
<i>ML-Senticon (ES)</i>	11.542
<i>SentiWordNet 3.0 (EN)</i>	206.941

De lo anterior se deduce que una nueva línea de investigación podría ser la construcción de un lexicon de contenidos emocionales en el idioma español que tuviese la misma completitud y generalidad que *SentiWordNet* para el idioma inglés. Habría que idear una metodología consistente para diseñar una base de datos de sentimientos de estas características. La traducción literal de las palabras existentes en *SentiWordNet* del inglés al español no funcionaría bien, ya que *SentiWordNet* incorpora valores de

⁵⁰ Multilingual, layered sentiment lexicons at lemma level. Disponible en: <http://www.lsi.us.es/~fermin/index.php?title=Datasets>. Última consulta realizada en Abril de 2015.

polaridad para palabras con contenido emocional que corresponden a sentidos muy específicos de la utilización de esa palabra en el idioma inglés.

Una vez que se dispusiese de un lexicón de palabras sentimentales para el idioma español de la misma calidad que SWN se podría validar la metodología propuesta en esta tesis doctoral utilizando un corpus de opiniones escritas en castellano. Por tanto, se podría llevar a cabo un estudio comparativo entre los resultados obtenidos aplicando la metodología en los mismos dominios pero en el idioma español.

Por otro lado, otra línea de investigación futura sería el análisis de los resultados obtenidos aplicando la metodología propuesta en esta tesis doctoral en otros dominios distintos del dominio de las películas cinematográficas o las finanzas. Por ejemplo, se podrían analizar los resultados obtenidos por la metodología al aplicarla a un dominio con unas características totalmente diferentes, como podría ser, por ejemplo, el dominio de la medicina.

5.3.5 Aplicación de la lógica difusa dentro de la metodología propuesta en esta tesis doctoral

La lógica difusa (o lógica *fuzzy*) se adapta muy bien al mundo real en el que vivimos, e incluso puede comprender y funcionar con expresiones del tipo "*hace mucho calor*", "*no es muy alto*", "*el ritmo del corazón está un poco acelerado*", etc. Su funcionamiento se basa en comprender los cuantificadores de cualidad (en los ejemplos de arriba: *mucho*, *muy* y *un poco*). Por tanto, una de las aplicaciones de la lógica difusa es la representación de la imprecisión en lenguaje natural (Kar & Mandal, 2011).

La lógica difusa se basa en reglas heurísticas de la forma "*SI (antecedente) ENTONCES (consecuente)*", donde el antecedente y el consecuente son también conjuntos difusos, ya sea puros o resultado de operar con ellos. Por ejemplo, se pueden tener reglas como "*SI hace muchísimo frío ENTONCES aumento drásticamente la temperatura*" o "*SI voy a llegar un poco tarde ENTONCES aumento levemente la velocidad*". Las reglas de las que dispone el motor de inferencia de un sistema difuso pueden ser formuladas por expertos o aprendidas por el propio sistema (Klir & Yuan, 1995).

De todo lo anterior se puede intuir que la lógica difusa puede desempeñar un papel importante dentro del proceso de análisis de sentimientos de opiniones de usuarios, ya que es frecuente que los usuarios utilicen expresiones imprecisas cuando dan su opinión. Además, es frecuente que este tipo de expresiones se utilicen para realzar las

emociones o sentimientos expresados en las opiniones. Los siguientes trabajos han estudiado la aplicación de la lógica difusa en el proceso de análisis de sentimientos de las opiniones de los usuarios en diferentes dominios (Martín-Bautista et al., 2002; Andreevskaia & Bergler, 2006; Abulaish et al., 2010; Kar & Mandal, 2011; Zhuo et al., 2014). Por ejemplo, el trabajo de Andreevskaia y Bergler (2006) se basa en el uso de lógica difusa para asignar etiquetas de sentimiento difusas a todas las palabras presentes en *WordNet*.

En la metodología propuesta en esta tesis doctoral no se tiene en cuenta el uso de cuantificadores indefinidos a la hora de ensalzar o disminuir los valores de polaridad asociados a una característica identificada en la opinión de un usuario. Una posible línea de investigación sería comprobar los efectos que tiene el uso de la lógica difusa dentro de la metodología expuesta en esta tesis doctoral para tratar este aspecto.

Por ejemplo, si consideramos una frase como “*La interpretación de Antonio Banderas es mala*” y una segunda frase como “*La interpretación de Antonio Banderas es radicalmente mala*”, se aprecia que en ambas frases se están mostrando emociones que afectan a la característica *Antonio Banderas* de manera negativa. Sin embargo en la segunda frase la carga de sentimiento es más alta que en la primera ya que se hace especial énfasis al afirmar que la interpretación es mala. El motivo es el uso del cuantificador de cualidad *radicalmente* que aparece en la segunda frase, que ensalza el carácter negativo del adjetivo *mala*. Sin embargo, en la metodología propuesta en esta tesis, el vector de polaridad asociado a la característica *Antonio Banderas* sería el mismo, ya que no se tendría en cuenta la mayor carga emocional de la segunda frase y solamente se mirarían los valores de sentimiento asociados al adjetivo calificativo *mala*.

Otro factor a tener en cuenta es el ámbito de influencia de los cuantificadores. En el trabajo de (Yllescas, 2005) se estudia el ámbito de actuación de los cuantificadores desde un punto de vista lingüístico.

En la metodología propuesta en esta tesis doctoral, el tener en cuenta la existencia de cuantificadores de cualidad en las opiniones de los usuarios implicará que se modifiquen los valores de los vectores de polaridad asociados a las características identificadas por el sistema. Esto causará cambios en la polaridad de las características, y por lo tanto, también en la polaridad de la opinión global de los usuarios, con lo cual, se obtendrán nuevos resultados que habrá que interpretar y comparar con los actuales.

Por ejemplo, si se analiza la frase en inglés “*this actor is more or less good*” y ahora se analiza otra opinión en este sentido: “*this actor is completely good*”, se ve que hay una

diferencia entre las emociones o sentimientos que se expresan en una frase y otra, a pesar de que se utiliza el mismo adjetivo calificativo para la característica *actor*, que en este caso es *good*. En la primera frase el cuantificador de cualidad o modificador *more or less* rebaja el sentido claramente positivo de la palabra *good*, mientras que en la segunda frase el modificador *completely* la realza. Por tanto, se puede emplear la lógica difusa en minería de opiniones para realzar o rebajar los sentimientos asociados a las palabras con contenido de sentimiento.

Por otro lado, si analizamos la frase “*this actor is not good*” vemos que se puede establecer otra clase de modificador que ni realza ni rebaja el contenido emocional de las palabras a las que va ligado, sino que en este caso lo invierte. En la metodología propuesta en esta tesis doctoral este tipo de modificador lingüístico ya está contemplado, ya que nos referimos a él cuando hablamos de presencia de ‘cláusula negativa’ en la frase de una opinión de un usuario.

En este contexto, la primera labor que se debería llevar a cabo sería la identificación de los posibles modificadores lingüísticos del idioma en se esté trabajando, y su clasificación en base a que realcen, rebajen o inviertan las emociones. Existen multitud de modificadores lingüísticos que alteran la carga emocional de las frases contenidas en las opiniones de los usuarios. En la tabla 5.7 se muestra una posible clasificación de modificadores que ensalzan, rebajan o invierten el contenido emocional de las palabras a las que van ligados. Se han definido tres clases de modificadores lingüísticos para el idioma inglés y dos categorías distintas dentro de cada clase en función de su intensidad. En total seis conjunto de modificadores denotados por M_1 , M_2 , M_3 , M_4 , M_5 y M_6 .

