Complex & Intelligent Systems (2023) 9:2893–2914
https://doi.org/10.1007/s40747-022-00693-x

ORIGINAL ARTICLE

Evaluating feature combination strategies for hate-speech detection in Spanish using linguistic features and transformers

José Antonio García-Díaz[1] · Salud María Jiménez-Zafra[2] · Miguel Angel García-Cumbreras[2] · Rafael Valencia-García[1]

Received: 30 June 2021 / Accepted: 11 February 2022 / Published online: 26 February 2022

Abstract
The rise of social networks has allowed misogynistic, xenophobic, and homophobic people to spread their hate-speech to intimidate individuals or groups because of their gender, ethnicity or sexual orientation. The consequences of hate-speech are devastating, causing severe depression and even leading people to commit suicide. Hate-speech identification is challenging as the large amount of daily publications makes it impossible to review every comment by hand. Moreover, hate-speech is also spread by hoaxes that requires language and context understanding. With the aim of reducing the number of comments that should be reviewed by experts, or even for the development of autonomous systems, the automatic identification of hate-speech has gained academic relevance. However, the reliability of automatic approaches is still limited specifically in languages other than English, in which some of the state-of-the-art techniques have not been analyzed in detail. In this work, we examine which features are most effective in identifying hate-speech in Spanish and how these features can be combined to develop more accurate systems. In addition, we characterize the language present in each type of hate-speech by means of explainable linguistic features and compare our results with state-of-the-art approaches. Our research indicates that combining linguistic features and transformers by means of knowledge integration outperforms current solutions regarding hate-speech identification in Spanish.

Keywords Hate-speech · Feature engineering · Knowledge integration · Text classification · Natural language processing

Introduction

Offensive speech is defined as speech that offends someone. A text is considered offensive if it includes any form of unacceptable language, that is, whether it contains insults, threats or bad language [3]. Offensive speech varies widely, from simple profanity to much more serious types of speech [53]. One of the most problematic types of offensive language is hate-speech, since the presence of hate-speech on social media platforms has been shown to correlate with real-life hate crimes [39].

It is quite difficult to distinguish between offensive language and hate-speech as there are few universal definitions [11]. However, all definitions agree in that hate-speech is the language that targets a person or group with the intent to be harmful or cause social chaos. This targeting is usually done on the basis of some characteristics such as race, gender, sexual orientation, nationality, or religion [52]. Offensive language, on the other hand, is a more general category containing any kind of profanity or insult. Hate-speech can, therefore, be classified as a subset of offensive language. Zampieri et al. [59] propose guidelines for classifying offensive language, as well as the type and target of offensive language. These guidelines collect the characteristics of offensive language in general, hate-speech, and other types of targeted offensive language, such as cyberbullying [25,31].

✉ Rafael Valencia-García
valencia@um.es

José Antonio García-Díaz
joseantonio.garcia8@um.es

Salud María Jiménez-Zafra
sjzafra@ujaen.es

Miguel Angel García-Cumbreras
magc@ujaen.es

[1] Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, Spain

[2] Computer Science Department, SINAI, CEATIC, Universidad de Jaén, 23071 Jaén, Spain

Given the enormous amount of user-generated content on social networks, it is not feasible to rely on manual supervision to stop hate-speech. In view of that, this study aims to contribute to the detection of hate-speech in Spanish, due to the growing need for research on this topic in languages other than English [15]. For this purpose, we compile the existing Spanish hate-speech corpora and analyze several classification techniques based on linguistic features, transformers and different integration mechanisms. We selected these approaches because they have enabled significant advances in most Natural Language Processing (NLP) classification tasks [54]. To this end, we define the following research questions:

- RQ1. Which individual features are most effective for hate-speech detection?
- RQ2. How can features be integrated for more robust systems?
- RQ3. Is it possible to characterize the language of the different hate-speech types by means of explainable linguistic features?
- RQ4. Do our methods improve the results of the state-of-the-art?

The main contributions of this work are:

1. Hate-speech detection in Spanish. We focus on Spanish, since the number of existing works in this language is small compared to English and it is very important to increase the reliability of hate-speech detection systems in other languages as well.
2. Compilation of all existing datasets in Spanish for hate-speech detection and experimentation with transformers and state-of-the-art features. Existing studies to date work with one or two of the best known datasets in Spanish (HaterNet and HatEval). However, there are more datasets that the scientific community should be aware of and that could help to advance the study of this phenomenon.
3. Use of the UMUTextStats tool [18,19] and fine-grain negation features [27,28] in order to characterize the language present in each type of hate-speech by means of explainable linguistic features.
4. A new method for hate-speech detection in Spanish that outperforms those of the state of the art. Our proposal based on the combination of linguistic features and transformers outperforms current solutions.

The rest of the paper is organized as follows: "State of the art" compiles novel studies and shared-tasks regarding hate-speech detection using NLP, with special interest in those oriented to Spanish. "Datasets" describes in detail the datasets involved in this study. In "Materials and methods", the reader can find a detailed description of our pipeline, including the feature sets and the deep-learning architectures evaluated. In "Results and analysis" we show and analyze the results of the experiments carried out to answer each of the research questions we formulated. Finally, "Conclusion and further work" summarizes the insights achieved and proposes promising research directions.

State of the art

Hate-speech detection using NLP is a recent task in Spanish, so the number of existing studies is limited. However, its importance has led to an increasing number of researchers focusing on this topic. For an overview on hate-speech detection we find two insightful surveys. On the one hand, in [52], the authors present the terminology employed to talk about this topic, and analyze the methods and features used for hate-speech classification. Moreover, they give some insights for data annotation and pose the context and the language as challenges, because most of the research are on English data. On the other, Fortuna and Nunes [15] also identify the need of studies in languages other than English. In their survey, they analyze the concept of hate-speech from different perspectives and provide a helpful definition for building automatic detection systems. In addition, they compare hate-speech with some related concepts: cyberbullying, abusive language, discrimination, toxicity, flaming, extremism and radicalization. Furthermore, they conduct a systematic literature review and analyze the works based on approaches, algorithms, features, and datasets, among others. Finally, they identify challenges and opportunities.

Most studies on hate-speech focus on the identification of racism [51,55], the detection of misogyny [17,18,47], the identification of xenophobia [7,47], and the recognition of hate in general [3,10,31]. In fact, we can find a large set of shared-tasks about these topics, such as the AMI shared-task on Automatic Misogyny Identification at IberEval 2018 [14] and Evalita 2018 [6], the 2019 and 2020 editions of the HASOC track on hate-speech and offensive content identification [30,32,35,36], and the HatEval shared-task on the Detection of hate-speech against Immigrants and Women [4], among others. Regarding the origin of the analyzed data, there are different sources, such as Twitter [26], Facebook [49], YouTube [50], and Yahoo! [57], being Twitter the most commonly used. With respect to the languages in which the studies are conducted, there are studies in Arabic [1], Croatian [33], Danish [53], Dutch [55], English [14], French [40], German [23], Hindi–English [5], Indonesian [2], Italian [6], Portuguese [16], Spanish [3] and Turkish [9], but by far the majority of them are in English. As our work focuses on Spanish, we present below a brief review of works on the automatic detection of hate-speech in Spanish.

Table 1 Features most commonly used in recent works on hate-speech

References	Language	LF	WE	SE	BERT-based
Romin et al. [50]	Bengali	–	✓	–	–
Plaza-del-Arco et al. [3]	Spanish	✓	✓	✓	✓
García-Díaz et al. [18]	Spanish	✓	✓	✓	–
Pamungkas et al. [41]	English, Italian, and Spanish	✓	✓	–	✓
Capozzi et al. [7]	Italian	✓	✓	–	✓
Huang et al. [26]	English, Italian, Polish, Portuguese, and Spanish	✓	✓	–	✓
Plaza-del-Arco et al. [47]	Spanish	✓	✓	–	–
Sigurbergsson and Derczynski [53]	Danish	✓	✓	–	–
Gómez et al. [21]	English	–	✓	–	–
Kapil and Ekbal [29]	English	✓	✓	✓	–
Çöltekin [9]	Turkish	✓	–	–	–

The majority of the studies found in Spanish are related to the participation in the shared-tasks AMI 2018 [6,14] and HatEval 2019 [4]. Regarding the techniques employed for hate-speech detection, some approaches are still based on traditional techniques, such as Support Vector Machines (SVM) [45,56], as well as traditional feature sets, such as n-grams and TF–IDF. In relation to neural networks models, Long Short-Term Memory (LSTMs) and Convolutional Neural Network (CNNs) are the most popular architectures employed by the teams [13,58]. As commented in [3], only a few of the participants evaluate the reliability of modern approaches based on transformers [20].

There are, however, some works out of these shared-tasks, such as those described in [3,18,44]. García-Díaz et al. [18] focus on misogyny identification and compile a dataset regarding three sub-types of misogyny in Spanish and evaluate a set of linguistic features and sentence embeddings. HaterNet [44] is a another recent work that evaluates hate-speech detection in Spanish. We describe in detail these datasets as well as the techniques employed in "Datasets" because they are the datasets used for answering the research questions proposed in this work. Last, the most recent work we are aware of is [3], in which the authors evaluate non-contextual and contextual embeddings including multilingual and monolingual pre-trained language models such as mBERT, XLM and BETO [8], over HaterNet [44] and HatEval [4] datasets. They conclude that BETO outperforms mBERT and XLM, pointing out that it is necessary to train a model on Spanish, since the system is capable of modulating more accurately the vocabulary. However, it is worth noting that the precision over the hate-speech label was higher with a simple logistic regression and TF–IDF on HaterNet dataset, whereas it was higher with the pre-trained word embeddings and a CNN on HatEval dataset.

Finally, we would like to point out that our work is related to that presented in [3] because of its relationship with the identification of hate-speech in Spanish and the evaluation of state-of-the-art transformers. However, our work differs from it in the following aspects. First, we analyze different feature sets separately, based on linguistic features and transformers, and take into account knowledge integration and ensemble learning strategies to build more robust solutions. Second, we analyze the reliability of the features employed to gain insights on whether these features are common in hate-speech categories. Third, we evaluate novel Spanish BERT models such as Spanish RoBERTa and BERTIN.

Table 1 summarizes the features most commonly used in the most recent works on hate-speech. We have classified them in linguistic features or n-grams (LF), pretrained word embeddings (WE), sentence embeddings (SE), and BERT-based models (BERT based). We can observe that the majority of works employ word embeddings and that BERT based features are being explored in the latest works as they have outperformed results in several tasks regarding NLP.

Datasets

This section describes the Spanish hate-speech datasets involved in this study. They focus on three topics: misogyny, xenophobia and hate in general. Table 2 summarizes the statistics of the datasets which are the (1) Spanish MisoCorpus 2020, (2) the AMI 2018, (3) HaterNet, and (4) the Spanish split of the HatEval 2019 dataset.

The **Spanish MisoCorpus 2020**[1] [18] is divided into three splits: (1) VARS, considering violence towards women in politics and public media; (2) SELA, on the understanding of the differences in misogynistic messages in Spanish from Spain and Spanish from Latin America; and (3) DDSS, that

[1] https://collaborativehealth.inf.um.es/corpora/misogyny/misocorpus-spanish-2020.rar.

