Spanish hate-speech detection in football

Detección de odio en fútbol en español

Esteban Montesinos-Cánovas,[1] **Francisco García-Sánchez,**[1] **José Antonio García-Díaz,**[1] **Gema Alcaraz-Mármol,**[2] **Rafael Valencia-García**[1]
[1]Departamento de Informática y Sistemas, Facultad de Informática, Universidad de Murcia
[2]Departamento de Filología Moderna, Universidad de Castilla-La Mancha
{esteban.montesinosc, frgarcia, joseantonio.garcia8, valencia}@um.es
Gema.Alcaraz@uclm.es

Abstract: In the last few years, Natural Language Processing (NLP) tools have been successfully applied to a number of different tasks, including author profiling, negation detection or hate speech detection, to name but a few. For the identification of hate speech from text, pre-trained language models can be leveraged to build high-performing classifiers using a transfer learning approach. In this work, we train and evaluate state-of-the-art pre-trained classifiers based on Transformers. The explored models are fine-tuned using a hate speech corpus in Spanish that has been compiled as part of this research. The corpus contains a total of 7,483 football-related tweets that have been manually annotated under four categories: aggressive, racist, misogynist, and safe. A multi-label approach is used, allowing the same tweet to be labeled with more than one class. The best results, with a macro F1-score of 88.713%, have been obtained by a combination of the models using Knowledge Integration.
Keywords: Hate speech detection, Large Language Models, Linguistic features, Interpretability.

Resumen: En los últimos años, el Procesamiento del Lenguaje Natural (PLN) se ha aplicado con éxito a diversas tareas, como la elaboración de perfiles de autor, la detección de negaciones o la detección de discursos de odio. Para la identificación de odio a partir de texto, es posible explotar modelos del lenguaje preentrenados que permitan construir clasificadores de alto rendimiento utilizando un enfoque de aprendizaje por transferencia (en inglés, transfer learning). En este trabajo, se presentan los resultados de entrenar y evaluar clasificadores preentrenados de última generación basados en Transformers. Los modelos explorados se ajustan (en inglés, fine tune) utilizando un corpus en español sobre el discurso de odio en el fútbol que se ha compilado como parte de esta investigación. El corpus contiene un total de 7.483 tuits relacionados con el fútbol que han sido anotados manualmente bajo cuatro categorías: agresivo, racista, misógino y seguro. Se utilizó un enfoque multi-etiqueta, que permite etiquetar el mismo tuit con más de una clase. Los mejores resultados, con un macro F1-score del 88,713%, se han obtenido mediante una combinación de los modelos utilizando la estrategia de Knowledge Integration.
Palabras clave: Discurso de odio, Modelos del lenguaje, Características lingüísticas, Interpretabilidad.

1 Introduction

The Cambridge Dictionary defines *hate speech* as "public speech that expresses hate or encourages violence towards a person or group based on something such as race, religion, sex, or sexual orientation"[1]. The use of hateful rhetoric has been linked to real-world violence[2] and, with the advent of social media, hatred has a much higher spreading velocity (Mathew et al., 2019; Paz, Montero-Díaz, and Moreno-Delgado, 2020). Tackling hate speech has become a recognized global

[1]https://dictionary.cambridge.org/us/dictionary/english/hate-speech

[2]https://www.un.org/en/hate-speech/understanding-hate-speech/hate-speech-and-real-harm

priority for the United Nations[3]. The identification of hate content in different media constitutes one of the first key steps towards such objective. The automation of this task and the early detection of such content would help media managers to prevent its dissemination and for such harmful messages to reach the masses.