Teniendo en cuenta lo anterior, y sabiendo que la lógica difusa se basa en reglas heurísticas, se podrían establecer reglas gramaticales en las que tanto el antecedente como el consecuente sean difusos. Lógicamente, las reglas se establecerán de manera genérica y se almacenarán en una base de conocimiento de reglas difusas. Estas reglas establecerán el grado en el que cada modificador, dependiendo de la clase y categoría a la que pertenezca, modificará la carga emocional de la palabra de sentimiento a la que va ligado.

Tabla 5.7. Ejemplo de clasificación de modificadores lingüísticos difusos para el idioma inglés.**Clase 1: Modificadores que realzan los sentimientos**

- RADICALES = $M_1 = \{"completely", "extremely", "radically", "absolutely", "totally", \dots\}$
- MODERADOS = $M_2 = \{"very", "much", "lot", "quite", \dots\}$

Clase 2: Modificadores que rebajan los sentimientos

- RADICALES = $M_3 = \{"almost nothing", "hardly", \dots\}$
- MODERADOS = $M_4 = \{"little", "bit", "few", "more or less", "almost", \dots\}$

Clase 3: Modificadores que invierten los sentimientos

- RADICALES = $M_5 = \{"not", "not at all", "nothing", "completely not", "absolutely not", \dots\}$
- MODERADOS = $M_6 = \{"practically not", "practically nothing", \dots\}$

Por ejemplo, cuando el módulo *fuzzy* procese la frase: “*this actor is more or less good*” llevará a cabo lo siguiente: mirará que en la frase existe un cuantificador de cualidad presente (*more or less*). A continuación tratará de aplicar alguna regla presente en la base de conocimiento de reglas difusas. Si existe una regla difusa en la base de conocimiento que se pueda aplicar, entonces se aplicará y se modificará la carga emocional de la palabra a la que va ligado el modificador, que en este caso es *good*. Por tanto, si los valores de sentimiento asociados a la palabra *good* en la base de datos de valores de sentimiento son $(0.95, 0, 0.05)$, el módulo *fuzzy* mirará si existe una regla a aplicar en la base de conocimiento de reglas difusas cuando existe un modificador perteneciente al conjunto de modificadores M_4 (modificadores moderados que rebajan la carga emocional de las palabras) que en este caso es *more or less*.

Para este ejemplo se supone que existe una regla para modificadores M_4 que hace que se rebaje la carga emocional de la palabra en cierto grado establecido en la propia regla. Entonces el resultado de aplicar este modificador hará que los nuevos valores de sentimiento asociados a la palabra *good* sean, por ejemplo, $(0.475, 0, 0.525)$. Es decir, el modificador *more or less* ha rebajado la carga emocional de la palabra *good*, con lo cual este hecho tendrá una implicación fundamental en el cálculo de polaridades de

características y por consiguiente en el análisis global de las opiniones de los usuarios. El grado en el que cada modificador (dependiendo de la clase y categoría a la que pertenezca) incidirá en los valores de sentimiento de la palabra a la que se asocia vendrá definido en la propia regla. Además, el trabajo de definición de las reglas de la base de conocimiento de reglas difusas tendrá que ser llevado a cabo por expertos lingüistas.

Teniendo en cuenta todo lo anterior, la arquitectura del sistema para la minería de opiniones basada en características propuesta al inicio del Capítulo II, que marca el hilo conductor de la presente tesis doctoral, quedaría tal y como muestra la figura 5.2. En rosa aparecen las partes añadidas a la arquitectura general del sistema anteriormente expuesta. Se han añadido el módulo *fuzzy* y un nuevo recurso externo: la base de conocimiento de reglas difusas. Esta base de conocimiento será utilizada por el módulo *fuzzy* para realizar sus inferencias y aplicar las reglas difusas antes de que se calcule la polaridad de las características identificadas por el sistema en las opiniones de los usuarios.

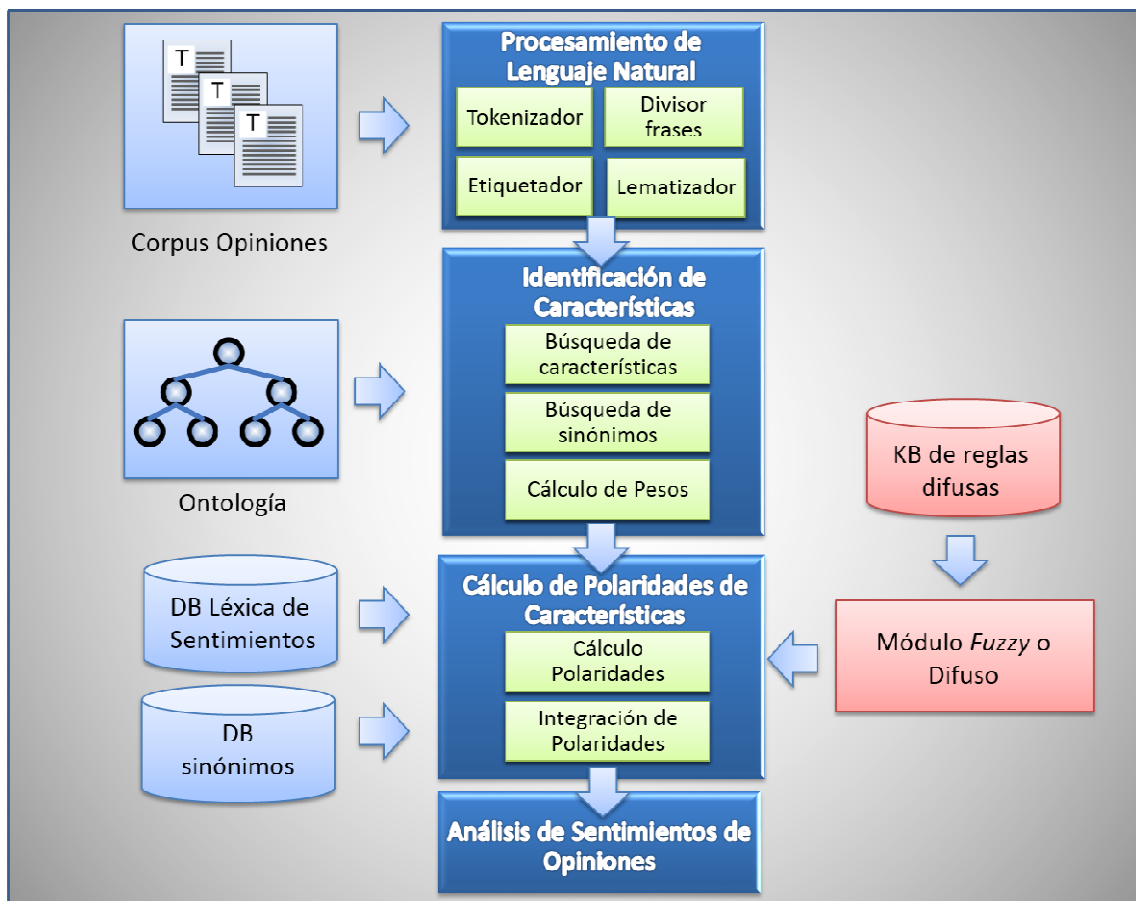


Figura 5.2. Arquitectura general del sistema con la inclusión del módulo *fuzzy*.

Capítulo VI

Contribuciones científicas

6.1 Publicaciones JCR

Peñalver-Martínez, I., García-Sánchez, F., Valencia-García, R., Rodríguez-García, M. A., Moreno, V., Fraga, A., Sánchez-Cervantes, J. L.

Feature-Based Opinion Mining through ontologies

Expert Systems with Applications (impact factor 2013: 1,965, Q1)

Clave: A Volumen: 41 (13) Páginas: 5995-6008 Fecha: 2014

Lugar de publicación: <http://dx.doi.org/10.1016/j.eswa.2014.03.022>

6.2 Publicaciones en revistas de otros índices de impacto

Peñalver-Martínez I., Valencia-García, R., García-Sánchez, F.