Table 2 Corpus statistics regarding size

Label	*Train*	*Development*	*Test*	Total
Spanish MisoCorpus 2020				
Non-misogyny	2797	948	945	4690
Misogyny	2237	730	733	3700
Total	5034	1678	1678	8390
AMI 2018				
Non-misogyny	1253	422	416	2074
Misogyny	1227	405	415	2064
Total	2480	827	831	4138
HaterNET				
Non-hateful	2667	875	891	3600
Hateful	933	325	309	1567
Total	3600	1200	1200	6000
HatEval 2019 (Spanish)				
Non-hateful	2643	278	939	3860
Hateful	1857	222	660	2739
Total	4500	500	1599	6599

contains general traits related to misogyny. For this work we consider the full dataset that contains 8390 tweets. This corpus is slightly imbalanced, with more tweets labeled as non-misogyny. It was manually annotated by three human annotators. It is worth mentioning that the experiments conducted in [18] were applied with a balanced dataset, in which the authors sub-sampled the *non-misogyny* class. However, we compile all the tweets in order to keep the imbalance which is, on the one hand, more realistic and, on the other hand, similar to that observed in the rest of the evaluated datasets. Moreover, the authors evaluated the reliability of their methods using tenfold cross-validation with traditional machine-learning. We consider, however, that for the correct comparison with the rest of the datasets is better to split the Spanish MisoCorpus 2020 into training, development, and testing. The best result achieved by the authors was an accuracy of 82.882% with a combination of linguistic features and sentence embeddings.

The second dataset is from the shared-task Automatic Misogyny Identification [14] (AMI 2018)[2], proposed in IberEval 2018. The dataset is multilingual, with 4138 tweets written in Spanish and English. It has two subtasks: a binary misogyny identification task and a twofold multi-classification of misogynistic behavior. The first subtask includes determining traits of misogyny stereotypes, dominance, derailing, sexual harassment, threats of violence and discredit. The second subtask aims to determining when the target of the misogynist commentary is a particular individual or a group. In the scope of our work, we focus on the binary classification problem. To solve this task, the participants of the shared-task submitted several proposals achieving the best result an accuracy of 81.4681% [42] with a SVM architecture that combined several features, including *lexicons of abusive words, with a special focus on sexist slurs and abusive words targeting women*.

The **HaterNet**[3] [44] dataset was compiled from Twitter. The authors started from an initial set of 2 million of tweets that were filtered automatically and manually tagged by four human annotators. HaterNet is the dataset that presents more imbalance, with 1567 documents annotated as hateful, and 3600 annotated as non-hateful. For the evaluation, the authors focused on the F1-score of the hateful class. During their research, the authors of HaterNet dataset proposed a combination of recurrent neural networks and multilayer perceptrons to combine embeddings, emojis, and other statistical features, achieving an area under the curve (AUC) of 0.828.

Last dataset is **HatEval 2019**[4] [4], provided in a shared-task in SemEval 2019. This dataset was released for evaluating the detection of hate-speech towards immigrants and women. According to the overview of the HatEval 2019 task "*most part of the training set of tweets against women has been derived from an earlier collection carried out in the context of two previous challenges on misogyny identification*". Those datasets are AMI and EVALITA 2018. This suggests that the HatEval 2019 concerning misogyny is highly biased to those datasets. Two subtasks were proposed in HatEval 2019: (1) hate-speech detection against immigrants and women, and (2) aggressive behavior and target classification, which tries to determine if the target is an individual or a group. The Spanish split of HatEval 2019 consists of 6599 tweets divided into training, validation, and testing. The best result achieved in the Spanish binary subtask was a macro-averaged F1-score of 73%. For the rest of the participants the average was 68.21% and the standard deviation 0.0521, suggesting that most of the results were competitive. Out of the participants, there was a tie in the first position between [45] and [56], both using SVMs but evaluating different feature sets including bag of words, linguistic features, and Part-of-Speech features among others.

Materials and methods

To answer the research questions of this study, we implement a system[5] based on linguistic features, transformers and dif-

[2] https://amiibereval2018.wordpress.com/important-dates/data/.

[3] https://zenodo.org/record/2592149#.YNBqJGj7SUl.

[4] https://competitions.codalab.org/competitions/19935.

[5] The source code and requirements are available in the following GitHub repository: https://github.com/Smolky/CAIS-HateSpeeechDetectionInSpanish-2021.

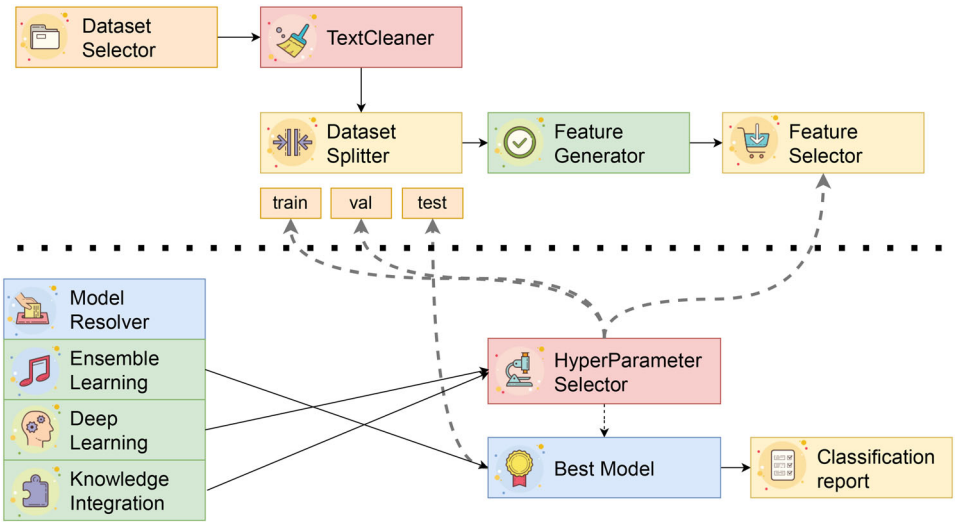

Fig. 1 System architecture

ferent integration mechanisms (knowledge integration with deep learning and ensemble learning). Figure 1 depicts the pipeline of our proposal. In a nutshell, it can be described as follows. First, the `DataResolver` module acts as input and is responsible of selecting one of the evaluated datasets. Second, the `TextCleaner` module cleans and pre-processes the texts to make them more uniform. Third, the `DatasetSplitter` module gets the training, validation, and testing splits. In addition, for training the models, the training split is used to fit the feature generation and feature selection processes, made by the `FeatureGenerator` and `FeatureSelector` modules, respectively. Fourth, the `ModelResolver` is the other input and is responsible to select one strategy for evaluate the datasets. Next, the `HyperParameterSelector` module is capable of evaluating different neural network architectures and hyperparameters to obtain the most suitable combination for each feature set and dataset. Note that the `ModelResolver` can select one these alternatives: (1) deep-learning, that handles the features separately; (2) knowledge integration, that uses two or more feature sets in the same neural network architecture, and (3) ensemble learning, that combines the predictions of the best models of each feature set. In the following sections, these modules are described in detail.

DatasetSelector

It loads the datasets and normalize them into a common format, standardizing the names of the labels, as some of these datasets use numbers as labels and others textual categories.

TextCleaner

It generates two clean versions of the texts. In the first clean version, urls, hashtags, mentions, emojis and punctuation symbols are removed. Digits are replaced with the token `[NUMBER]`, and elongations of certain letters are removed. This version is used for extracting some linguistic features related to Part-of-Speech, specially those related to proper names, such as surnames based on toponymics (Sevilla, Madrid), professions (Herrero[6], Zapatero[7]), or physical traits (Rubio[8]), which are common in Spanish. Next, it generates another clean version by lowercasing the first clean version. The second version is used to generate the word and sentence embeddings as it is explained in "FeatureGenerator". In addition, the original version of the text is kept to obtain features regarding the usage of uppercase or misspellings.

DatasetSplitter

It splits each dataset into training, validation, and test sets. We adopt the following strategy: (1) if the dataset has these three splits, we use them; (2) if the dataset contains only training and test sets, we split the training set to also generate a validation set by randomly selecting a 20% of the training split and keeping the test set as it is; (3) if the dataset is not partitioned, we perform a random split in a ratio of 60–20–20, keeping the splits balanced.

FeatureGenerator

This module generates the features used to represent the texts. We evaluate the following feature sets: linguistic features (LF), pretrained word embeddings (WE), sentence embeddings (SE), and fine-tuned BERT embeddings (BF). These features have been selected because they are the most com-

[6] In English: blacksmith.

[7] In English: shoemaker.

[8] In English: blonde.

monly used in existing works on hate-speech as can be seen in Table 1.

We extract the **linguistic features (LF)** from UMUTextStats [18,19] and extend them with fine-grain negation features from [27,28]. Concerning the negation features we get the list of negation cues appearing in each text (simple cues (e.g., "no"/ *not*), continuous cues (e.g., "en mi vida"/ *in my life*) and discontinuous cues (e.g., "ni...ni"/ *nor...nor*) and compute their total. Regarding the UMUTextStats, it is a text analysis tool focused on Spanish. It collects a total of 365 linguistic features that are organized with the following categories: phonetics, morphosyntax, correction and style, semantics, pragmatics, stylometry, lexical, psycho-linguistic processes, register, and social media jargon.

We extract three feature sets based on embeddings. First, we evaluate pre-trained **word embeddings (WE)** from word2vec [37], GloVe [43], and fastText [38]. Pre-trained word embeddings are a form of transfer learning in which the embeddings are learned from other general NLP tasks. They allow networks to converge faster as the representation of the words starts already clustered. Second, we obtain **fixed sentence embeddings from fastText (SE)** from its Spanish model [22], in which every document is represented as a fixed vector of length 300. Word and sentence embeddings from the pretrained models have the drawback that they do not take into account polysemy, so words have an unique representation regardless their context. Contextual word embeddings, on contrast, take into account the surrounding words in order to convey the embeddings. For this, for the third kind of embeddings, we evaluate different BERT models: BETO [8], the Spanish adaption of BERT [12], multilingual BERT (m-BERT) [46], Spanish RoBERTa [24], trained from National Library of Spain, and BERTIN.[9] BERT, and consequently BETO, use bidirectional transformers to learn contextual embeddings. Our approach to obtain these vectors is the following: we use the HuggingFace library to **fine-tune BETO with each dataset separately (BF)**. Then, we extract the sentence embeddings, as suggested in [48], applying a mean pooling on-top of the contextualized word embeddings, obtaining a fixed-vector representation of length of 768 for each document in the corpus. The advantage of this representation is that it is easier to combine with other feature sets but keeping the performance.

FeatureSelector

The `FeatureSelector` module is responsible for normalizing the features and selecting the best ones. Regarding normalization, we scale each LF individually in a range of 0 and 1. We apply this strategy as linguistic features are more heterogeneous, including some features that measures percentages and other raw counts. Next, we obtain the Mutual Information (MI) to observe the dependency of each feature with the label. With this information, we perform a feature selection by discarding those features that were ranked in the last quartile. We apply this process to LF and SE. We do not apply it to BF, as we observe that feature selection is not effective on those features.

ModelResolver

The `ModelResolver` selects the strategy used to train the models and the feature combinations. To address RQ1, we evaluate the feature sets separately. To answer RQ2, we evaluate two strategies. On the one hand, knowledge integration, which consists of combining several neural networks into a bigger one. Each feature set could work independently with a series of hidden layers, and then combine their inputs to output the final prediction or even to feed some more networks. On the other hand, we evaluate the combination of feature sets by means of ensembles. An ensemble is the combination of the outputs of two or more algorithms in order to make the final prediction. Ensembles are less sensitive to the training data, and usually provides better performance. Specifically, four strategies of ensemble learning are considered: (1) hard voting, which consists of selecting the label with a majority vote from the individual models; (2) highest probability, which consists of selecting the highest prediction probability among all the models; (3) average probability, which averages the probabilities of each model; and (4) logistic regression, which involves training a logistic regression classifier from the probabilities of the training splits.