Textual hate speech detection systems usually rely on Natural Language Processing (NLP) techniques based on automatic classifiers trained with Large Language Models (LLMs) (Alkomah and Ma, 2022). Hate speech detectors are usually trained with a set of manually labeled data. The system learns from these examples, finding relationships between the expressions in the input texts and other discriminatory features. The process leads to creating a model capable of classifying new samples into the considered categories (e.g., hateful, non hateful). Hate speech identification is a hot topic, and the number of research works published in the last few years dealing with this problem is increasing (Oliveira and Azevedo, 2022; Bilal et al., 2022; Mehta and Passi, 2022; Chiril et al., 2022; Ali et al., 2022; Roy, Bhawal, and Subalalitha, 2022; Husain and Uzuner, 2022; Wullach, Adler, and Minkov, 2022; Plaza del Arco et al., 2021). However, it is far from solved (Arango, Pérez, and Poblete, 2022). First, most works are focused on texts written in English and only a minority are dedicated to other languages such as Arabic (Husain and Uzuner, 2022), Urdu (Bilal et al., 2022), Spanish (Plaza del Arco et al., 2021) or Dravidian languages (Roy, Bhawal, and Subalalitha, 2022). Therefore, one of the main difficulties when facing this challenge is to obtain a large enough corpus to train the model. Second, the classifiers built upon LLMs are hard to interpret, as these models do not provide interpretable features.

In this work we aim to facilitate social media content moderators their mission to avoid the spread of hatred in sports. With that purpose, we assess the suitability of several state-of-the-art LLMs to tackle hate speech detection from texts written in Spanish. In particular, we compare the performance of nine LLMs plus a model based on interpretable Linguistic Features (LFs). We evaluate Spanish models such as BETO (Cañete et al., 2020), MarIA (Gutiérrez-Fandiño et al., 2022), BERTIN (de la Rosa et al., 2022), DistilBETO (Cañete et al., 2022) and AlBETO (Cañete et al., 2022), and multilingual LLMs such as multilingual BERT (Devlin et al., 2019), XLM (Conneau et al., 2020), multilingual DeBERTa (He, Gao, and Chen, 2021) and TwHIN-BERT (Zhang et al., 2022). On the other hand, to investigate the particularities of hate speech in sports, which can have cultural and contextual components, we have compiled a multi-label corpus with 7,482 tweets written in Spanish and related to football (a.k.a., soccer), annotated as safe, aggressive, misogynist and racist. In football, large masses of people that support one team can generate controversy with other teams' fans, and the amount of content generated in social media around each match is massive. Therefore, it is quite common to find aggressive comments on Twitter associated to a match either towards the referee or some players. On contrast, some keywords concerning sports can introduce topic bias in generic datasets. In this sense, in (Poletto et al., 2021) the authors identify datasets in which words concerning football were discerning features to mark a document as hate-speech.

The rest of the manuscript is organized as follows. Section 2 provides background information concerning hate-speech detection in Spanish. A detailed description of the corpus and its compilation process is presented in Section 3. Then, in Section 4 the system architecture is described along with a complete overview of all its components. The evaluation of the feature sets in isolation or combined and the error analysis conducted for the evaluation of the overall system is shown in Section 5. Finally, conclusions and future work are put forward in Section 6.

2 State of the art

Pre-trained language models have become a popular solution for NLP due to their ability to learn contextualized representations of language (Min et al., 2021). They are a type of neural network that is trained on massive amounts of text data, often using unsupervised learning techniques. The goal is to learn general language representations that can be fine-tuned for specific NLP tasks such as sentiment analysis or hate speech detection (Omar et al., 2022). One

[3] https://www.un.org/en/hate-speech

of the most popular pre-trained language models is the Transformer architecture, originally introduced in (Vaswani et al., 2017). The Transformer model was trained on a large corpus of text using a technique called self-supervised learning, resulting in BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019), which outperformed previous state-of-the-art models on a wide range of NLP tasks.

LLMs have shown promising results for hate speech detection (Alkomah and Ma, 2022). These models have been pre-trained on large amounts of text data and can be fine-tuned on a smaller labeled dataset to perform specific NLP tasks, including hate speech detection. For this, a classifier is trained on top of the pre-trained model to predict whether a given piece of text is hateful or not based on the features extracted by the model. The ability of pre-trained models to capture context and nuances in language use can thus be leveraged by the hate speech identification system. Besides, LLMs have the ability to learn complex linguistic patterns, which can help improve the accuracy of hate speech detection.