Minería de Opiniones basada en características guiada por Ontologías

Procesamiento del Lenguaje Natural (Latindex, 32 criterios)

Clave: A Volumen: 46 Páginas: 91-98 Fecha: Marzo, 2011

Lugar de publicación: SEPLN 848/703

6.3 Publicaciones en congresos internacionales

Peñalver-Martínez I., Valencia-García R., García-Sánchez F.

Ontology-guided approach for Feature-Based Opinion Mining

Congreso: 16th International Conference on Applications of Natural Language to Information Systems, NLDB 2011, Alicante, Spain

Fecha: 28/06/2011-30/06/2011

Proceedings: Lecture Notes in Computer Science Volume 6716, 2011, pp 193-200

Referencias

Bibliografía consultada y
referenciada

Referencias

- Abascal-Mena, R. (2009). Towards a semantic Web: ontology development based on the extraction of semantic concepts from digital documents. In Proceedings of the WSEAES 13th international conference on Computers, Rodos, Greece.
- Abulaish, M., Ahmad, T., & Doja, M. N. (2010). Opinion-Based imprecise query answering. In *Advances in Knowledge Discovery and Data Mining* (pp. 238-248). Springer Berlin Heidelberg.
- Airoldi, E. M., Cohen, W. W., Fienberg, S. E. (2005). Bayesian methods for frequent terms in text: Models of contagion and the Delta square statistic. Proceedings of the CSNA & INTERFACE Annual Meetings (2005).
- Alekh, A. and Bhattacharyya, P. (2005). Sentiment analysis: a new approach for effective use of linguistic knowledge and exploiting similarities in a set of documents to be classified. ICON 2005.
- Akerkar, R.A., Sajja, P.S. (2010). Knowledge-Based system, capítulo Natural Language Interface: Question Answering System, pp. 323–330. Jones and Barlett Publishers (2010).
- Alexopoulou, D., Wächter, T., Pickersgill, L., Eyre, C. & Schroeder, M. (2008). Terminologies for text-mining; an experiment in the lipoprotein metabolism domain. *BMC Bioinformatics*. 9 (Suppl 4). p.pp. 1-12.
- Amorim, M.T., Cury, D., Menezes, C.S. (2012). Um helpdesk inteligente baseado em ontologias. In: *Simpósio Brasileiro de Informática na Educação, SBIE 2012*, Rio de Janeiro (2012).
- Amorim, M.T., Cury, D., Menezes, C.S. (2011a). Um sistema inteligente baseado em ontologia para apoio ao esclarecimento de dúvidas. In: *Simpósio Brasileiro de Informática na Educação, SBIE 2011*, Aracaju-SE (2011).
- Amorim, M.T., Cury, D., Menezes, C.S. (2011b). Uma abordagem arquitetônica para um sistema pergunta-resposta. In: *Conferencia IADIS Ibero Americana*, Rio de Janeiro (2011).
- Andreevskaia, A. & Bergler, S. (2006) Mining wordnet for a fuzzy sentiment: Sentiment tag extraction from wordnet glosses, In Proceedings of the 11rd Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006), pp. 209–216, 2006.
- Antoniou, G. & vanHarmelen, F. (2004). *A Semantic Web Primer*. Cambridge, MA, USA: MIT Press.
- Ahmad, T., & Doja, M. N. (2012). Rule Based System for Enhancing Recall for Feature Mining from Short Sentences in Customer Review Documents. *International Journal on Computer Science & Engineering*, 4(6).
- Apostol, T. M. (2006). *Mathematical Analysis*, Addison-Wesley Publishing Company, Inc. Reading, Massachusetts, U.S.A. 2006.
- Arfken, G. (1985). *Mathematical Methods for Physicists*, 3rd ed. Orlando, FL: Academic Press, pp. 530-534, 1985.
- Avancini, H., (2000). *Information Retrieval & Automated Web Page Categorization*. ISISTAN – Research Report RRxx-00. Tandil (Bs. As.).
- Baccianella, S., Esuli, A., & Sebastiani, F. (2009). Multi-facet rating of product reviews, In *ECIR '09: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval* (pp. 461–472).

- Baccianella, S., Esuli, A. & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation* (pp. 2200–2204). European Language Resources Association.
- Balahur, A. & Montoyo, A. (2008). Applying a Culture Dependent Emotion Triggers Database for Text Valence and Emotion Classification. In *Proceedings of the AISB 2008 Convention “Communication, Interaction and Social Intelligence”*, 2008.
- Balahur, A., & Montoyo, A. (2010). Semantic approaches to fine and coarse-grained feature-based opinion mining. In *Natural Language Processing and Information Systems* (pp. 142-153). Springer Berlin Heidelberg.
- Balahur, A. (2011). *Methods and Resources for Sentimental Analysis in Multilingual Documents of Different Text Types*. PhD Thesis. University of Alicante. Alicante, Spain.
- Balakrishnan, A. V. (1981). *Applied Functional Analysis*, 2nd ed. New York: Springer-Verlag, 1981.
- Baldoni, M., Baroglio, C., Patti, V. & Rena, P. (2012). From tags to emotions: Ontology-driven sentiment analysis in the social semantic web. *Journal of Intelligenza Artificiale*. 10.3233/IA-2012-0028. Volume 6, Number 1 / 2012. May 16, 2012, 41-54. Italy.
- Balduccini, M., Baral, B.C., Lierler, Y. (2008). Knowledge representation and question answering. In: *Handbook of Knowledge Representation*, ch. 20. Elsevier (2008).
- Barforush, A. A., & Rahnama, A. (2012). Ontology learning: revisited. *Journal of Web Engineering*, 11(4), 269-289.
- Bautista-Thompson, E., Guzmán-Ramírez, E., & Figueroa-Nazuno, J. (2004). Predicción de Múltiples Puntos de Series de Tiempo Utilizando Support Vector Machines. *Computación y Sistemas*, 7(3), 148-155. México.
- Beineke, P., Hastie, T. and Vaithyanathan, S. (2004). The Sentimental Factor: Improving Review Classification via Human-Provided Information. *Proceedings of the ACL*, 2004.
- Berezansky, Y. M., Us, G. F. and Sheftel, Z. G. (1996). *Functional Analysis*, Vol. 1. Boston, MA: Birkhäuser, 1996.
- Berland, M., Charniak, E. (1999). *Finding Parts in Very Large Corpora*. Providence: Brown University.
- Bernal, D., Castro, A. & González, J. (2014). Web Semántica, más de una década de su aparición. *Puente. Revista científica*. Volumen 8, N°1.
- Berners-Lee, T. & Hendler, J. (2001). Publishing on the Semantic Web, *Nature*, April 26 2001 (pp. 1023-1025).
- Berners-Lee, T., Hendler, J. & Lassila, O. (2001). The semantic Web. *Scientific american*. 284 (5). p.pp. 28-37.
- Birkhoff, G. and Kreyszig, E. (1984). The Establishment of Functional Analysis. *Historia Math.* 11, 258-321, 1984.
- Borshevich, V., Mustyatsa, A., Oleinik, W. (1993). Fuzzy Spectral Analysis of Heart's Rhythms, Fifth IFSA World Congress, pp. 561-563.
- Borst, W.N. (1997). *Construction of Engineering Ontologies for Knowledge Sharing and Reuse*. CTIT Ph.D-thesis series No.97-14. University of Twente. Enschede, The Netherlands.
- Burgun, A., Bodenreider, O. (2001). *Mapping the UMLS Semantic Network into General Ontologies*. Bethesda, Maryland: National Library of Medicine.