HyperParameterSelector

As neural networks are highly configurable, we conduct an hyper-optimisation stage to evaluate which neural networks architectures are more suitable for hate-speech recognition. For SE or LF, we rely mostly on multilayer perceptrons. Word embeddings, however, allow to feed several kinds of neural networks architectures based on convolutional and recurrent networks. On the one hand, CNNs are more popular for solving computer vision tasks, but they could also be applied for conducting NLP tasks such as document classification [34]. The idea behind CNNs is the usage of filters based on pooling layers that are capable of generating high-order features. In this sense, CNNs exploit spatial time dimension of natural language, clustering joint words or expressions that may convey different meaning from the meaning of each word separately. Recurrent Neural Networks (RNNs), on the other hand, exploit the time dimension of the text, as they read the embeddings sequentially keeping information of past or even future words in case of bidirectional recurrent neural networks. Specifically, we evaluate two recurrent bidirec-

[9] https://huggingface.co/bertin-project/bertin-roberta-base-spanish.

tional networks: Long-Short Memory Units (BiLSTM), and Gated Recurrent Units (BiGRU).

We also evaluate the number of hidden layers and neurons. We distinguish among shallow neural networks, composed by one or two hidden layers and the same number of neurons per layer, and deep-learning architectures, between three and eight hidden layers. For deep-learning architectures we test different number of neurons disposed in several shapes: brick, funnel, long funnel, diamond and triangle. In addition, we evaluate a dropout to avoid overfitting, and different activation functions, including *ReLu*, *eLu*, *Sigmoid*, *Tanh*, and *Linear*. For the learning rate, we evaluate 10e−3 and 10e−4 with a scheduler using a time-based decay. Due to the size of the datasets and their slightly imbalance, we decided to evaluate the batch size of 32 and 64 but including larger batch sizes (128, 256, 512) for HaterNet, to ensure that every batch contains an enough number of hate-speech instances. All models were trained during 1000 epochs maximum with an early-stopping mechanism to avoid overfitting. For word embeddings we evaluate the usage of Spanish pre-trained word embeddings from word2vec, fastText, and GloVe or leaving the embeddings be learn from scratch. The hyperparameters explored are included in Table 3.

Table 3 Hyperparameter options for the neural networks architectures evaluated

Parameter	Ranges
Shared hyperparameters	
Batch size	[32, 64]
Dropout	[False, 0.1, 0.2, 0.3]
Neurons per layer	[4, 8, 16, 48, 64, 128, 256, 512, 1024]
Shallow neural networks	
Activation	[linear, relu, sigmoid, tanh]
Numbers of layers	[1, 2]
Shape	[brick]
Deep neural networks	
Activation	[sigmoid, tanh, selu, elu]
Numbers of layers	[3, 4, 5, 6, 7, 8]
Shape	[funnel, rhombus, longfunnel, brick, diamond, triangle]
Convolutional neural networks	
Activation	[sigmoid, tanh, selu, elu]
Numbers of layers	[1, 2]
Shape	brick
kernel size	[3, 5, 7]
Recurrent neural networks	
Bidirectional	[True, False]
Activation	[sigmoid, tanh, selu, elu]
Numbers of layers	[1, 2]
Shape	[brick]
kernel size	[3, 5, 7]

Classification report

This module reports the results of the best model with the test split. We evaluate the precision (see Eq. 1), recall (see Eq. 2), and F1-score (see Eq. 3) of the hate-speech label (*misogynous* for the Spanish MisoCorpus 2020 and AMI, and *hateful* for HaterNet and HatEval 2019), and of the non hate-speech label (*non_misogynous* for the Spanish MisoCorpus 2020 and AMI, and *non_hateful* for HaterNet and HatEval 2019), as well as the weighted versions of precision, recall, and F1-score for the overall comparison. In addition, to compare our approach with other systems, we use the accuracy (see Eq. 4) because is the metric used to rank the Spanish MisoCorpus-2020 and AMI 2018, and the macro F1-score, used to rank HatEval 2019.

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}), \tag{1}$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}), \tag{2}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{3}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{4}$$

where TP (True Positives) are those assessments where the system and human experts agree for a label assignment, FP (False Positives) are those labels assigned by the system that do not agree with the expert assignment, FN (False negatives) are those labels that the system failed to assign as they were given by the human expert, and TN (True Negatives) are those non-assigned labels that were also discarded by the expert.

Results and analysis

This section presents the results of the experiments conducted to answer the research questions formulated. In the following, each of them is addressed in a separate subsection.

RQ1. Which individual features are most effective for hate-speech detection?

The objective of this research question is to determine which feature set, in isolation, performs best in detecting hate-speech messages.

Discussion

Table 4 shows the results obtained for the individual features of each dataset. As expected, the fine-tune versions of BETO (BF) outperform the rest of features in a great extent.

Table 4 Performance of the individual features regarding hate-speech detection

	LF			SE			WE			BF		
	P	*R*	F1	*P*	*R*	F1	*P*	*R*	F1	*P*	*R*	F1
Spanish MisoCorpus 2020												
Misogynous	78.37	74.62	76.45	83.31	68.76	75.34	81.48	81.04	81.26	88.92	87.59	88.25
Non_misogynous	81.02	84.02	82.49	78.66	89.31	83.65	85.35	85.71	85.53	90.48	91.53	91.00
Weighted avg	79.86	79.92	79.85	80.69	80.33	80.02	83.66	83.67	83.67	89.80	89.81	**89.80**
AMI 2018												
Misogynous	74.15	78.80	76.40	78.63	71.81	75.06	80.66	70.36	75.16	82.19	83.37	82.78
Non_misogynous	77.44	72.60	74.94	74.12	80.53	77.19	73.77	83.17	78.19	83.17	81.97	82.57
Weighted avg	75.79	75.69	75.67	76.37	76.17	76.13	77.21	76.77	76.68	82.68	82.67	**82.67**
HaterNET												
Hateful	74.65	17.15	27.89	55.95	60.84	58.29	68.18	48.54	56.71	70.44	62.46	66.21
Non_hateful	77.33	97.98	86.44	86.00	83.39	84.67	83.78	92.14	87.76	87.47	90.91	89.16
Weighted avg	76.64	77.17	71.36	78.26	77.58	77.88	79.76	80.92	79.77	83.09	83.58	**83.25**
HatEval 2019												
Hateful	59.79	61.06	60.42	60.00	82.73	69.55	59.03	75.76	66.36	66.15	84.09	74.05
Non_hateful	72.22	71.14	71.67	83.45	61.24	70.64	78.72	63.05	70.02	86.18	69.76	77.10
Weighted avg	67.09	66.98	67.03	73.77	70.11	70.19	70.60	68.29	68.51	77.92	75.67	**75.84**

The results in bold highlight the higher scores

Therefore, we decide to evaluate other Spanish BERT models in order to compare the performance of different embeddings based on BERT. Table 5 presents the results of different pretrained contextual embeddings from BERT. Specifically, we evaluate BERTIN, BETO, the Spanish RoBERTa trained from National Library of Spain (BNE), and multilingual BERT (M-BERT). The best results are obtained by BETO, so we select this pretrained model to combine with the rest of the features to answer RQ2 and RQ3.

Next, we focus on evaluating the network complexity. Table 6 depicts the network architecture and the hyperparameters per feature set. It can be noticed that all neural network architectures are shallow neural networks with one or two hidden layers maximum with the same number of neurons in each layer. We can also observe that larger learning rates (10e−3) behave better than smaller ones (10e−4). However, there are no clear clues as to whether the learning rate is correlated with the feature set or the dataset. Regarding the activation function, ReLu is the one that appears mostly for achieving the best results, specially in AMI 2018. Tanh appears in complex networks with greater number of neurons. Finally, we can observe that the Spanish MisoCorpus 2020 and HaterNet share the learning rate and the activation function for LF, SE, and WE. In case of BF, however, the learning rate is the same but not the activation function.

Response

After analyzing the performance of the evaluated feature sets, we observe that the fine-tuned embeddings from BETO (BF) outperform the rest of feature sets. They achieved a significant increase of 4-5% regarding the second best feature set (WE). In relation to the reliability of LF, they achieve competitive results in all datasets except in HaterNet, in which they obtain limited results regarding the recall of the *hateful* category. Finally, with respect to the neural network architecture, we observe that shallow neural networks with few neurons and few hidden layers behave better than deep neural networks.

RQ2. How can features be integrated for more robust systems?

To answer this research question, we evaluate two different strategies to combine the feature sets. On the one hand, by combining them into the same neural network from a knowledge integration strategy (see "Knowledge integration strategy") and, on the other hand, by combining the results of the best model for each feature set using ensemble learning (see Sect. "Ensemble learning").

Knowledge integration strategy

To evaluate the knowledge integration strategy, we combine the features in the same neural network and perform the hyperparameter optimisation again. For this, we use the Keras API function to develop a multi-input neural network. Each feature set is used as input and connected to a dense layer. Then, all features are combined to produce the final prediction. Table 7 shows the results achieved by the linguistic

Table 5 Performance of BERT embeddings

	BERTIN			BETO			BNE			M-BERT		
	P	*R*	F1	*P*	*R*	F1	*P*	*R*	F1	*P*	*R*	F1
Spanish MisoCorpus 2020												
Misogynous	82.30	79.95	81.11	88.92	87.59	88.25	85.58	76.13	80.58	89.82	83.08	86.32
Non_misogynous	84.78	86.67	85.71	90.48	91.53	91.00	82.94	90.05	86.35	87.60	92.70	90.08
Weighted avg.	83.70	83.73	83.70	89.80	89.81	89.80	84.10	83.97	83.83	88.57	88.50	88.44
AMI 2018												
Misogynous	82.65	67.71	74.44	82.19	83.37	82.78	77.26	72.05	75.16	79.95	82.65	81.28
Non_misogynous	72.71	85.82	78.72	83.17	81.97	82.57	73.87	78.85	76.28	82.09	79.33	80.68
Weighted avg.	77.67	76.77	76.58	82.68	82.67	82.67	75.57	75.45	75.42	81.02	80.99	80.98
HaterNET												
Hateful	62.20	42.07	50.19	70.44	62.46	66.21	58.70	43.69	50.09	67.78	52.43	59.12
Non_hateful	81.94	91.13	86.29	87.47	90.91	89.16	82.06	89.34	85.55	84.70	91.36	87.90
Weighted avg	76.86	78.50	77.00	83.09	83.58	83.25	76.05	77.58	76.42	80.35	81.33	80.49
HatEval 2019												
Hateful	58.98	70.15	64.08	66.15	84.09	74.05	59.32	68.94	63.77	65.31	80.15	71.97
Non_hateful	75.80	65.71	70.39	75.67	75.67	75.67	75.36	66.77	70.81	83.40	70.07	76.16
Weighted avg.	68.86	67.54	67.79	77.92	75.67	75.84	68.74	67.67	67.90	75.93	74.23	74.43

Table 6 Feature hyperparameter results of the individual features regarding hate-speech detection

Feature-set	Architecture	Shape	# of layers	# of neurons	Dropout	lr	Activation
Spanish MisoCorpus 2020							
LF	MLP	Brick	2	31	0.3	10e−3	Sigmoid
SE	MLP	Brick	2	64	0.3	10e−3	ReLu
WE	MLP	Brick	2	8	0.3	10e−3	Linear
BF	MLP	Brick	2	48	0.3	10e−4	ReLu
AMI 2018							
LF	MLP	Brick	1	8	0.3	10e−4	ReLu
SE	MLP	Brick	1	8	0.3	10e−4	ReLu
WE	CNN	Brick	2	64	–	10e−3	ReLu
BF	MLP	Brick	2	2	0.2	10e−3	Sigmoid
HaterNet							
LF	MLP	Brick	1	8	–	10e−3	Sigmoid
SE	MLP	Brick	2	64	0.3	10e−3	ReLu
WE	MLP	Brick	2	48	0.2	10e−3	Tanh
BF	MLP	Brick	1	16	–	10e−4	Sigmoid
HatEval 2019							
LF	MLP	Brick	2	31	–	10e−3	Tanh
SE	MLP	Brick	2	27	–	10e−4	ReLu
WE	CNN	Brick	1	16	–	10e−3	ReLu
BF	MLP	Brick	1	37	–	10e−3	Tanh