Twitter characterizes hateful conduct as promoting "violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease"[4]. The number of studies on hate speech detection in Twitter and other social media is growing dramatically (Mansur, Omar, and Tiun, 2023). In (Mehta and Passi, 2022), the authors present an attempt to face hate speech detection using explainable artificial intelligence. Different deep learning models and variants of Transformers-based models are explored to extract useful interpretable results from two datasets, the Google Jigsaw dataset (comprised of user discussions from talk pages of English Wikipedia) and the HateXplain dataset (comprised of posts from Twitter and Gab). The performance in terms of macro F1-score of the best models is well over 90%.

In languages other than English, the scarcity of linguistic resources makes hate speech detection more challenging. The authors in (Mozafari, Farahbakhsh, and Crespi, 2022) propose a cross-lingual few-shot approach for hate speech and offensive language detection that is evaluated using 15 publicly available datasets in 8 languages. Instead of transfer learning, a meta learning approach following optimization-based and metric-based methods is used to train a model able to generalize quickly to new languages with few labeled data. Focusing on Spanish, different machine learning, deep learning and transfer learning-based hate speech detection models are compared in (Plaza del Arco et al., 2021). Pre-trained language models outperform previous approaches, with monolingual LLMs obtaining the best results. In this work, we extend the evaluation to the most up-to-date language models, both mono- and multilingual, and set the focus on a sector particularly hard hit by the scourge of hatred such as football. In previous works, we have successfully combined LFs with other classification models to deal with different NLP tasks such as author profiling (García-Díaz, Colomo-Palacios, and Valencia-García, 2022), misogyny identification (García-Díaz et al., 2021) or sentiment analysis (García-Díaz, García-Sánchez, and Valencia-García, 2023). In this work, we also incorporate LFs to enhance the interpretability of the results.

As far as our knowledge goes, few studies have considered hate-speech in football as an automatic document classification problem. In (Cleland, 2014), the authors conducted a discourse analysis from around 500 online texts concerning racism in football. They discovered that message boards allow racist messages to spread but that most of the racist posts were confronted by other supporters. In (Vasconcelos et al., 2019), the authors compiled tweets concerning two editions of the FIFA Soccer World Cup (2014 and 2018). Then, the authors measured the sentiment polarity of the tweets and identified the tweets containing hate-speech content. This latter challenge was treated as a multi classification problem with three labels, namely, hate speech, offensive language, and regular messages.

3 Corpus

In this section, the compiled corpus is described. We decided to compile a novel corpus given that, as far as we are concerned, there are no hate-speech datasets in Span-

[4] `https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy`

ish focused on football. The source of this dataset is Twitter. We decided to use Twitter because it is a forum for discussion of current topics and, at the time of the compilation process, it had a public API. In addition, its syndication mechanism through hashtags allows us to focus on specific debates.

The main topic covered are tweets related to the Spanish men's football league. This league was chosen because of the large number of comments in Spanish that each game provokes. In order to obtain large volumes of data, the search for matches of each match day is filtered by using hashtags. That is, if for example Real Madrid was playing against Levante, then that match was monitored using a search filter that contained the hashtag `#RealMadridLevante`. It should be noted that there are Spanish football teams, such as Real Madrid, Barcelona or Sevilla, which belong to large cities and have a higher number of fans than other teams, and this is reflected in the number of comments obtained for each match.

We also included in our search aggressive tweets. For this, more specific queries are used to capture types of hate speech. These queries include aggressive keywords such as *dumb*, *asshole*, or *bad ass*, along with the keyword *football*. Similarly, tweets with likely racist imagery and language are obtained. For this, we compiled a list with the names of several football players of color who play for top and second division teams of the Spanish football league. Then, searches for tweets whose content included these players' names and expressions with, depending on the context, racist meaning (e.g., *black*, *monkey*, or *orangutan*) were defined. In a similar fashion, to complete the dataset with misogynistic tweets we focused on women's football matches and on keywords that could indicate a case of misogyny. In this case, the word *woman* and *football* always appeared along with words or phrases with a probably sexist or derogatory meaning, such as *scrubbing*, *ironing* or *washing clothes*.