- Cambria, E., Schuller, B., Liu, B., Wang, H., & Havasi, C. (2013). Knowledge-Based Approaches to Concept-Level Sentiment Analysis. *IEEE Intelligent Systems*, 28(2), 12-14.
- Cardoso, J. & Pinto A.M. (2015). The Web Ontology Language (OWL) and its Applications, *Encyclopedia of Information Science and Technology*, editor Mehdi Khosrow-Pour, Information Science Pub, 2015, pages 754-766.
- Carrión, M. (2014). Resolución de anáforas que requieren conocimiento cultural con la herramienta FunGramKB. *Revista de Lingüística y Lenguas Aplicadas* Vol. 9 año 2014, 1-13 EISSN 1886-6298.
- Castells, P., Foncillas, B., Lara, R., Rico, M. & Alonso, J. L. (2004) "Semantic Web technologies for economic and financial information management," in *Proceedings of the 1st European Semantic Web Symposium (ESWS '04)*, pp. 473–487.
- Cerini, S., Compagnoni, V., Demontis, A., Formentelli, M., & Gandini, G. (2007). Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining. *Language resources and linguistic theory: Typology, second language acquisition, English linguistics*, 200-210.
- Chang-Shing, L., Yuan-Fang, K., Yau-Hwang, K., Mei-Hui, W. (2007). Automated ontology construction for unstructured text documents, *Data & Knowledge Engineering*, 60(3), pp. 547-566.
- Chang-Su, K., Sung-Han, K. and Hoe-Kyung J. (2015). A study on Web standard-based *RDF* converter by applying Linked Data and using *RDF/XML* standard format for data. *International Journal of Software Engineering and Its Applications* Vol. 9, No. 1 (2015), pp. 1-12.
- Chen, L., Qi, L. & Wang, F. (2012). Comparison of feature-level learning methods for mining online consumer reviews. *Expert Systems with Applications*, 39, 9588–9601.
- Chen, H., & Zimbra, D. (2010). Ai and opinion mining. *IEEE Intelligent Systems*, 25, 74-80.
- Chowdhury, G.G. (2003). Natural language processing. *Annual Review of Information Science and Technology*. 37 (1). p.pp. 51-89.
- Clarke, S. J., & Willett, P. (1997). Estimating the recall performance of Web search engines. In *Aslib Proceedings* (Vol. 49, No. 7, pp. 184-189). MCB UP Ltd.
- Codina, L.; Rovira, C. (2006). La Web Semántica. En: Tramullas, J. *Tendencias en documentación digital*. Gijón: Ediciones Trea, pp. 9-54.
- Connolly, D., van Harmelen, F., Horrocks, I., McGuinness, D.L., Patel-Schneider, P.F., & Stein, L.A. (2001). DAML+OIL (March 2001) reference description. *W3C Note*, 18 December 2001. Consultado en (Julio 2014): <http://www.w3.org/TR/daml+oil-reference>.
- Criado, L. (2009). Procedimiento semi-automático para transformar la web en web semántica. Tesis Doctoral. Universidad de Navarra. Navarra, España.
- Cruz, F. L., Troyano, J. A., Pontes, B. & Ortega, F. J. (2014). Building layered, multilingual sentiment lexicons at synset and lemma levels. *Expert Systems with Applications*, 2014.
- Cruz, F. L., Troyano, J. A., Enríquez, F., Ortega, F. J. & Vallejo, C. G. (2013). 'Long autonomy or long delay?' The importance of domain in opinion mining. *Expert Systems with Applications*, 40, 3174–3184.
- Cunningham, H., Tablan, V., Roberts A. & Bontcheva, K. (2013) Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. *PLoS Comput Biol* 9(2): e1002854.

- Cunningham, H. et al. (2011). Text Processing with GATE (Version 6). University of Sheffield Department of Computer Science. 15 April 2011. ISBN 0956599311.
- Darriba, V.M. (2007). Universidad de Vigo: Asignatura Lenguajes Naturales. Campus de Orense. Chap. Tema 1 y 2.
- Dave, K., Lawrence, S., & Pennock, D. M. (2003) Mining the peanut gallery: opinion extraction and semantic classification of product reviews, In WWW '03: Proceedings of the 12th international conference on World Wide Web, pp. 519–528, 2003.
- David Beckett & Tim Berners-Lee (2008). Turtle - Terse *RDF* Triple Language. [Online]. Available from: <http://www.w3.org/TeamSubmission/2011/SUBM-turtle-20110328/>. [Accessed: 10 September 2014].
- Davis, B. P. (2013). On Applying Controlled Natural Languages for Ontology Authoring and Semantic Annotation. National University of Ireland Galway.
- Deitel, P. and Deitel, H. (2006). JavaTM. How to program, seventh edition. Prentice Hall Press. Upper Saddle River, NJ, USA.
- Deng, Z. H., Luo, K. H., & Yu, H. L. (2014). A study of supervised term weighting scheme for sentiment analysis. *Expert Systems with Applications*, 41(7), 3506-3513.
- Ding, X., Liu, B., & Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. In *Proceedings of the conference on Web search and Web data mining (WSDM'08)* (pp. 231–239).
- Doove, S., Gibcus, P., Kwaak, T., Smit, L., & Span, T. (2014). Survey on the access to finance of enterprises (SAFE).
- Durán, E., Zachman, P., Álvarez, M., Aguilera, J., Nuñez, J., Umaño, M. (2014). Metadatos y ontologías en el diseño de repositorios institucionales. XVI Workshop de Investigadores en Ciencias de la Computación. Red de Universidades con Carreras en Informática (RedUNCI). (p.p. 435-438).
- Durbin, S. D., Richter, J. N. and Warner, D. (2003). A System for Affective Rating of Texts. *Proceedings of the KDD Workshop on Operational Text Classification Systems (OTC-3)*, 2003.
- Durley, I., Guzmán-Luna, J. & Becerra, M.A. (2013). Semi-automatic extraction of information model for semantic markup of research object. *Institución universitaria Salazar y Herrera. Quid*, N°. 22, pp. 71-76, Ene-Jun, 2013, ISSN: 1692-343X, Medellín-Colombia.
- Edgar-Serna M. & Alexei-Serna A. (2014). Ontology for knowledge management in software maintenance. *International Journal of Information Management* Volume 34, Issue 5, October 2014, Pages 704–710.
- Eirinaki, M., Pisl, S., & Singh, J. (2012). Feature-based opinion mining and ranking. *Journal of Computer and System Sciences*, 78(4), 1175-1184.
- Esuli, A. (2008). Automatic generation of lexical resources for opinion mining: models, algorithms and applications. In *ACM SIGIR Forum*(Vol. 42, No. 2, pp. 105-106). ACM.
- Esuli, A. & Sebastiani, F. (2006). Sentiwordnet: A publicly available resource for opinion mining. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pp.417-422.
- Esuli, A. & Sebastiani, F. (2005). Determining the semantic orientation of terms through gloss classification, In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management* (pp. 617–624).
- Fakhrahmad, S.M., Sadreddini, M.H. and Zolghadri Jahromi, M. (2014). A proposed expert system for word sense disambiguation: deductive ambiguity resolution based on data mining and forward chaining. *Expert Systems*. doi: 10.1111/exsy.12075.