Table 7 Performance of the features regarding hate-speech detection applying knowledge integration strategy

	LF, BF			SE, WE, BF			LF, SE, WE, BF		
	P	*R*	F1	*P*	*R*	F1	*P*	*R*	F1
Spanish MisoCorpus 2020									
Misogynous	89.50	88.40	88.95	89.42	88.81	89.12	90.99	86.77	88.83
Mon_misogynous	91.09	91.96	91.52	91.37	91.85	91.61	90.09	93.33	91.68
Weighted avg.	90.40	90.41	90.40	90.52	90.52	**90.52**	90.48	90.46	90.44
AMI 2018									
Misogynous	82.86	83.86	83.35	81.99	85.54	83.73	85.47	83.61	84.53
Non_misogynous	83.70	82.69	83.19	84.92	81.25	83.05	84.00	85.82	84.90
Weighted avg.	83.28	83.27	83.27	83.46	83.39	83.39	84.73	84.72	**84.72**
HaterNET									
Hateful	74.39	59.22	65.95	68.13	70.55	69.32	76.25	59.22	66.67
Non_hateful	86.79	92.93	89.76	89.66	88.55	89.10	86.88	93.60	90.11
Weighted avg.	83.60	84.25	83.62	84.11	83.92	84.01	84.14	84.75	**84.08**
HatEval 2019									
Hateful	69.50	79.39	74.12	65.44	83.79	73.49	65.10	83.94	73.33
Non_hateful	83.91	75.51	79.48	85.81	68.90	76.43	85.83	68.37	76.11
Weighted avg.	77.96	77.11	**77.27**	77.40	75.05	75.22	**77.27**	74.80	74.96

The results in bold highlight the higher scores

Table 8 Feature hyperparameter results of the knowledge integration strategy regarding hate-speech detection

Feature-set	Architecture	Shape	# of layers	# of neurons	Dropout	lr	Activation
Spanish MisoCorpus 2020							
LF, BF	MLP	Brick	2	512	0.2	10e−3	Linear
SE, WE, BF	MLP	Brick	1	512	0.3	10e−3	Linear
LF, SE, WE, BF	MLP	Brick	1	512	–	10e−3	Linear
AMI 2018							
LF, BF	MLP	Brick	2	2	0.3	10e−3	Tanh
SE, WE, BF	MLP	Brick	1	4	0.3	10e−3	Tanh
LF, SE, WE, BF	MLP	lfunnel	7	64	0.1	10e−3	eLu
HaterNet							
LF, BF	MLP	Brick	2	512	0.2	10e−3	Linear
SE, WE, BF	MLP	Brick	1	1024	False	10e−3	Linear
LF, SE, WE, BF	MLP	Brick	1	1024	False	10e−3	Linear
HatEval 2019							
LF, BF	MLP	Brick	1	256	0.2	10e−3	Tanh
SE, WE, BF	MLP	Brick	1	1024	0.2	10e−4	ReLu
LF, SE, WE, BF	MLP	Brick	2	64	0.1	10e−4	Linear

features and the best embedding (LF, BF), all the embeddings (SE, WE, and BF), and all features (LF, SE, WE, BF).

For the Spanish MisoCorpus 2020, the best result is achieved with the combination of the embeddings (SE, WE, and BF), with an 90.52% of weighted F1-score. However, the results of the combination of LF and BF as well as the combination of all features (LF, SE, WE, BF) are very similar: 90.40% and 90.44% of weighted F1-score, respectively. In addition, it can be seen that precision and recall become more stable with the feature combinations than with the feature sets separately (see Sect. "RQ1. Which individual features are most effective for hate-speech detection?"). In relation to AMI 2018, the best result is achieved by combining all the feature sets, with an 84.72% of weighted F1-score. In this case, the other combinations also get similar results: 83.27% for LF and BF, and 83.39% for SE, WE, and BF. HaterNet also obtain the best result with the combination of all features (84.08%). Regarding the last dataset, HatEval 2019, the best overall result is achieved with the combination of LF and BF, with a weighted F1-score of 77.27%.

Next, we focus on evaluating the network complexity. Table 8 lists the network architecture and the hyperparameters per feature set. The combination of different feature sets within the same network achieves better results with simpler neural networks, except for AMI 2018. Despite the fact that the number of hidden layers is similar, the number of neurons is quite superior, except for AMI 2018. For the learning rate, we can observe that larger learning rates (10e−3) behave better regardless the feature set combination for the Spanish MisoCorpus 2020, the AMI 2018, and the HaterNet datasets. However, small learning rates (10e−4) get better results in HatEval 2019 with the combination of all the embeddings (SE, WE, and BF) and all the features (LF, SE, WE, BF). Regarding the activation function, it can be observed that the Spanish MisoCorpus-2020 and HaterNet perform better with no activation function.

Ensemble learning

For evaluating the ensemble learning strategy, we exploit the best model per feature set developed for answering RQ1 (see Sect. "RQ1. Which individual features are most effective for hate-speech detection?"). Next, we generate a new prediction based on different approaches. First, using the hard voting strategy that consists of getting the majority vote of the models in the ensemble. Second, based on the prediction output of the model with the highest probability in the output layer. Third, by computing the average of the predictions of the last layer. Forth, by training a logistic regression classifier that learns to predict hate-speech or not based on the probabilities output by each model separately. We experiment with three combination sets: (1) linguistic features and the fine-tuned embeddings from BETO (LF, BE); (2) all the embeddings (SE, WE, and BF); and (3) linguistic features and all the embeddings (LF, SE, WE, BF). Due to the large number of datasets and strategies, we split the results in the following tables: (1) Table 9 for the Spanish MisoCorpus 2020, (2) Table 10 for the AMI 2018, (3) Table 11 for HaterNet, and (4) Table 12 for the HatEval 2019.

Regarding the Spanish MisoCorpus 2020 (see Table 9), the best overall result is obtained with the logistic regression strategy and all the feature sets, reaching a weighted F1-score of 90.14%. We can observe that this strategy also gets very good results with the other feature combinations evaluated. However, it is noteworthy that the highest probability strategy obtains the best precision regarding *misogyny* identification, but with a considerable sacrifice of the recall. This strategy considers one text as hate-speech when any of its classifiers outputs a probability higher than the 50%. Due to larger precision achieved, we conclude that this strategy is reliable for predicting a text as misogyny. However, the low recall indicates that there are many FN. Therefore, it would be necessary tuning the threshold of this strategy in order to adjust the compromise between precision and recall. On the other hand, the ensemble based on averaging probabilities performs slightly worse than the ensemble based on logistic regression, outperforming only when combining LF and BF. Finally, the ensemble learning based on the hard voting strategy penalizes the models with LF, getting the ensemble of SE, WE, BF better results. When comparing those results to the ones achieved with the knowledge integration strategy (see Table 7), we can observe that the knowledge integration strategy improves the results of the best ensemble learning strategy (logistic regression) with all the combinations: (1) 90.40% vs 89.74 with LF and BF, (2) 90.52% vs 89.96% with SE, WE, and BF; and (3) 90.44% vs 90.14% when combining all the features. However, the network architecture is complex.

Concerning AMI 2018 (see Table 10), the best ensemble learning result is provided by the strategy based on averaging the probabilities of LF and BF, with a weighted F1-score of 83.30%. As we observed when analyzing the Spanish MisoCorpus 2020 (see Table 9), the precision of the *misogyny* label with the ensemble based on the highest probability is high, with 92.90% by combining all feature sets, but with a great recall loss (53.50%). In the same line with the Spanish MisoCorpus 2020, the highest probability strategy of LF and BF achieves a reliable precision (87.10%), but with a slight drop in recall (69.90%). However, on the contrary, those ensembles based on averaging probabilities get better results than the logistic regression strategy.

As for HaterNet (see Table 11), the best result corresponds to the logistic regression strategy combining the features based on embeddings (SE, WE, and BF). This model achieves a weighted F1-score of 84.34%. The same strategy, but combining all the features (LF, SE, WE, and BF) obtains slightly worse weighted F1-score (84.25%), and 83.16% when combining LF and BF. When comparing the logistic regression strategy with the average probabilities strategy, we can observe that the weighted F1-scores are similar, but there are important differences among the precision and recall values of the *hateful* class. These differences were not observed in the Spanish MisoCorpus 2020 (see Table 9), the AMI 2018 datasets (see Table 10) nor the HatEval datasets (see Table 12). Regarding the strategy based on the highest probability, it can be observed that the combination of all feature sets (LF, SE, WE, and BF) achieves a perfect precision of the 12.60% identified instances. As we observed in the other datasets, the highest probability strategy using LF and BF achieves high precision (87.10%) but limited recall (69.90%). With respect to the hard voting strategy, it reaches lower results regarding the *hateful* label, but similar precision and recall in all the ensembles with LF (67.30% precision, 65.40% recall for LF and BF, and 69.66% precision, 60.19% recall when LF is combined with the rest of features).

Table 9 Performance of the features regarding hate-speech detection applying an ensemble learning strategy over the Spanish MisoCorpus 2020

Strategy	LF, BF			SE, WE, BF			LF, SE, WE, BF		
	P	*R*	F1	*P*	*R*	F1	*P*	*R*	F1
Spanish MisoCorpus 2020									
Hard voting									
Misogynous	77.70	93.50	84.80	89.15	82.95	85.94	84.49	88.40	86.40
Nonmisogynous	94.40	79.20	85.90	87.45	92.17	89.75	90.67	87.41	89.01
Weighted avg	86.80	85.40	85.40	88.19	88.14	88.08	87.97	87.84	87.87
High prob.									
Misogynous	93.70	68.80	79.30	94.20	59.80	73.10	95.50	52.50	67.80
Nonmisogynous	79.90	96.40	87.40	75.70	97.10	85.10	72.70	98.10	83.50
Weighted avg	85.90	84.30	83.90	83.80	80.80	79.90	82.70	78.20	76.60
Avg. prob.									
Misogynous	88.80	87.60	88.30	89.80	85.00	87.30	89.10	85.10	87.10
Nonmisogynous	90.70	91.40	91.00	88.80	92.50	90.60	88.90	92.00	90.40
Weighted avg	89.90	89.90	89.90	89.20	89.20	89.20	89.00	89.00	89.00
Log. regression									
Misogynous	88.80	87.59	88.19	90.42	86.22	88.27	90.34	86.77	88.52
Nonmisogynous	90.47	91.43	90.95	89.68	92.91	91.27	90.04	92.80	91.40
Weighted avg.	89.74	89.75	89.74	90.00	89.99	89.96	90.17	90.17	90.14

Table 10 Performance of the features regarding hate-speech detection applying an ensemble learning strategy over the AMI 2018 dataset

Strategy	LF, BF			SE, WE, BF			LF, SE, WE, BF		
	P	*R*	F1	*P*	*R*	F1	*P*	*R*	F1
AMI 2018									
Hard voting									
Misogynous	72.40	92.30	81.10	83.03	76.63	79.70	79.24	85.54	82.27
Non_misogynous	89.40	64.90	75.20	78.35	84.38	81.25	84.33	77.64	80.85
Weighted avg.	80.90	78.60	78.20	80.69	80.51	80.48	81.79	81.59	81.56
High prob.									
Misogynous	87.10	69.90	77.50	90.70	56.60	69.70	92.90	53.50	67.90
Non_misogynous	74.90	89.70	81.60	68.50	94.20	79.40	67.40	95.90	79.20
Weighted avg.	81.00	79.80	79.60	79.60	75.50	74.50	80.10	74.70	73.50
Avg. prob.									
Misogynous	82.40	84.60	83.50	84.60	78.30	81.40	84.60	78.10	81.20
Non_misogynous	84.20	82.00	83.10	79.90	85.80	82.70	79.70	85.80	82.60
Weighted avg.	83.30	83.30	83.30	82.20	82.10	82.20	82.10	91.90	81.90
Log. regression									
Misogynous	81.69	83.86	82.76	80.88	75.42	78.06	80.93	75.66	78.21
Non_misogynous	83.46	83.46	82.34	77.03	82.21	79.54	77.20	82.21	79.63
Weighted avg.	82.58	82.55	82.55	78.95	78.82	78.80	79.06	78.94	78.92