Once the tweets were obtained, the annotation phase began. For this, two collaborators of our research group annotated whether each tweet is aggressive and/or racist and/or misogynistic. The disagreements between annotators were resolved individually. Out of the 7483 tweets processed, 1739 were annotated as aggressive, 345 as racist and 350 as misogynistic, leaving the rest annotated as safe. The dataset was divided into three splits, namely, training, validation and test, in a ratio of 60-20-20. This separation is necessary as two hyperparameter tuning stages are conducted. The statistics of the dataset are depicted in Table 1. It is noticeable that the number of aggressive tweets is quite higher than the rest and this causes a strong imbalance in the labels that will condition the models evaluation process. The dataset is available at `https://pln.inf.um.es/corpora/hate-speech/hate-football-2023.zip`

Label	Train	Val	Test	Total
aggressive	1042	347	350	1739
misogyny	209	69	72	350
racist	206	69	70	345
safe	3355	1118	1119	5592
Total	4812	1603	1611	8026

Table 1: Dataset statistics.

Table 2 includes some examples of the dataset. As it can be observed, some of the documents are marked with more than one label, indicating that the same tweet can be, at the same time, aggressive, misogynistic, and racist. Note that the Safe (S) label is not necessary, as one tweet that does not contain any of the hate-speech traits can be considered safe. However, we decided to include this label explicitly on the dataset for the sake of clarity.

4 System architecture

For evaluating the different feature sets, the system depicted in Figure 1 was developed. In short, the pipeline can be described as follows. First, a data-cleaning stage takes place over the dataset to produce different versions of the texts that are used to extract different feature sets. Second, a two-fold feature extraction process is applied. On the one hand, we extract LFs using UMUTextStats (García-Díaz et al., 2022) and, on the other, sentence embeddings from nine LLMs. The third stage is the training of several deep learning classifiers. We trained a model for each feature set separately plus two strategies for combining all the feature sets. The first strategy is based on Knowledge Integration (KI), which consists in training a new

Tweet	A	R	M	S
No deseo que ganes nada, pero de madridista a madridista que te vaya bien y no te lesiones	-	-	-	✓
EXISTE ALGO PEOR QUE SER MUJER Y ENCIMA SER NEGRA???	✓	✓	✓	-
Este finde 3 al paleti pa debutar y el próximo clásico voltereta y a mamar	✓	-	-	-
Sois una puta lacra! Qué asco dais a todo aquel que no es ultra madridista	✓	-	-	-
El PSG está FRACASANDO ESTREPITOSAMENTE aun con Messi	-	-	-	✓
Hoy Kiko la ha cogido con el negro senegalés diakhaby	-	✓	-	-
Diakhaby es igual de malo que negro es, y es negro negro el cabron.	✓	✓	-	-
Creo que las palabras mujer y fútbol no pueden estar en la misma frase	-	-	✓	-
Que posibilidad hay q un negro como villa.. sin ser jugador de fútbol pueda estar con una mujer asi?	-	✓	-	-
Ojalá que el narizón de scaloni no se esté refiriendo al pecho Frionel Messi	✓	-	-	-

Table 2: Examples of the multi-label annotated corpus for the Aggressive (A), Racism (R), Misogyny (M), and Safe (S) labels.

multi-input neural network with all the feature sets. The second strategy is Ensemble Learning (EL), which combines the outputs and predictions of the models trained using only one feature set. Finally, we evaluate the performance of each classifier using the test split and the macro F1-score.