- Farman, A., Eun, K. & Yong-Gi K. (2015). Type-2 fuzzy ontology-based opinion mining and information extraction: A proposal to automate the hotel reservation system. *Applied Intelligence*. April 2015, Volume 42, Issue 3, pp 481-500.
- Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4), 82-89.
- Feldman, S. (1999). NLP Meets the Jabberwocky: Natural Language Processing in Information Retrieval. *ONLINE-WESTON THEN WILTON-*. 23. p.pp. 62-73.
- Fensel, D., van Harmelen, F., Horrocks, I., McGuinness, D.L. & Patel-Schneider, P.F. (2001) OIL: An ontology infrastructure for the semantic Web. *IEEE Intelligent Systems*, 16(2), pp. 38–45.
- Fernández, I. P. (2014). *La Crisis Financiera En El Cine: Orígenes, Desarrollos, Consecuencias*. Contribuciones a las Ciencias Sociales, (2014-02).
- Flanagan, D. (1999). *Java™ in a Nutshell: A Desktop Quick Reference*, Third edition, published November 1999. ISBN 1-56592-487-8E.
- Fox, M.S., Barbuceanu, M., Gruninger, M. & Lin, J. (1998) An organizational ontology for enterprise modelling, in *Simulating Organizations: Computational Models of Institutions and Groups*, pp. 131–152, MIT Press, Cambridge, Mass, USA.
- Gabrilovich, E. and Markovitch, S. (2004). Text Categorization with Many Redundant Features: Using Aggressive Feature Selection to Make SVMs Competitive with C4.5. *ICML 2004*.
- Gamma, E. (2003). *Patrones de Diseño - Elementos de Software orientado a objetos reutilizable*. ISBN: 8478290591. Pearson Educación S.A.
- Gamon M. (2005). Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis, In *COLING 2005* (pp. 841—847).
- Gantz, J.F., Reinsel, D., Chute, C., Schlichting, W., McArthur, J., Minton, S., Xheneti, I., Toncheva, A., Manfrediz, A. (2008), *The Expanding Digital Universe: A Forecast of Worldwide Information Growth Through 2008*, IDC & EMC. Consultado en (Marzo 2014).
- Ghiassi, M., Skinner, J., & Zimbra, D. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with Applications: An International Journal*, 40(16), 6266-6282.
- Giuliani, M., Marschall, T., & Isard, A. (2014). Using Ellipsis Detection and Word Similarity for Transformation of Spoken Language into Grammatically Valid Sentences. In *15th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (p. 243).
- Gladun, A., Rogushina, J., Valencia-García, R., & Béjar, R. M. (2013). Semantics-driven modelling of user preferences for information retrieval in the biomedical domain. *Informatics for Health and Social Care*, 38(2), 150-170.
- Goldberg A. B. & Zhu X. (2006) Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization, In *TextGraphs '06: Proceedings of TextGraphs: the First Workshop on Graph Based Methods for Natural Language Processing*, pp. 45–52, 2006.
- Goodchild, M.F. (2010). Twenty years of progress: GIScience in 2010. *Journal of spatial information science*. Number 1 (2010), pp. 3–20.
- Gruber, T. (2004). Interview Tom Gruber. *SIGSEMIS Bulletin*. 1 (3). p.pp. 4-9.
- Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5, pp. 199-220.

- Guarino, N. (1995). Formal Ontology, Conceptual Analysis and Knowledge Representation. *International Journal of Human and Computer Studies*, 43(5/6), pp. 625-640.
- Gubern, R. (2014). *Historia del cine*. Anagrama.
- Hai, Z., Chang, K., Kim, J. J., & Yang, C. C. (2014). Identifying Features in Opinion Mining via Intrinsic and Extrinsic Domain Relevance. *IEEE Trans. Knowl. Data Eng.*, 26(3), 623-634.
- Hatzivassiloglou, V. & McKeown, K. R. (1997) Predicting the semantic orientation of adjectives, In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pp. 174–181, 1997.
- Heflin, J., Hendler, J. & Luke, S. (1998). Reading between the lines: Using SHOE to discover implicit knowledge from the Web. In: *AAAI-98 Workshop on AI and Information Integration*. 1998.
- Heflin, J., Hendler, J. & Luke, S. (1999). Coping with changing ontologies in a distributed environment. In: *AAAI-99 Workshop on Ontology Management*. 1999.
- Hernández-González, Y., García-Moreno, C., Rodríguez-García, M.A., Valencia-García, R. & García-Sánchez, F. (2014) A semantic-based platform for R&D project funding management. *Computers in Industry*, 65(5): 850–861.
- Hirst, G. (1999). Review of "EuroWordNet: a multilingual database with lexical semantic networks" by Piek Vossen. *Kluwer Academic Publishers* 1999. *Computational Linguistics* 25(4). 628-630.
- Hofweber, T. (2009). "Logic and Ontology", in Edward N. Zalta, *Stanford Encyclopedia of Philosophy* (Spring 2009 Edition).
- Horrocks, I., Patel-Schneider, P.F. & Harmelen, F.V. (2003). From SHIQ and *RDF* to *OWL*: The Making of a Web Ontology Language. *Journal of Web Semantics*. 1. p.p. 2003.
- Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. USA, Seattle, WA.
- Hutson, V. & Pym, J. S. (1980). *Applications of Functional Analysis and Operator Theory*. New York: Academic Press, 1980.
- Hutton, A., Liu, A., Martin, C. (2012). Crowdsourcing Evaluations of Classifier Interpretability. *AAAI Spring Symposium on Wisdom of the Crowds*, 2012.
- Hye-Jin Min & Jong C. Park (2012). Identifying helpful reviews based on customer's mentions about experiences. *Expert Systems with Applications* 39 (2012) 11830–11838.
- Kruchten, P. (2004). *The rational unified process: an introduction*. Addison-Wesley Professional.
- James H. & Deborah L. M. (2000). The DARPA Agent Markup Language. 16 (6). p.pp. 67-73.
- Janowicz, K., Raubal, M. and Kuhn W. (2011). The semantics of similarity in geographic information retrieval. *Journal of spatial information science*. Number 2 (2011), pp. 29–57.
- Jindal, N., & Liu, B. (2006). Mining comparative sentences and relations. In *AAAI'06*.
- Jones, D. (1998). *Developing Shared Ontologies in Multi-agent Systems*. Liverpool: Department of Computer Science, University of Liverpool.
- Kamps J. & Marx, M. (2002) Words with attitude, In *Proc 1st International WordNet Conference*, pp. 332–341, 2002.

- Kar, A., & Mandal, D. P. (2011). Finding opinion strength using fuzzy logic on web reviews. *International Journal of Engineering and Industries*, 2(1), 37-43.
- Kauppinen, T., Henriksson, R., Väättäin, J., Deichstetter, C. & Hyvönen, E. (2006). Ontology-based modeling and visualization of cultural spatio-temporal knowledge. In *Developments in Artificial Intelligence and the Semantic Web - Proceedings of the 12th Finnish AI Conference STeP 2006*.
- Kietz, J.-U., Volz, R., Maedche, A. (2000). Extracting a Domain-Specific Ontology from a Corporate Intranet. *Proceedings of CoNLL-2000 and LLL-2000*, Lisbon, Portugal, 2000. 167-175.
- Kilgariff, A., Yallop, C. (2000). *What's in a thesaurus*. ITRI Reports Series. Brighton: ITRI. University of Brighton.
- Kim, Y., Jeong, M., & Jeong, S. R. (2014). Using Big Data Opinion Mining to Predict Rises and Falls in the Stock Price Index. *Handbook of Research on Organizational Transformations Through Big Data Analytics*, 30.
- Kim, S., Li, F., Lebanon, G. and Essa, I. (2012). Beyond sentiment: The manifold of human emotions. *arXiv:1202.1568v1*, 2012.
- Kim S. & Hovy, E. (2004) Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, pp. 1267–1373, 2004.
- Klir, G., & Yuan, B. (1995). *Fuzzy sets and fuzzy logic (Vol. 4)*. New Jersey: Prentice Hall.
- Klyne, G. & Carroll, J. (2004). *Resource Description Framework (RDF): Concepts and Abstract Syntax*. [Online]. Available from: <http://www.w3.org/TR/rdf-concepts/>.
- Kontopoulos, E., Berberidis, C., Dergiades, T. & Bassiliades, N. (2013) Ontology-based sentiment analysis of twitter posts. *Expert Systems with Applications*, 40, 4065–4074.
- Köthe, G. (1979). *Topological Vector Spaces*. New York: Springer-Verlag, 1979.
- Koza Orellana, W. (2015). Proposal for automatic extraction of medical term candidates with linguistic information processing description and evaluation of results. *Alfa: Revista de Linguística (São José do Rio Preto)*, 59(1), 113-128.
- Krauthammer, M. & Nenadic, G. (2004). Term identification in the biomedical literature. *Journal of biomedical informatics*. 37 (6). p.pp. 512-526.
- Kreyszig, E. (1989). *Introductory Functional Analysis with Applications*. New York: Wiley, 1989.
- Kreutzer, J. & Witte, N. (2014). *Opinion Mining Using SentiWordNet*. Semantic Analysis HT 2013/14. Uppsala University
- Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: *DBpedia A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia*. In: *Semantic Web (2012)*.
- Li, G. and Liu, F. (2012). Application of a clustering method on sentiment analysis. *Journal of Information Science* 38(2):127--139, 2012.
- Li, Q., Chen, Y.P., Myaeng, S., Jin, Y., Kang, B.Y. (2009). Concept unification of terms in different languages via Web mining for Information Retrieval. *Information Processing & Management*, 45(2), pp. 246-262.
- Liddy, Elizabeth D. (2001): "Natural language processing" en *Encyclopedia of Library and Information Science*, segunda edición, Nueva York: Marcel Decker.
- Liu, B. (2006). *Web Data Mining book*. Chapter 11.
- Liu, B. (2007). *Web Data Mining. Exploring Hyperlinks, Contents and Usage Data*. Springer, first edition.