With respect to HatEval (see Table 12), the best overall result is reached with the logistic regression strategy of LF and BF, with a weighted F1-score of 76.66%. Regarding the ensemble learning based on the hard voting strategy, we can observe that the precision of identifying the *hateful* label is limited when comparing with the rest of the ensemble strategies. The highest probability strategy gets superior precision than the rest of the strategies in a similar way as we have observed for the rest of the datasets. However, we can see that the balance between precision and recall is not as high as we noticed in the Spanish MisoCorpus 2020, AMI and HaterNet. As we are able to discern which HatEval tweets focus on women and which on immigrants, we analyzed the results separately. We observed that the subset of HatEval 2019 focused on misogyny gets higher precision with the hate-speech label but limited recall, regardless of the

Table 11 Performance of the features regarding hate-speech detection applying an ensemble learning strategy over the HaterNet dataset

Strategy	LF, BF			SE, WE, BF			LF, SE, WE, BF		
	P	*R*	F1	*P*	*R*	F1	*P*	*R*	F1
HaterNet									
Hard voting									
Hateful	67.30	65.40	66.30	70.43	58.58	63.96	69.66	60.19	64.58
Non_hateful	88.10	89.00	88.60	86.43	91.47	88.89	86.82	90.91	88.82
Weighted avg.	82.80	82.90	82.80	82.31	83.00	82.46	82.40	83.00	82.58
High prob.									
Hateful	87.10	69.90	77.50	83.50	32.70	47.00	100.00	12.60	22.40
Non_hateful	74.90	89.70	81.60	80.70	97.80	88.40	76.70	100.00	86.80
Weighted avg.	79.80	79.80	79.60	81.40	81.00	77.80	82.70	77.50	70.30
Avg. prob.									
Hateful	82.40	84.60	83.50	73.50	56.60	64.00	74.90	54.00	62.80
Non_hateful	84.20	82.00	83.10	86.10	92.90	89.40	85.50	93.70	89.40
Weighted avg	83.30	83.30	83.30	82.80	83.60	82.80	82.70	83.50	82.50
Log. regression									
Hateful	65.39	71.52	68.32	68.75	71.20	69.95	68.65	70.87	69.75
Non_hateful	89.79	86.87	88.31	89.89	88.78	89.33	89.78	88.78	89.28
Weighted avg	83.51	82.92	83.16	84.44	84.25	84.34	84.34	84.17	84.25

ensemble strategy. On the other hand, the analysis of the split focused on migrants suggests the opposite, lower precision but higher recall. However, as the subset of HatEval 2019 towards women is biased with AMI 2018, a more thorough analysis of these differences would be necessary.

Response

For the combination of the features, we evaluated two strategies, one consisting of knowledge integration and the other of ensemble learning with multiple criteria. We realized that the results obtained with knowledge integration are, in general, superior to those achieved with ensemble learning, although there is not a great difference. However, we observed a higher complexity on the neural networks that requires more neurons than those of the best models of each feature set independently.

Concerning the ensemble learning study, the highest probability strategy achieves the best precision over the *misogynous* and *hateful* classes in all the datasets. However, this comes at a cost with respect to recall. We observe this specially with the HaterNet dataset, in which we obtained a perfect precision but a recall of 12.60%. For systems in which the precision is more important than recall, we recommend to focus on the highest probability strategy but selecting fewer feature sets as we observe better F1-score. In general, we can say that the strategies that provide competitive results regardless the datasets are knowledge integration and ensemble learning based on logistic regression using LF and BF as features.

RQ3. Is it possible to characterize the language of the different hate-speech types by means of explainable linguistic features?

For addressing this research question, we obtain the mutual information per linguistic feature of the different categories. In order to observe how linguistic features from different categories contribute to the identification of hate-speech, we rank those features and we organize them in groups of 5 according to the category. Figures 2, 3, 4, and 5 represent this classification for the Spanish MisoCorpus 2020, AMI 2018, HaterNet, and HatEval 2019 datasets respectively. Note that there are some categories, such as semantics, in which there are fewer than five categories.

Discussion

Regarding the **Spanish MisoCorpus 2020** (see Fig. 2) we can observe that the categories related to register are the most discriminatory. Register includes the usage of strong offensive speech, swear, colloquialisms and, to a lesser extent, non-fluent speech. Correction and style is another relevant category, highlighting features related to orthographic errors and misspelled words. Concerning morphosyntax, those features related to grammatical feminine words, nouns, prepositions, and suffixes are strong discriminatory features. Stylometry and social media are also effective regarding misogyny identification, including mentions or replies than include female name accounts. The usage of hashtags and social media jargon also stand out. As for stylometry, we

Table 12 Performance of the features regarding hate-speech detection applying an ensemble learning strategy over the HatEval 2019 dataset

Strategy	LF, BF			SE, WE, BF			LF, SE, WE, BF		
	P	*R*	F1	*P*	*R*	F1	*P*	*R*	F1
HatEval 2019									
Hard voting									
Hateful	58.30	89.70	70.60	63.88	85.88	73.22	61.05	88.33	72.20
Non_hateful	88.30	54.80	67.70	86.82	65.92	74.94	88.04	60.38	71.64
Weighted avg.	75.90	69.20	68.90	74.11	74.11	74.23	76.90	71.92	71.87
High prob.									
Hateful	73.60	55.50	63.30	71.90	62.40	66.80	77.30	45.50	57.30
Non_hateful	73.30	86.00	79.20	75.80	82.90	79.20	70.30	90.60	79.20
Weighted avg	73.50	73.40	73.40	74.20	74.40	74.10	73.20	72.00	70.10
Avg. prob.									
Hateful	67.90	80.60	73.70	63.60	84.70	72.60	65.70	82.60	73.20
Non_hateful	84.30	73.30	78.40	86.00	65.90	74.60	85.10	69.80	76.70
Weighted avg.	77.60	76.30	76.50	76.70	73.70	73.80	77.10	75.00	75.20
Log. regression									
Hateful	67.32	83.64	74.60	60.48	76.52	67.56	60.94	76.82	67.96
Non_hateful	86.14	71.46	78.11	79.71	64.86	71.52	80.05	65.39	71.98
Weighted avg.	78.37	76.49	76.66	71.77	69.67	69.89	72.16	70.11	70.32

can observe that features related to readability, quotations, and the length of the text can contribute in some extent to the identification of misogynistic messages. With respect to pragmatics, discourse markers used to argue, structure, or add information are relevant features. It also appears as relevant the usage of similes in figurative language and the usage of courtesy forms to introduce oneself in the conversation. The usage of negations appears both in misogynous and non misogynous documents, being the most relevant *ni...ni*, *no...no*, *nunca...nadie*, *sin...ni*, and *casi nadie*.

Next, we analyze the correlation between the linguistic features with the **AMI 2018 dataset** (see Fig. 3). The analysis reveals that linguistic features within register category and, specifically, related to offensive speech are the strongest discriminatory features. AMI and the Spanish MisoCorpus 2020, which both focus on misogyny identification, share this feature. Correction and style, however, is less relevant in AMI 2018 than in Spanish MisoCorpus 2020, as in AMI 2018 there are differences in the number of misspelled words among misogynous and non misogynous classes but we observe no differences in orthographic errors. Regarding morphosyntax, we can observe that verbs in subjunctive simple or in singular are discriminatory features as well as words in masculine and nominal suffixes. Lexical is another relevant feature but also differs from Spanish MisoCorpus 2020. In AMI 2018, the most relevant topics are related to animals, female and male persons and groups, and topics related to sex and risk. In the Spanish MisoCorpus 2020, however, the relevant topics are related with locations, organizations and also with analytic thinking and tentativeness. Only sex topics appear in both datasets as one of the most discriminating feature. This finding suggests that the context in which the tweets were collected can have a relevant role. In AMI, the documents were compiled using the following strategies: (1) using offensive representative words, (2) observing accounts from potential victims, and from (3) people who explicitly declared their hate against women. The Spanish MisoCorpus-2020 shares two of the three strategies mentioned, not taking into account misogynist accounts but taking more attention to certain events, like the arrival of Greta Thunberg in Madrid at the UN Climate Change Conference, or a case of rape of a minor that occurred in Spain related to a local soccer team. Focusing on those events may force the relevance of lexical features related to locations and organizations. It is surprising, however, that in the Spanish MisoCorpus-2020, animals is not a relevant feature for misogyny identification (in Spanish, the name of some female animals are usually misogynistic insults). In the same line, the usage of male and female groups of persons are relevant features in AMI but not in the MisoCorpus. This fact suggests that those terms can appear in misogynous and non-misogynous texts, so their are not good indicators of misogynous content. Another relevant difference of AMI 2018 with the Spanish MisoCorpus 2020 is social media, that have major impact in the Spanish MisoCorpus 2020. With respect to pragmatics, mutual information in AMI 2018 suggests that figurative language plays an important role to discern among misogynist messages from the usage of metaphors and understatements.

Similar to the Spanish MisoCorpus 2020 and AMI, we can observe from **HaterNet** (see Fig. 4) that offensive

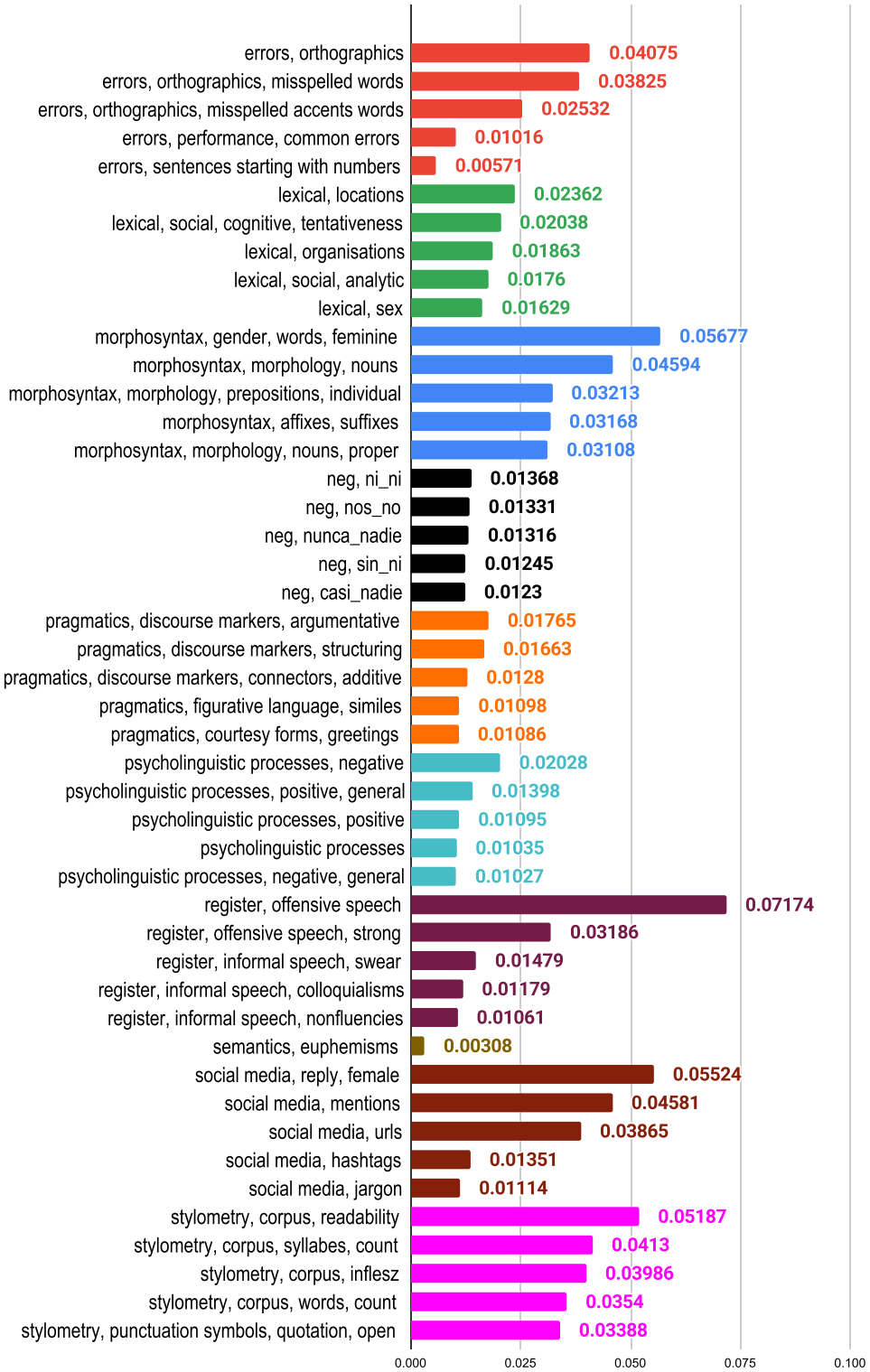

Fig. 2 Mutual information of the five-ranked features per category from Spanish MisoCorpus 2020