4.1 Data-cleaning

We conduct a data-cleaning stage in which acronyms are expanded, elongations and digits are removed along with hyperlinks, hashtags, quotations, and other punctuation symbols. Both the original and the cleaned version of the dataset are employed to extract the LFs as some categories, such as the ones related to correction and style, require the use of the original text. For the LLMs, however, only the original version of the texts is used, as each LLM has its own tokenizer and can contain special rules for Out of Vocabulary (OoV) words.

4.2 Feature extraction

The next stage in our pipeline is the feature extraction phase. Two families of features are considered: LFs, obtained with the UMUTextStat tool, and sentence embeddings, obtained for each of the LLMs.

UMUTextStats (García-Díaz et al., 2022) is a tool for extracting LFs, similar to LIWC (Tausczik and Pennebaker, 2010) but designed from scratch for Spanish. UMUTextStats considers a total of 364 LFs grouped in the following categories: i) phonetics, ii) psycho-linguistic processes, iii) morphosyntax, iv) stylometry, v) correction and style, vi) register, vii) semantics, viii) lexis, ix) pragmatics, and x) social media. Used in isolation, the performance of the LFs is more limited than the obtained with LLMs. However, LFs have two main advantages. On the one hand, they are interpretable, as it is possible to correlate the features with the target labels. On the other hand, LFs can be combined with other feature sets, such as the ones obtained with LLMs, improving the overall performance.

For extracting the features related to the LLMs, we proceeded as follows. We fine-tuned each LLM and then obtained the sentence embeddings of the dataset from the classification token, in a similar fashion as described at (Reimers and Gurevych, 2019).

In this research, a large variety of LLMs have been evaluated. Three of them are trained only with Spanish documents and four models are multilingual. Besides, two lightweight models are included, as we want to measure the performance drop between the regular and the lightweight models. Next, some relevant details about each model are provided.

- **MarIA** (Gutiérrez-Fandiño et al., 2022). It is a Spanish LLM based on the RoBERTa architecture. Therefore, it is a Masked Language Model (MLM), as it was trained with the objective of predicting a word that is hidden in a sentence. The dataset employed for pre-training MarIA was crawled from the National Library of Spain (Biblioteca Nacional de España) and contains 570

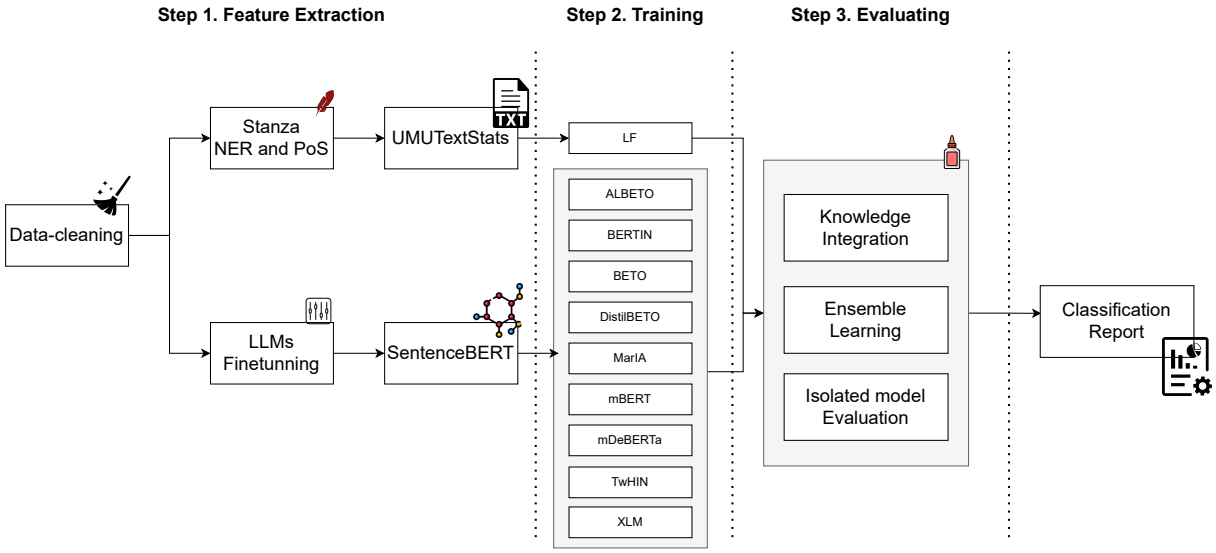

Figure 1: System architecture.