- Losada, A. S., Bas, J. L., Bellido, S., Contreras, J., Benjamins, R. & Gomez, J. M. (2005) WP10: Case Study eBanking D10. 7 Financial Ontology. Data, Information and Process Integration with Semantic Web Services, Tech. Rep. FP6-507483.
- Lu, Y., Castellanos, M., Dayal, U., & Zhai, C. (2011). Automatic construction of a context-aware sentiment lexicon: an optimization approach. In Proceedings of the 20th international conference on World wide web (pp. 347-356). ACM.
- Lu, W.H., Chien, L.F. & Lee H.J. (2002). Translation of Web queries using anchor text mining. ACM Transactions on Asian Language Information Processing (TALIP), 1(2), pp. 159–172.
- Luke, S., Spector, L., Rager, D. & Hendler, J. (1997). Ontology-based Web Agents. In: Proceedings of the First International Conference on Autonomous Agents. AGENTS '97. [Online]. New York, NY, USA: ACM, pp. 59-66. Available from: <http://doi.acm.org/10.1145/267658.267668>. [Accessed: 05 January 2015].
- Malková, E. (2014). Ellipsis and sentence fragments in Ian McEwan's Amsterdam: their effect on meaning.
- Malik, S. K., Prakash, N., & Rizvi, S. (2010). Semantic annotation framework for intelligent information retrieval using KIM architecture. International Journal of Web & Semantic Technology (IJWest), 1(4), 12-26.
- Mandala, R.; Tokunaga, T.; Tanaka, H. (1999). Session: Complementing WordNet with Roget's and corpus-based thesauri for information retrieval. [en línea]. En: Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics. Morristown: Association for Computational Linguistics. http://portal.acm.org/ft_gateway.cfm?id=977049&type=pdf&coll=portal&dl=ACM&CFID=19358554&CFTOKEN=67326162 [Consulta: 29 abril 2014]
- Martí Antonín, M. A., Castellón Masalles, I. (2000). Lingüística computacional. Barcelona: Edicions Universitat de Barcelona.
- Martí Antonín, M. A., Taulé Delor, M., Màrquez Villodre, L., Bertran Ibarz, M. (2007) "Anotación semiautomática con papeles temáticos de los corpus CESS-ECE". Procesamiento del lenguaje natural. N. 38 (abr. 2007). ISSN 1135-5948, pp. 67-76.
- Martín-Bautista, M. J., Kraft, D. H., Vila, M. A., Chen, J., & Cruz, J. (2002). User profiles and fuzzy logic for web retrieval issues. Soft computing, 6(5), 365-372.
- Martín-Valdivia, M. T., Martínez-Cámara, E., Perea-Ortega, J. M. & Ureña-López, L. A. (2013). Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches. Expert Systems with Applications, 40, 3934–3942.
- Mazón, E. (2008). El valor de los contenidos no estructurados en la gestión empresarial, Revista Auditoria y seguridad, N 21, Perspectiva Empresarial, pp. 86-87.
- McCarthy, J. (1959). Programs with Common Sense. In Mechanisation of Thought Processes, Proceedings of the Symposium of the National Physics Laboratory, pages 77-84, London, U.K.
- McGuinness, D.L. & Van Harmelen, F. (2004). OWL Web ontology language overview. W3C recommendation. 10 (10). p.p. 2004.
- Meyer, B. (1998). Construcción de software orientado a objetos. ISBN: 9788483220405. Prentice-Hall.
- Miao, Q., Li, Q., & Dai, R. (2008). An integration strategy for mining product features and opinions. In Proceedings of the 17th ACM conference on information and knowledge management (pp. 1369–1370). Napa Valley, California.
- Milios, E., Zhang, Y., He, B. & Dong, L. (2003). Automatic Term Extraction And Document Similarity In Special Text Corpora. 6th Conference of the Pacific Association for Computational Linguistics. p.pp. 275-284.

- Miller, P., & Hemforth, B. (2013). Verb phrase ellipsis with nominal antecedents. Ms. Université Paris Diderot.
- Miller, G. (1995). WordNet: a lexical database for English. *Communications of the ACM* 38(11). 39-41.
- Min, H. J., & Park, J. C. (2012). Identifying helpful reviews based on customer's mentions about experiences. *Expert Systems with Applications*, 39(15), 11830-11838.
- Modoni, G.E., Sacco, M. & Terkaj, W. (2014). A survey of *RDF* store solutions. In *Engineering, Technology and Innovation (ICE)*, 2014 International ICE Conference on. 14516690. Bergamo, Italy.
- Moraes, R., Valiati, J.F., Neto, W.P. (2013) Document-level sentiment classification: An empirical comparison between SVM and ANN. In *Proceedings of Expert Systems with Applications: An International Journal* (pp. 621-633). Tarrytown, NY, USA.
- Moreno, A., Pineda, F. & Hidalgo, R. (2010). Análisis de Valoraciones de Usuario de Hoteles con Sentitext: un sistema de análisis de sentimiento independiente del dominio, *Procesamiento del Lenguaje Natural*, 45, 31–39.
- Moreo, A., Romero, M., Castro, J.L., Zurita J.M. (2012). Lexicon-based Comments-oriented News Sentiment Analyzer system. *Expert Systems with Applications* 39 (2012) 9166–9180.
- Mukras R. & Carroll J. (2004) A comparison of machine learning techniques applied to sentiment classification, R Mukras in Masters thesis University of Sussex Falmer Brighton (2004) 2004.
- Mullen, T. and Collier, N. (2004). Sentiment Analysis using Support Vector Machines with Diverse Information Sources. *Proceedings of EMNLP*, 2004.
- Narock, T., Zhou, L. and Yoon, V. (2012). Semantic similarity of ontology instances using polarity mining. *Journal of the American Society for Information Science and Technology* Volume 64, Issue 2.
- Neches, R., Fikes, R.E. Finin, T., Gruber, T.R., Senator, T., Swartout, W.R. (1991). Enabling technology for knowledge sharing. *AI Magazine*, 12(3), pp. 36-56.
- Negacy, D., Bretonnel, C., Lawrence E. (2014). Ontology Translation: A Case Study on Translating the Gene Ontology from English to German. *Natural Language Processing and Information Systems. Lecture Notes in Computer Science* Volume 8455, 2014, pp 33-38.
- Ochoa, J. L., Valencia-García, R., Perez-Soltero, A., & Barceló-Valenzuela, M. (2013). A semantic role labelling-based framework for learning ontologies from Spanish documents. *Expert Systems with Applications*, 40(6), 2058-2068.
- Ohana, B. & Tierney, B. (2009). Sentiment classification of reviews using SentiWordNet. 9th. IT & T Conference.
- Padró, L. (2011). Analizadores Multilingües en FreeLing. *Linguamatica*, vol. 3, n. 2, pg. 13-20. December, 2011.
- Padró, L. & Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)* ELRA.Istanbul, Turkey. May, 2012.
- Pareja, I. V. (2003). Decisiones empresariales bajo riesgo e incertidumbre. Editorial Norma.
- Partridge, C. & Stefanova, M. (2001) A Synthesis of State of the Art Enterprise Ontologies, Lessons Learned, The BORO Program, LADSEB CNR.