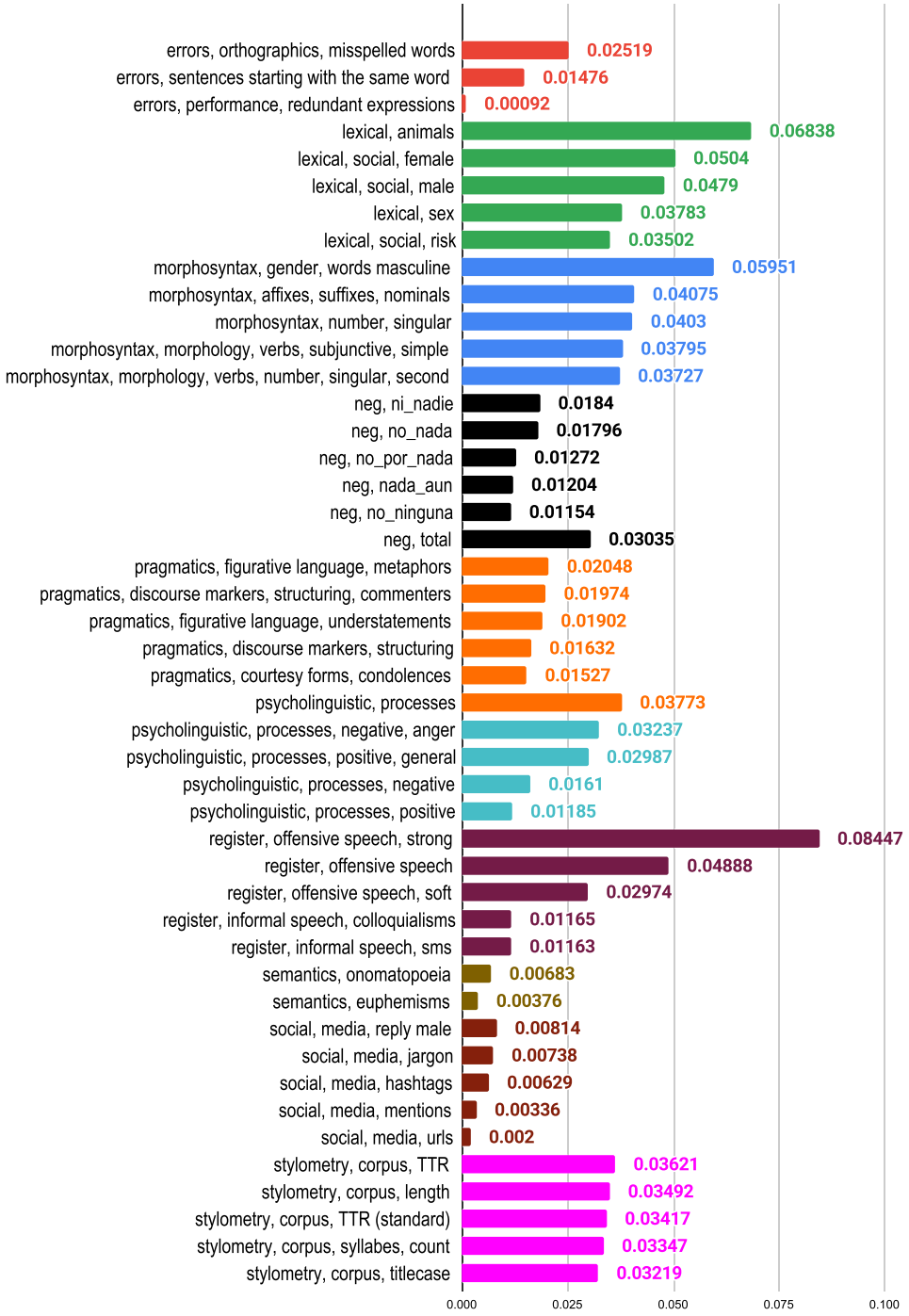

Fig. 3 Mutual information of the five-ranked features per category from AMI 2018

speech (register) is the most discriminatory feature. However, in HaterNet the presence of swear, SMS language and cultism appear as relevant features. With regard to the rest of categories all features behave similar. We note common performance errors for the correction and style category, and topics related to clothes and body for the lexical category. Intransitive verbs, as well as verbs in indicative (simple or compound), impersonal pronouns, and articles are also relevant features within the morphosyntactic category. Regarding pragmatics, the discourse markers related to reformulate and argument, as well as figurative language related to similes and idioms, are relevant for hate-speech detection. In addition, we can observe that social media jargon and the usage of hyperlinks can be useful for this dataset.

Concerning **HatEval 2019** (see Fig. 5), we can confirm the importance of offensive language (register), as this feature shows similar behavior in all datasets. It is worth mentioning that this analysis is biased because some of the documents

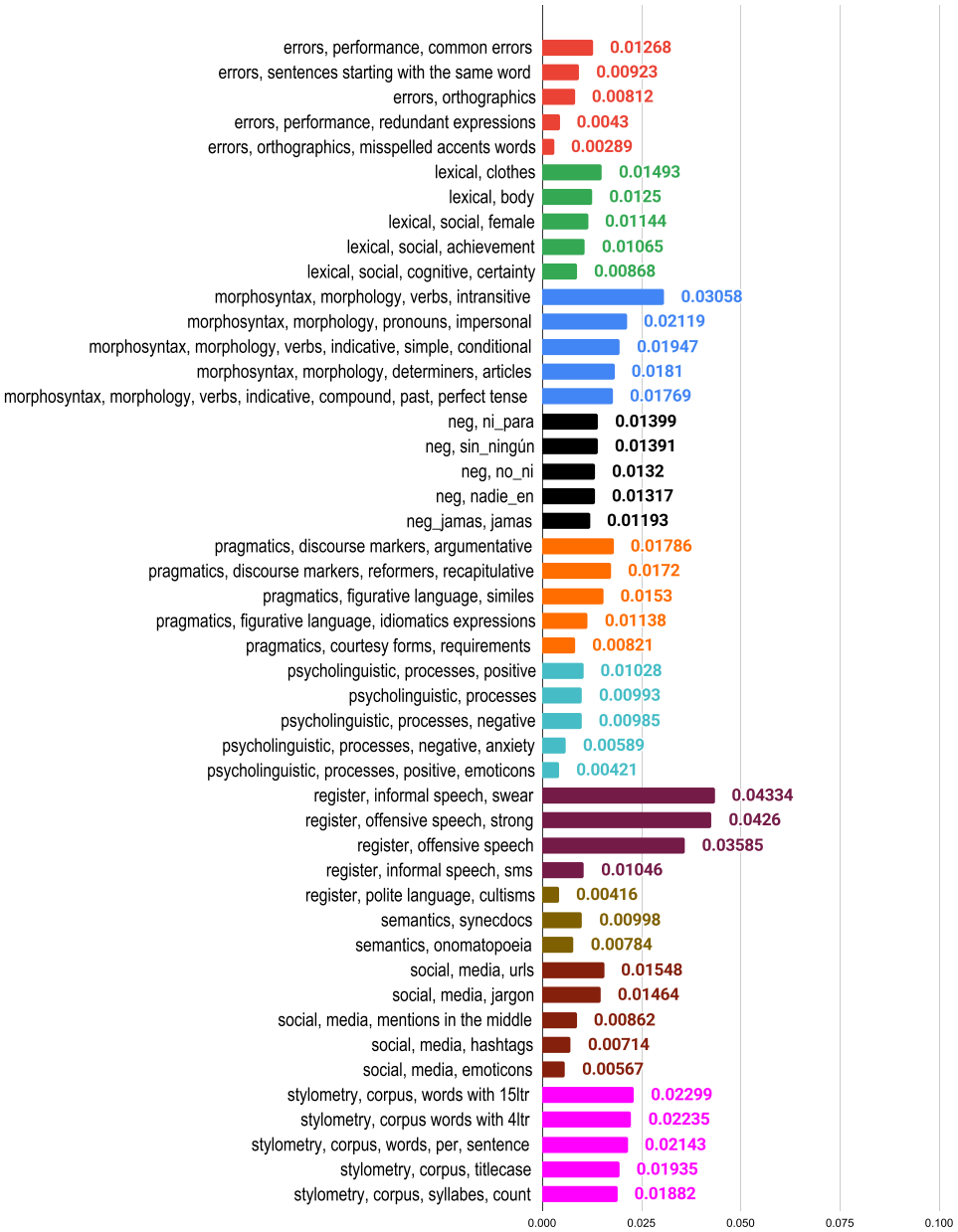

Fig. 4 Mutual information of the five-ranked features per category from HaterNet

from the misogyny split in HatEval 2019 also appear in AMI 2018. Consequently, we rank the linguistic features with information gain, but only with the subset of the HatEval 2019 dataset regarding to immigrants (not shown). The mutual information on the immigrants split also indicates that strong offensive speech is a relevant feature, but also the usage of colloquialisms and softer offensive language. Regarding lexical, the linguistic features are similar to the ones that appear for the AMI 2018 dataset including sex, common names referring to women or groups of women, inclusive language, and exploration. When we remove from our analysis the documents towards women and we focus only on hate-speech towards immigrants, we observe topics concerning sex, home, friendship, perceptual processes, and discrepancies. Some negation cues also appears as discriminatory features, including *nada más*, *jamás*, *casi nadie*, *ni...ni*, *tampoco...tan*. In fact, HatEval 2019 is the only dataset of those studied in which the total of negations appears as discriminatory feature, although with little impact regarding the identification of hate-speech. In terms of morphosyntax, we can observe that suffixes, including adjectivizers and nominals, as well as prepositions and copulative verbs, are discriminatory features. With respect to pragmatics, we can observe connectors, reformers, and words and expressions used to order the clauses. This fact suggests that a reflection, discussion and/or debate occurs when speaking towards women or immigrants within a conversation in hate-speech context. Regarding psycho-linguistic processes, those related