GB of data. MarIA uses the Byte-Pair Encoding (BPE) tokenizer.

- **BERTIN** (de la Rosa et al., 2022). It is a Spanish LLM and, similar to MarIA, is based on the RoBERTa architecture. However, BERTIN was trained using the Spanish split of the mC4 dataset (Xue et al., 2021).
- **BETO** (Cañete et al., 2020). It is a Spanish LLM trained on a large Spanish corpus. BETO is based on the BERT architecture. In fact, it has a similar size but BETO was trained as a MLM. BETO uses the BPE tokenizer, with a size of 31k tokens. BETO has two lightweight versions. One is **AlBETO**, which is based on the ALBERT architecture, and the other is **DistilBETO**, based on distillation. Both models are described in (Cañete et al., 2022).
- **Multilingual BERT (mBERT)** (Devlin et al., 2019). It is the multilingual version of BERT, trained in a self-supervised way with a total of 104 different languages with the largest Wikipedia. Apart from MLM, mBERT was also trained using Next Sentence Prediction (NSP).
- **XLM** (Conneau et al., 2020). It is a multilingual LLM based on the RoBERTa architecture. XLM was trained with more than 2 terabytes of texts from CommonCrawl with around 100 different languages.
- **TwHIN-BERT** (Zhang et al., 2022). It is a multilingual LLM trained with around 7 billion tweets written in 100 different languages. The training objective of TwHIN-BERT includes a social objective based on the rich social engagements within a Twitter Heterogeneous Information Network. In this work, we have used the base model of TwHIN, with 280M parameters.
- **Multilingual DeBERTa (mDeBERTa)** (He, Gao, and Chen, 2021). Currently, DeBERTa is in its 3rd version. It is a multilingual LLM that uses disentangled attention and a novel mask decoder. This LLM is trained with multilingual data from the CC100 dataset. It has 86M parameters with a vocabulary containing 250K tokens, which introduces 190M parameters in the embedding layer.

As some of the classifiers evaluated are based on the combination of the features, a representation of the LLMs that can be easily combined with the LF is required. In this sense, we decided to obtain the sentence embeddings; that is, every document in the corpus is encoded as vector for representing each document in the corpus. This representation can be combined with the fixed-size representation of the LFs.

Before obtaining the sentence embeddings, we conducted an hyperparameter optimization stage for each LLM separately.

The evaluated parameters include the learning rate (lr), the number of epoch, the batch size, the warm up steps and the weight decay. The results of this process are depicted in Table 3. It can be observed that the number of epochs varies depending on the LLM but, in general, is always larger or equal to 3. Concerning the batch size, most of the models benefit from small sizes, except BETO and mBERT. The warm-up steps, that modify the learning rate, is 1000 (the maximum evaluated) only for XLM and 500 for TwHIN. The rest of the LLMs, however, achieve better performance with smaller or no warm-up steps. Finally, for the weight decay, the values differ in the group of Spanish LLMs and in the group of multilingual LLMs, but are kept similar in the group of lightweight models.

After the hyperparameter tuning of the LLMs, the sentence embeddings for each document in the dataset was obtained from the classification token, in a similar fashion as Sentence BERT (Reimers and Gurevych, 2019) does.