- Pang B., Lee L. & Vaithyanathan S. (2002) Thumbs up? Sentiment Classification using machine learning methods. In EMNLP-2002. pp 79-86.
- Pang B. & Lee L. (2004). A Sentiment Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In Proceedings of the ACL, (pp. 271–278).
- Pang B. & Lee L. (2008). Opinion mining and sentiment analysis. *Foundation and Trends in Information Retrieval*, 2(1-2), 1-135.
- Patel, S.K., Bhadka, H.B. (2015). Semantic Web technology and ontology designing for e-learning environments. (*IJCSIT*) *International Journal of Computer Science and Information Technologies*, Vol. 6 (1) , 2015, 48-51
- Pedraza-Jiménez, R., Codina, L., Rovira, C. (2007). Web Semántica y ontologías en el procesamiento de la información documental. En “El profesional de la información”, 2007, noviembre-diciembre, 16 (6), pp. 569-578.
- Peis, E., Herrera-Viedma, E., Hassan, Y., Herrera, J. C. “Análisis de la web semántica: estado actual y requisitos futuros”. En: *El profesional de la información*, 2003, septiembre-octubre, v. 12, n. 5, pp. 368-376.
- Peng, F. (2003). *Language Independent Text Learning with Statistical n-gram Language Models*. PhD thesis, University of Waterloo, 2003.
- Pérez, A. L. G., Rodríguez, A. C., & Molina, M. A. (2002). Factores determinantes de la rentabilidad financiera de las PYMES. *Spanish Journal of Finance and Accounting/Revista Española de Financiación y Contabilidad*, 31(112), 395-429.
- Perez-Rosas, V., Banea, C. & Mihalcea, R. (2012). Learning Sentiment Lexicons in Spanish, in *Proceedings of the International Conference on Language Resources and Evaluations (LREC 2012)*, Istanbul, Turkey, May 2012.
- Pulido, J.R.G., Ruiz, M.A.G., Herrera, R., Cabello, E., Legrand, S. and Elliman, D. (2006). Ontology languages for the semantic Web: A never completely updated review. *Knowledge-Based Systems*, 19(7), pp. 489-497.
- Qiong Wu & Songbo Tan (2011). A two-stage framework for cross-domain sentiment classification. *Expert Systems with Applications* 38 (2011) 14269–14275.
- Quine, W.O. (1961). *From a Logical Point of View, Nine Logico-Philosophical Essays*. Cambridge, Massachusetts: Harvard University Press.
- Reed, M. and Simon, B. (1972). *Methods of Modern Mathematical Physics, Vol. 1: Functional Analysis*. New York: Academic Press, 1972.
- Refoufi, A. (2014). Pronominal Anaphora Resolution Using XML Tagged Documents. *J Comput Eng Inf Technol* 3: 1. doi: <http://dx.doi.org/10.4172/2324.9307>, 2.
- Restrepo, J. A., & Medina, S. (2014). Operational Risk Estimation Uncertainty on Environment: A Case Study (Estimación Del Riesgo Operativo Bajo Ambiente De Incertidumbre: Estudio De Caso). *Revista Internacional Administración & Finanzas*, 7(7), 39-54.
- Richard Cyganiak, David Wood & Markus Lanthaler (2004). *RDF Vocabulary Description Language 1.0: RDF Schema*. Changes. [Online]. Available from: <http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/>. [Accessed: 08 January 2015].
- Rivero, D., Vila, J., Rivero, Y. (2014). Ontologías. Integración de esquemas. *Tlatemoani Revista Académica de Investigación Editada por Eumed.net* No. 17 Diciembre 2014 España ISSN: 19899300.
- Rodríguez-García, M.A., Valencia-García, R., García-Sánchez, F. & Samper-Zapater, J.J. (2014a) Creating a semantically-enhanced cloud services environment through ontology evolution. *Future Generations in Computer Systems*. vol. 32, pp. 295–306.

- Rodríguez-García, M. A., Valencia-García, R., García-Sánchez, F. & Samper-Zapater, J. J. (2014b). Ontology-based annotation and retrieval of services in the cloud. *Knowledge-Based Systems*, 56, 15-25.
- Ruiz F., Vizcaino A., Piattini M. & Garcia F. (2004). An Ontology for the Management of Soft-ware Maintenance Projects. *International Journal of Software Engineering and Knowledge Engineering*, Vol. 14, No. 3, (2004) pp 1-27.
- Ruiz-Martínez, J.M (2012). Metodología para la población automática de ontologías. Aplicación en los dominios de medicina y turismo. *Procesamiento de Lenguaje Natural*, Revista nº 48 marzo de 2012, pp 123-126.
- Ruiz-Martínez, J.M., Valencia-García, R., García-Sánchez, F. (2012). Semantic-Based Sentiment analysis in financial news. In *Proceedings of 1st International Workshop on Finance and Economics on the Semantic Web (FEOSW 2012)*, In conjunction with 9th Extended Semantic Web Conference (ESWC 2012).
- Rumbaugh, J., Jacobson, I., & Booch, G. (2004). *Unified Modeling Language Reference Manual*, The. Pearson Higher Education.
- Rushdi-Saleh, M., Martín-Validvia, M. T., Montejo-Ráez, A. & Ureña-López, L.A. (2011). Experiments with SVM to classify opinions in different domains. *Expert Systems with Applications*, 38, 14799–14804.
- Saif, H., He, Y. and Alani, H. (2012) Alleviating data sparsity for Twitter sentiment analysis. *Workshop on Making Sense of Microposts*, 2012.
- Sala, A., Picó, J., Bondia, J. (2000). *Tratamiento de la incertidumbre en modelado y control borrosos*. Universidad Politécnica de Valencia, 2000.
- Salton, G. & McGill, M.J. (1983). *Introduction to modern information retrieval*.
- Salveti, F., Lewis, S. and Reichenbach, C. (2004). Impact of Lexical Filtering on Overall Opinion Polarity Identification. *Proceedings of the AAAI Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, 2004.
- Santorini, B. (1990). Par-of-Speech tagging guidelines for the Penn Tree Bank Project. Technical report MS-CIS-90-47 y Santorini B. (1993).
- Sasaki, Y. (2005). Question Answering as Question-Biased Term Extraction: A New Approach toward Multilingual QA. In *Proceedings of 43th Annual Meeting of the Association for Computational Linguistics*, Michigan, USA, June 2005.
- Shaban-Nejad, A., Menon, S. & Buckeridge, D. (2015). A Semantic Web Platform for Online Vaccine Sentiment Surveillance. *ISDS-International Society Disease Surveillance. Conference Abstracts*.
- Shadbolt, N., Berners-Lee, T. & Hall, W. (2006) 'The Semantic Web Revisited'. *IEEE Intelligent Systems* 21(3): 96-101 (2006).
- Shimada K. & Endo T. (2008) Seeing several stars: A rating inference task for a document containing several evaluation criteria, In *PAKDD 2008: Proceedings of Advances in Knowledge Discovery and Data Mining*, 12th Pacific-Asia Conference, pp. 1006–1014, 2008.
- Singh, V. K., Piryani, R., Uddin, A., & Waila, P. (2013, March). Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. In *Automation, Computing, Communication, Control and Compressed Sensing (iMac4s)*, 2013 International Multi-Conference on (pp. 712-717).
- Sista, S. P. and Srinivasan, S. H. (2004). Polarized Lexicon for Review Classification. *Proceedings of the International Conference on Machine Learning; Models, Technologies & Applications*, 2004.