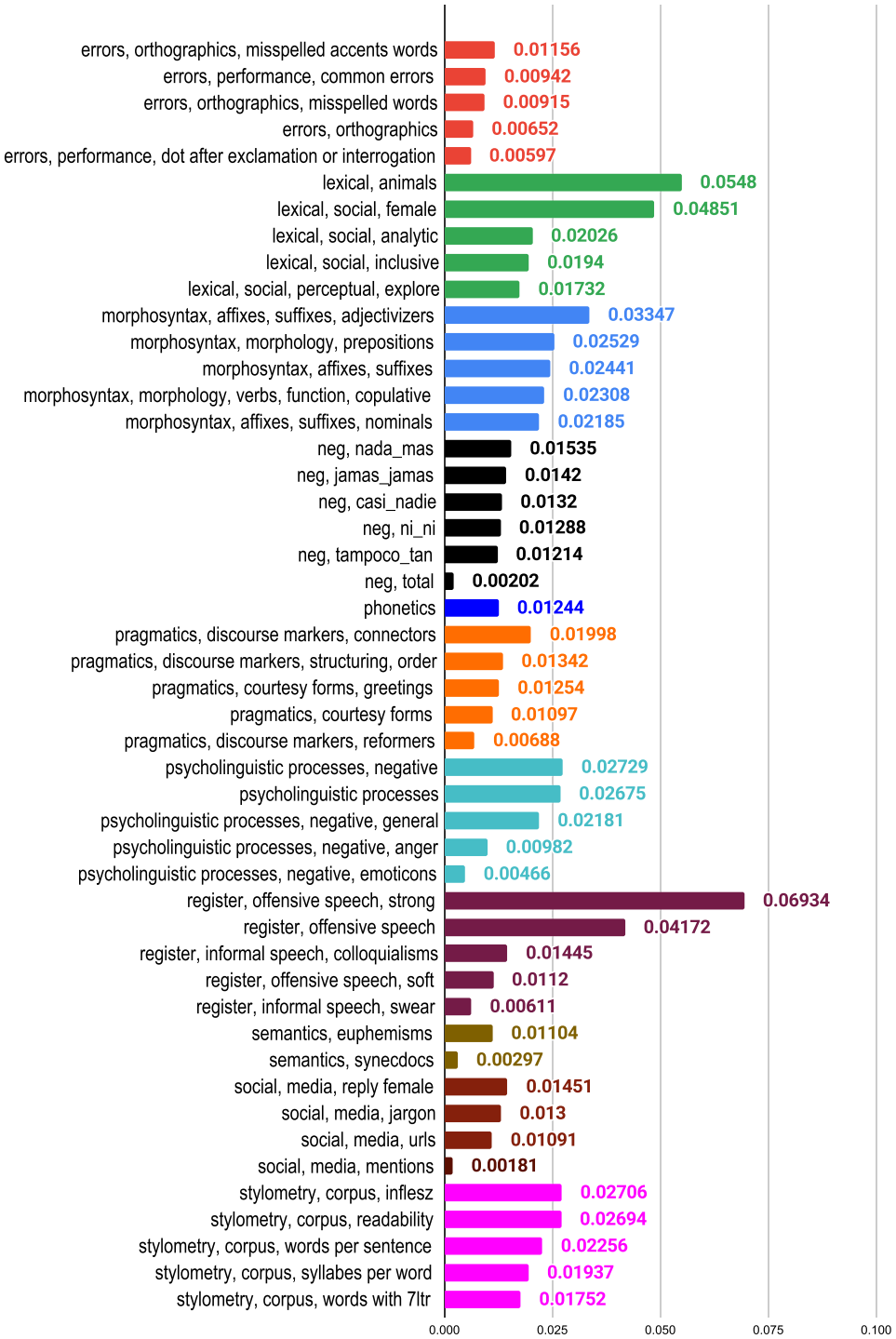

Fig. 5 Mutual information of the five-ranked features per category from HatEval 2019

with negative sentiments and especially anger are discriminatory features. We can also observe the presence of negative emojis. Concerning social media usage, we observe that reply to females is a discriminatory feature. We analyzed if this fact appears also in the subsets of the HatEval 2019 but we noticed that this feature is specially relevant to the misogyny subset. In fact, replying to males is also a relevant feature in this context.

Response

The analysis of the interpretability of the features leads to the following findings:

1. We observe common traits in all datasets regarding the register category, such as the features related to hard offensive speech. Moreover, swear and colloquialisms

also appear as discriminatory features, but in different degree.

2. From the misogyny identification, we note that the percentage of misspelled words is relevant for the Spanish MisoCorpus 2020 and AMI 2018 dataset. This finding does not appear in HaterNet, and is less so in HatEval 2019, which largely shares documents with AMI 2018.
3. Pragmatics and, specifically, discourse markers, appear frequently as discriminatory features. We observe that these features are more frequent in hate-speech or non hate-speech classes. We notice that argumentative markers are more common on non misogynous texts in the Spanish MisoCorpus 2020, but more common in misogynous texts in AMI 2018. Connectors used before to state a consequence are more common in non hateful documents as well as discourse markers used for structuring the text.
4. Linguistic features concerning the usage of social media show different behavior in the two corpus related to misogyny. The usage of mentions, hyperlinks, hashtags, and the usage of specific jargon appear as relevant features in the Spanish MisoCorpus 2020. However, social media features are no relevant in AMI 2018. HaterNet and HatEval indicate an intermediate value but not relevant.
5. Topics are not shared among the datasets focused on misogyny. We observe a strong presence of topics related to locations, organizations, and analytic thinking on the Spanish MisoCorpus 2020 whereas in AMI the topics are more related to animals (as the names of female animals species are common insults in Spanish), male and female social groups, and risk.
6. The usage of negations are not discriminatory features for hate-speech identification. We conduct a deep analysis of a total of 121 negations including simple, continuous, and discontinuous cues. However, the only dataset in which these features appears to be relevant is the HatEval 2019, with more statements with negations in the *hateful* class.

RQ4. Do our methods improve the results of the state-of-the-art?

To address this research question we compare our results with the best state-of-the-art results obtained for each particular dataset. Specifically, we compare our two best strategies, consisting of knowledge integration and ensemble learning based on logistic regression of LF and BF, with the best approaches of the state-of-the-art. These models are selected because achieved competitive results regardless the dataset. It is worth mentioning the limitations of this comparison. First, when comparing with HaterNet and the Spanish MisoCorpus 2020, the results were evaluated using ten-fold cross validation but in our approach we use the test set. Second, the results described in [3] regarding HaterNet use a training-test split, which is not the same split than ours nor the one used during the original experiment by the authors, since they did not release the splits. Third, not all shared-tasks and research focus on the same metrics, as those focusing on misogyny use accuracy, HaterNet compares with the F1-score of the *hateful* label, and HatEval 2019 with the Macro F1-score. Accordingly, we have include in Table 13 all the metrics and all the available results.

Discussion

When comparing the results for the **Spanish MisoCorpus 2020**, we can observe that our proposal, grounded on the usage of linguistic features and transformers, outperforms the accuracy achieved in [18], from 85.2% to 90.4% with knowledge integration and to 89.7% with the ensemble learning based on logistic regression. It should be noted that, to the best of our knowledge, this dataset has not been evaluated in other research works, so the conclusions are limited.

Regarding **AMI 2018**, the best result obtained during the shared-task was an accuracy of 81.4681% by [42], outperformed slightly in [18] with an 81.5217%. These results were achieved using Support Vector Machines and similar strategies for the features. The results reported by our systems outperform both results, but not significantly. Our proposal based on knowledge integration gets an accuracy of 83.3% and the ensemble learning based on logistic regression an 82.5%. Although our results are the best we are aware of, we consider that the novelty of the models employed based on transformers should have improved the state-of-the-art results even more.

Regarding **HaterNet**, we focus on F1-score of the *hateful* class. In the original experiment with this dataset [44], the authors achieved a F1-score for the *hateful* label of 61.1%. This result was outperformed by [3] with their proposal based in BETO, with a 65.8%. Our proposal based on linguistic features with a knowledge integration strategy outperforms slightly these results, achieving 65.9% of F1-score for the *hateful* label, but the results are superior applying the ensemble learning based on logistic regression, with a 68.3%.

Finally, for comparing **HatEval 2019** we rely on the macro F1-score. During the competition, the best results were achieved by [45,56], both with an accuracy of 73%. These results were outperformed by [18] and [3] with a macro F1-score of 75.4% and 75.5%, respectively. Similar as we observe in AMI 2018, the results of our proposal outperforms slightly these results: 76.8% of macro F1-score with knowledge integration of LF and BF, and 76.5% with ensemble learning based on logistic regression.

Table 13 Comparison of our approaches with the state-of-the-art for the Spanish MisoCorpus 2020, AMI 2018, HaterNET, and HatEval 2019, using accuracy (Acc), F1-Score of the hate-speech class (F1_HS), and Macro F1-score (M_F1)

Dataset	Approaches: Algorithms, features, and references	Acc	F1_HS	M_F1
Spanish MisoCorpus 2020	SVM, lf, awe [18]	85.2	–	–
	Knowledge integration (LF–BF)	**90.4**	**88.9**	**90.2**
	Ensemble learning (LF–BF, with log. regression)	89.7	88.2	89.6
AMI 2018	SVM, bag-of-words, lexicons [42]	81.5	–	–
	SVM, lf, awe [18]	81.5	–	–
	Knowledge integration (LF–BF)	**83.3**	**83.4**	**83.3**
	Ensemble learning (LF–BF, with log. regression)	82.5	82.8	82.5
HaterNet	LSTM and MLP [44]	–	61.1	–
	BETO [3]	–	65.8	77.2
	Knowledge integration (LF–BF)	**84.3**	65.9	77.9
	Ensemble learning (LF–BF, with log. regression)	82.9	**68.3**	**78.3**
HatEval 2019	SMO, n-grams, lf, PoS features [45,56]	–	–	73.0
	SMO, lf, awe [18]	–	–	75.4
	BETO [3]	–	**77.6**	75.5
	Knowledge integration (LF–BF)	**77.1**	76.8	**76.8**
	Ensemble learning (LF–BF, with log. regression)	76.5	74.6	**76.5**

The results in bold highlight the higher scores

Response

Taking into account the results provided by our methods and after comparing them with those of the state-of-the-art, we can say that our methods outperform those of the state-of-the-art.

Conclusion and further work

In this paper we have conducted a study of different datasets regarding hate-speech identification in Spanish, in order to determine which kind of individual features are most effective for hate-speech detection, how these features can be combined, if linguistic features could provide insights regarding the identification of hate-speech, and if the methods proposed here outperforms the state-of-the-art results.

As future lines of research, we plan two strategies, one related to further experimentation on the hate-speech topic and the other to an in-depth analysis of the system presented herein. On the one hand, in terms of experimentation, we will include the cross-validation strategy in our pipeline. Moreover, we will work on hate-speech related subtasks, such as determining the target and taking into account contextual features and media features, such as images or hyperlinked content. In addition, we will also try to focus on longer documents. On the other hand, the analysis strategy will be directed towards error analysis and the use of explainability tools. First, we will perform an error analysis to determine which cases are misclassified by each of the explored feature types and why, and whether the combination of them improves the classification. Finally, regarding explainability, we plan as future work to use tools like SHAP to see the contribution of each feature within the neural network. In this work, we have evaluated the reliability of using linguistic features to characterize hate-speech using model agnostic metrics, but these features are evaluated outside the neural network.