4.3 Model training

The next step in our pipeline is the training of different classifiers. Three different approaches have been explored. First, classifiers trained using one feature set only: nine classifiers for each LLM and one classifier for the LFs. Second, a unique classifier combining all the feature sets using KI. This approach consists in training from scratch a new classifier that is multi-input. In deep layers of this model, the different neurons of each LLM are combined to produce the final classification probabilities. The main strength of KI is that it can learn when two or more feature sets produce high-order features. The last approach considered is based on EL. A total of four EL strategies have been evaluated. The first one is based on the mode of the predictions (MODE), the second one is based on averaging the probabilities output for each model (MEAN), the third one is based on obtaining the model that outputs the highest probability (HIGHEST) for each document, and the final strategy consists in a weighted mode (WEIGHTED), in which the importance of each model is measured in terms of the performance using the validation split.

For each deep learning classifier, except the ones based on EL, we conducted an hyperparameter optimization stage. For this, we evaluated the shape of the neural network layer. This shape is made up of a number of hidden layers, a number of neurons per layer, and how the neurons are arranged in different hidden layers. For the neural networks that have one or two hidden layers, all layers have the same number of neurons (brick shape). However, for deep neural networks other forms are also evaluated, such as funnel shape or triangle, in which the number of neurons is reduced as the network gets deeper. A dropout mechanism is also evaluated in ratios of .1, .2, and .3, along with different learning rates, batch sizes, and activation functions between the layers of the network.

The results of the hyperparameter tuning stage are shown in Table 4. It can be noted that most of the best neural networks are shallow neural networks, i.e., neural networks with only one or two hidden layers. This is the case of LF, in which the best model is obtained with a very simple neural network with a strong dropout mechanism. From the LLMs, it can be observed that all the multilingual LLMs, except XLM, require large number of neurons. XLM, however, achieved its best results with fewer neurons but larger number of hidden layers. In case of KI, the best results are obtained with only two hidden layers with 95 neurons in each layer, arranged in a brick shape. Regarding the dropout mechanism, all models benefit from using it but with different degrees, being LF, mBERT, TWHin, and ALBETO the models with higher dropout size.

5 Results and discussion

In this section, the results obtained for each classifier are compared and discussed. This comparison is made based on the macro-weighted F1-score, as the labels in the dataset are unbalanced. Precision and recall are also included for further analysis.

The results of the LFs and each individual LLM are depicted in Table 5 along with the results of combining all the classification models using both KI and EL. First, it is possible to observe that the performance of LF is acceptable, both in terms of precision and recall, but is quite limited as compared to the results achieved by the LLMs. In fact, the average macro F1-score of the nine

LLM	lr	epochs	batch size	warmup steps	weight decay
MarIA	2.9e-05	4	8	0	0.1
BETO	4.3e-05	3	16	250	0.062
BERTIN	2.7e-05	5	8	250	0.13
mBERT	4.2e-05	5	16	250	0.26
mDeBERTa	3e-05	3	8	0	0.081
TwHIN	3.4e-05	4	8	500	0.29
XLM	3.5e-05	3	8	1000	0.24
AlBETO	2.4e-05	5	8	250	0.2
DilstilBETO	4.8e-05	3	8	0	0.19

Table 3: Hyperparameter tuning of the LLMs. A total of nine models were evaluated. The LLMs are organized as focused on Spanish (MarIA, BETO and BERTIN), multilingual (mBERT, mDeBERTa, TwHIN, and XLM) and lightweight (ALBETO and DistilBETO).

Model	shape	layers	neurons	dropout	lr	batch size	activation
LF	brick	1	16	0.3	0.01	256	linear
BETO	funnel	5	37	0.2	0.01	256	selu
MarIA	brick	2	256	0.1	0.001	512	relu
BERTIN	brick	2	37	0.1	0.001	512	linear
mBERT	funnel	4	256	0.3	0.001	256	selu
mDeBERTa	brick	2	512	0.2	0.001	512	tanh
TwHIN	brick	2	512	0.3	0.01	256	linear
XML	funnel	5	37	0.2	0.01	256	selu
ALBETO	brick	2	48	0.3	0.01	128	sigmoid
DistilBETO	brick	2	64	0.1	0.01	256	linear
KI	brick	2	95	0.2	0.01	128	relu

Table 4: Hyperparameter optimization stage for the models trained with Keras. The results are grouped by the Linguistic Features (FL), the sentence embeddings for each LLM, and the Knowledge Integration (KI) strategy.