- Smailovic, J., Grcar, M., Lavrac, N. & Znidarsic, M. (2014). Stream-based active learning for sentiment analysis in the financial domain. *Information Sciences* 285 (2014) 181–203.
- Sosa, E. (1997). Procesamiento del lenguaje natural: revisión del estado actual, bases teóricas y aplicaciones (Parte I). *El profesional de la información*.
- Stojanovic, L., Maedche, A., Motik, B., & Stojanovic, N. (2002). User-driven ontology evolution management. In *Knowledge engineering and knowledge management: ontologies and the semantic Web* (pp. 285-300). Springer Berlin Heidelberg.
- Stone, P., Dumphy, D. C., Smith, M. S., and Ogilvie, D. M. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press, Cambridge, USA.
- Strang, G. (2009). *Introduction to Linear Algebra*. 4th ed. Wellesley, MA: Wellesley-Cambridge Press, February 2009. ISBN: 9780980232714.
- Strapparava, C. & Valitutti, A. (2004). Wordnet-affect: an affective extension of wordnet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pp.1083-1086, Lisbon, Portugal.
- Studer, R., Benjamins, V. R. & Fensel D. (1998). *Knowledge engineering: Principles and methods*, *Data Knowl. Eng.*, 25, 161–197.
- Talitha, M., Davidson, C., Silva de Menezes, C. (2014). On the use of ontologies to guide agents answering questions. *New Perspectives in Information Systems and Technologies, Volume 1. Advances in Intelligent Systems and Computing Volume 275*, 2014, pp 525-533.
- Toledo-Gómez, I., Valtierra-Romero, E., Guzmán-Arenas, A., Cuevas-Rasgado, A., & Méndez-Segundo, L. (2014). AnaPro, Tool for Identification and Resolution of Direct Anaphora in Spanish. *Journal of Applied Research and Technology*, 12(1).
- Toutanova K. & Manning, C. (2000). Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora EMNLP/VLC-2000* (pp. 63–70). Hong Kong.
- Toutanova, K., Klein, D., Manning, C. & Singer, Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of HLT-NAACL 2003* (pp. 252–259).
- Trillas, E., Alsina, C., Terricabras, J. (1995). *Introducción a la Lógica Borrosa*, Barcelona, 1995.
- Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings 40th Annual Meeting of the Association for Computational Linguistics, ACL 2002*, 417-424, Philadelphia, USA.
- Turney, P. D. & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association, *ACM Trans. Inf. Syst.*, vol. 21, no. 4, pp. 315–346, 2003.
- Ureña, L.A. (2001). Resolución de la ambigüedad léxica en tareas de clasificación automática de documentos. *Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN)*. Editorial Club Universitario, 1 de dic. de 2001. 156 páginas.
- Vapnik, V., Guyon, I., Boser, B.E. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of COLT '92 Proceedings of the fifth annual workshop on Computational learning theory* (pp. 144-152). New York, USA.
- Valencia-García R, Ruiz-Sánchez JM, Vivancos-Vicente PJ, Fernández-Breis JT, Martínez-Béjar R, (2004). An incremental approach for discovering medical knowledge from texts. *Expert Systems with Applications*. 26(3), pp. 291-299.

- Valencia-García, R. (2005). Un entorno para la extracción incremental de conocimiento desde texto en lenguaje natural. Tesis doctoral. ISSN: 84-689-5046-7, University of Murcia, Spain.
- Valencia-García, R., Fernández-Breis, J.T., Ruiz-Martínez, J.M., García-Sánchez, F., Martínez-Béjar, R. (2008). A knowledge acquisition methodology to ontology construction for information retrieval from medical documents. *Expert Systems: The Knowledge Engineering Journal* 25(3), pp. 314-334.
- Varela, M. D. L. L. C. (2014). Resolución de anáforas que requieren conocimiento cultural con la herramienta Fungramkb. *Revista de Lingüística y Lenguas Aplicadas*, 9(1), 01-13.
- Vázquez, S., A. Montoyo, and G. Rigau. (2003). Método de desambiguación léxica basada en el recurso léxico: dominios relevantes. *Procesamiento del Lenguaje Natural*, 31(2003): 141-148.
- Vega, W. & Umaña, H. (2014). Diseño de Servicios Web Semánticos utilizando el desarrollo de software dirigido por modelos. En: *Ventana Informática* No. 30 (ene-jun). Manizales (Colombia): Facultad de Ciencias e Ingeniería, Universidad de Manizales. p. 97-108. ISSN: 0123-9678.
- Voorhees, E. M. (1994). On expanding query vectors with lexically related words. *Proceedings of the 2nd Text Retrieval Conference (TREC-2)*.
- W3C (2012a). Owl 2 functional syntax. [Online]. Available from: <http://www.w3.org/TR/2012/REC-owl2-syntax-20121211/>. [Accessed: 10 September 2014].
- W3C (2012b). Owl 2 xml serialization. [Online]. Available from: <http://www.w3.org/TR/owl2-xml-serialization/>. [Accessed: 10 September 2014].
- W3C (2014). Owl 2 rdf/xml syntax specification. [Online]. Available from: <http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/>. [Accessed: 10 September 2014].
- Wang, Z. (2014). Extraction system for Personal Attributes Extraction of CLP2014. In *Proc. The Third SIGHAN Workshop on Chinese Language Processing*.
- Whitelaw, C., Garg, N., and Argamon, S. (2005). Using appraisal groups for sentiment analysis. In *Proceedings of the CIKM 2005*: 625-631, Bremen, Germany.
- Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3), 165-210.
- XBRL International (2009). XBRL: eXtensible Business Reporting Language, <http://www.xbrl.org>.
- Yang, Y. & Liu, X. (1999). A Re-examination of Text Categorization Methods. In: *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '99*. 1999, New York, NY, USA: ACM, pp. 42-49.
- Yllescas, J. C. T. (2005). El ámbito de los cuantificadores I. Lo que la gramática sí codifica. *Interlingüística*, (16), 1105-1117.
- Yoshida, K. (1971). *Functional Analysis and Its Applications*. New York: Springer-Verlag, 1971.
- Zeidler, E. (1989). *Nonlinear Functional Analysis and Its Applications*. New York: Springer-Verlag, 1989.
- Zeidler, E. (1995). *Applied Functional Analysis: Applications to Mathematical Physics*. New York: Springer-Verlag, 1995.
- Zhai, Z., Xu, H., Kang, B., & Jia, P. (2011). Exploiting effective features for chinese sentiment classification. *Expert Systems with Applications*, 38, 9139–9146.

- Zhang, F., Ma, Z.M. & Li, W. (2015). Storing *OWL* ontologies in object-oriented databases. *Knowledge-Based Systems*. Volume 76, March 2015, Pages 240–255.
- Zhao, L. & Li, C. (2009). Ontology based opinion mining for movie reviews, In *Proceedings of the 3rd International Conference on Knowledge Science, Engineering and Management* (pp. 204–214).
- Zhou, L. & Chaovalit, P. (2008). Ontology-supported polarity mining, *Journal of the American Society for Information Science and Technology*, 59, 1, 98–110.
- Zhuo, S., Wu, X., & Luo, X. (2014). Chinese text sentiment analysis based on fuzzy semantic model. In *Cognitive Informatics & Cognitive Computing (ICCI* CC)*, 2014 IEEE 13th International Conference on (pp. 535-540). IEEE.
- Zimmer, R. (1990). *Essential Results in Functional Analysis*. Chicago: University of Chicago Press, pp. 13-17, 1990.