Acknowledgements This work was supported by project LaTe4PSP (PID2019-107652RB-I00) funded by MCIN/AEI/10.13039/501100011033, Project AlInFunds (PDC2021-121112-I00) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, Project LIVING-LANG (RTI2018-094653-B-C21) funded by MCIN/AEI/10.13039/501100011033 and by ERDF A way of making Europe, Project PID2020-116118GA-I00 supported by MICINN/AEI/10.13039/501100011033, Project PID2020-119478GB-I00 supported by MICINN/AEI/10.13039/501100011033, Banco Santander and University of Murcia through the industrial doctorate program, Fondo Social Europeo and Administration of the Junta de Andalucía (DOC_01073), Grant P20_00956 (PAIDI 2020) from the Andalusian Regional Government and grant 1380939 (FEDER Andalucía 2014-2020) from the Andalusian Regional Government.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Albadi N, Kurdi M, Mishra S (2018) Are they our brothers? Analysis and detection of religious hate speech in the Arabic twittersphere. In: 2018 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), pp 69–76. IEEE
2. Alfina I, Mulia R, Fanany M.I, Ekanata Y (2017) Hate speech detection in the Indonesian language: a dataset and preliminary study. In: 2017 international conference on advanced computer science and information systems (ICACSIS), pp 233–238. IEEE
3. Plaza-del Arco FM, Molina-González MD, Ureña-López LA, Martín-Valdivia MT (2021) Comparing pre-trained language models for Spanish hate speech detection. Expert Syst Appl 166:114120
4. Basile V, Bosco C, Fersini E, Debora N, Patti V, Pardo F.M.R, Rosso P, Sanguinetti M et al (2019) Semeval-2019 task 5: multilingual detection of hate speech against immigrants and women in twitter. In: 13th international workshop on semantic evaluation, pp 54–63. Association for Computational Linguistics
5. Bohra A, Vijay D, Singh V, Akhtar SS, Shrivastava M (2018) A dataset of Hindi-English code-mixed social media text for hate speech detection. In: Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media, pp 36–41
6. Bosco C, Felice D, Poletto F, Sanguinetti M, Maurizio T (2018) Overview of the evalita 2018 hate speech detection task. In: EVALITA 2018-sixth evaluation campaign of natural language processing and speech tools for Italian, vol. 2263, pp 1–9. CEUR
7. Capozzi AT, Lai M, Basile V, Poletto F, Sanguinetti M, Bosco C, Patti V, Ruffo G, Musto C, Polignano M et al (2020) "contro l'odio": a platform for detecting, monitoring and visualizing hate speech against immigrants in Italian social media. IJCoL Ital J Comput Linguist 6(6–1):77–97
8. Cañete J, Chaperon G, Fuentes R, Ho JH, Kang H, Pérez J (2020) Spanish pre-trained bert model and evaluation data. In: PML4DC at ICLR 2020
9. Çöltekin Ç (2020) A corpus of Turkish offensive language on social media. In: Proceedings of the 12th language resources and evaluation conference, pp 6174–6184
10. Corazza M, Menini S, Cabrio E, Tonelli S, Villata S (2020) A multilingual evaluation for online hate speech detection. ACM Trans Internet Technol (TOIT) 20(2):1–22
11. Davidson T, Warmsley D, Macy M, Weber I (2017) Automated hate speech detection and the problem of offensive language. In: Proceedings of the international AAAI conference on web and social media, vol 11
12. Devlin J, Chang M.W, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
13. Ding Y, Zhou X, Zhang X (2019) Ynu_dyx at semeval-2019 task 5: a stacked bigru model based on capsule network in detection of hate. In: Proceedings of the 13th international workshop on semantic evaluation, pp 535–539
14. Fersini E, Rosso P, Anzovino M (2018) Overview of the task on automatic misogyny identification at ibereval 2018. IberEval@ SEPLN vol 2150, pp 214–228
15. Fortuna P, Nunes S (2018) A survey on automatic detection of hate speech in text. ACM Comput Surv (CSUR) 51(4):1–30. https://doi.org/10.1145/3232676
16. Fortuna P, da Silva JR, Wanner L, Nunes S et al. (2019) A hierarchically-labeled Portuguese hate speech dataset. In: Proceedings of the third workshop on abusive language online, pp 94–104
17. Frenda S, Ghanem B, Montes-y Gómez M, Rosso P (2019) Online hate speech against women: automatic identification of misogyny and sexism on twitter. J Intell Fuzzy Syst 36(5):4743–4752
18. García-Díaz JA, Cánovas-García M, Palacios RC, Valencia-García R (2021) Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings. Future Gener Comput Syst 114:506–518. https://doi.org/10.1016/j.future.2020.08.032
19. García-Díaz JA, Cánovas-García M, Valencia-García R (2020) Ontology-driven aspect-based sentiment analysis classification: an infodemiological case study regarding infectious diseases in Latin America. Future Gener Comput Syst 112:641–657. https://doi.org/10.1016/j.future.2020.06.019
20. Gertner AS, Henderson J, Merkhofer E, Marsh A, Wellner B, Zarrella G (2019) Mitre at semeval-2019 task 5: transfer learning for multilingual hate speech detection. In: Proceedings of the 13th international workshop on semantic evaluation, pp 453–459
21. Gomez R, Gibert J, Gomez L, Karatzas D (2020) Exploring hate speech detection in multimodal publications. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 1470–1478
22. Grave E, Bojanowski P, Gupta P, Joulin A, Mikolov T (2018) Learning word vectors for 157 languages. In: Proceedings of the international conference on language resources and evaluation (LREC 2018)
23. Guillermo Carbonell BM, Michael Wojatzki BN (2016) Measuring the reliability of hate speech annotations: the case of the European refugee crisis. Bochumer Linguistische Arbeitsberichte, pp 6–9
24. Gutiérrez-Fandiño A, Armengol-Estapé J, Pàmies M, Llop-Palao J, Silveira-Ocampo J, Carrino C.P, Gonzalez-Agirre A, Armentano-Oller C, Rodriguez-Penagos C, Villegas M (2021) Spanish language models
25. Hinduja S, Patchin JW (2010) Bullying, cyberbullying, and suicide. Arch Suicide Res 14(3):206–221. https://doi.org/10.1080/13811118.2010.494133 (**(PMID: 20658375)**)
26. Huang X, Xing L, Dernoncourt F, Paul M (2020) Multilingual twitter corpus and baselines for evaluating demographic bias in hate speech recognition. In: Proceedings of the 12th language resources and evaluation conference, pp 1440–1448
27. Jiménez-Zafra SM, Morante R, Blanco E, Valdivia MTM, Lopez LAU (2020) Detecting negation cues and scopes in Spanish. In: Proceedings of the 12th language resources and evaluation conference, pp 6902–6911
28. Jiménez-Zafra SM, Taulé M, Martín-Valdivia MT, Urena-López LA, Martí MA (2018) Sfu review sp-neg: a Spanish corpus annotated with negation for sentiment analysis. A typology of negation patterns. Lang Resour Eval 52(2):533–569
29. Kapil P, Ekbal A (2020) A deep neural network based multi-task learning approach to hate speech detection. Knowl-Based Syst 210:106458
30. Kumari K, Singh J (2019) Ai ml nit patna at hasoc 2019: deep learning approach for identification of abusive content
31. Kumari K, Singh JP (2020) AI_ML_NIT_Patna @ TRAC - 2: deep learning approach for multi-lingual aggression identification. In: Proceedings of the second workshop on trolling, aggression and cyberbullying, pp 113–119. European Language Resources Asso-

ciation (ELRA), Marseille, France. https://aclanthology.org/2020.trac-1.18
32. Kumari K, Singh JP (2020) Ai_ml_nit_patna @hasoc 2020: Bert models for hate speech identification in indo-European languages. In: FIRE
33. Ljubešić N, Erjavec T, Fišer D (2018) Datasets of Slovene and Croatian moderated news comments. In: Proceedings of the 2nd workshop on abusive language online (ALW2), pp 124–131
34. Lopez MM, Kalita J (2017) Deep learning applied to NLP. CoRR arXiv:1703.03091
35. Mandl T, Modha S, Kumar MA, Chakravarthi BR (2020) Overview of the hasoc track at fire 2020: hate speech and offensive language identification in Tamil, Malayalam, Hindi, English and German. In: Forum for information retrieval evaluation, pp 29–32
36. Mandl T, Modha S, Majumder P, Patel D, Dave M, Mandlia C, Patel A (2019) Overview of the hasoc track at fire 2019: hate speech and offensive content identification in Indo-European languages. In: Proceedings of the 11th forum for information retrieval evaluation, pp 14–17
37. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781
38. Mikolov T, Grave E, Bojanowski P, Puhrsch C, Joulin A (2018) Advances in pre-training distributed word representations. In: Proceedings of the international conference on language resources and evaluation (LREC 2018)
39. Müller K, Schwarz C (2018) Fanning the flames of hate: Social media and hate crime. J Eur Econ Assoc
40. Ousidhoum ND, Lin Z, Zhang H, Song Y, Yeung DY (2019) Multilingual and multi-aspect hate speech analysis. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)
41. Pamungkas EW, Basile V, Patti V (2020) Misogyny detection in twitter: a multilingual and cross-domain study. Inf Process Manag 57(6):102360
42. Pamungkas EW, Cignarella AT, Basile V, Patti V et al. (2018) 14-exlab@ unito for ami at ibereval2018: exploiting lexical knowledge for detecting misogyny in english and spanish tweets. In: 3rd workshop on evaluation of human language technologies for Iberian languages, IberEval 2018, vol. 2150, pp. 234–241. CEUR-WS
43. Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543
44. Pereira-Kohatsu JC, Quijano-Sánchez L, Liberatore F, Camacho-Collados M (2019) Detecting and monitoring hate speech in twitter. Sensors 19(21):4654
45. Pérez JM, Luque FM (2019) Atalaya at semeval 2019 task 5: robust embeddings for tweet classification. In: Proceedings of the 13th international workshop on semantic evaluation, pp 64–69
46. Pires T, Schlinger E, Garrette D (2019) How multilingual is multilingual bert? arXiv preprint arXiv:1906.01502
47. Plaza-Del-Arco FM, Molina-González MD, Ureña López LA, Martín-Valdivia MT (2020) Detecting misogyny and xenophobia in Spanish tweets using language technologies. ACM Trans Internet Technol. https://doi.org/10.1145/3369869
48. Reimers N, Gurevych I (2019) Sentence-bert: sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084
49. Rodríguez A, Argueta C, Chen Y.L (2019) Automatic detection of hate speech on Facebook using sentiment and emotion analysis. In: 2019 international conference on artificial intelligence in information and communication (ICAIIC), pp 169–174. IEEE
50. Romim N, Ahmed M, Talukder H, Islam M.S (2021) Hate speech detection in the Bengali language: a dataset and its baseline evaluation. In: Proceedings of international joint conference on advances in computational intelligence. Springer, pp 457–468
51. Sap M, Card D, Gabriel S, Choi Y, Smith NA (2019) The risk of racial bias in hate speech detection. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 1668–1678
52. Schmidt A, Wiegand M (2017) A survey on hate speech detection using natural language processing. In: Proceedings of the fifth international workshop on natural language processing for social media, pp 1–10
53. Sigurbergsson GI, Derczynski L (2020) Offensive language and hate speech detection for Danish. In: Proceedings of The 12th language resources and evaluation conference, pp 3498–3508
54. Sun C, Qiu X, Xu Y, Huang X (2019) How to fine-tune bert for text classification? In: Sun M, Huang X, Ji H, Liu Z, Liu Y (eds) Chinese computational linguistics. Springer International Publishing, Cham, pp 194–206
55. Tulkens S, Hilte L, Lodewyckx E, Verhoeven B, Daelemans W (2016) A dictionary-based approach to racism detection in Dutch social media. In: Workshop programme, pp 11–17
56. Vega LEA, Reyes-Magaña JC, Gómez-Adorno H, Bel-Enguix G (2019) Mineriaunam at semeval-2019 task 5: detecting hate speech in twitter using multiple features in a combinatorial framework. In: Proceedings of the 13th international workshop on semantic evaluation, pp 447–452
57. Warner W, Hirschberg J (2012) Detecting hate speech on the world wide web. In: Proceedings of the second workshop on language in social media, pp 19–26
58. Winter K, Kern R (2019) Know-center at semeval-2019 task 5: multilingual hate speech detection on twitter using cnns. In: Proceedings of the 13th international workshop on semantic evaluation, pp 431–435
59. Zampieri M, Malmasi S, Nakov P, Rosenthal S, Farra N, Kumar R (2019) Predicting the type and target of offensive posts in social media. In: Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: human language technologies, vol 1 (Long and Short Papers), pp 1415–1420