LLMs is 86.26%, with a standard deviation of 1.58. This constitutes an improvement over the performance of the LF of 13.79%. Out of the LLMs taken separately, MarIA is the one that achieves a better performance with a macro F1-score of 87.690%. Nonetheless, BETO and its lightweight version based on ALBERT (ALBETO) achieve a quite similar performance (87.240% and 87.440%, respectively). This is interesting since ALBETO and DistilBETO can provide faster inferences, being more suitable to real-time environments. It can also be noted that the performance of Spanish LLMs is, in general terms, superior to the multilingual models. However, the case of TwHIN must be highlighted as it outperformed BETO and BERTIN. We hypothesize that this model benefits from the fact that it has been trained with tweets, the same data source as the one used in our dataset.

Focusing on the results of the combined models, KI is the model that achieved the best overall performance, with a macro F1-score of 88.713%. This fact suggests that the integration of LLMs and LFs is beneficial for Spanish hate-speech detection in football. However, this model is not the one that achieved the best precision nor recall. The results of the EL are more limited than KI in terms of macro F1-score, but the EL based on highest probability achieved almost a perfect recall (98.287%) but with a very limited precision, and the EL based on the mode achieves the best precision (90.257%).

Next, an error analysis of the best model (KI) was conducted. We analyzed the results per label. First, regarding aggressive documents our model identified as aggressive some tweets that does not use aggressive lan-

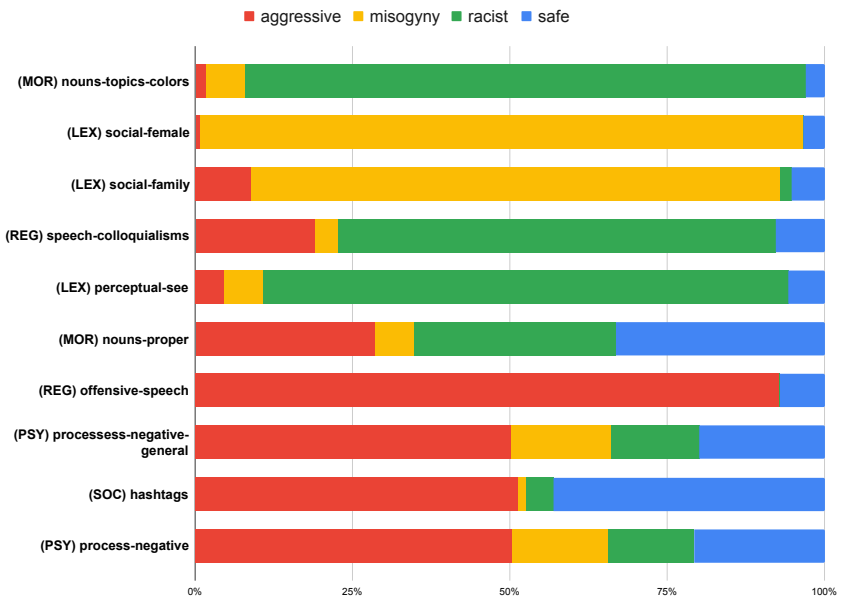

Figure 2: Information gain of the LFs grouped by label.

guage but that contain expressions concerning *woman*. There are also wrong classifications in tweets containing insults to some players indicating that they are stupid or calling them 'big heads'. Second, with respect to misogyny, we found out several misogynistic texts associated to the belief that a person cannot know about football just for being a woman. For example, in one of the samples a Twitter user is against having a woman presiding over German soccer. Third, concerning racism, we identified false-positives in which the word *negro* appears together with the name of African players. However, in such false-positive cases the term *negro* was not directed at the players but to other elements in the context. For example, some tweets in the dataset contain expressions such as *futuro negro*, which alludes to bad expectations; the idiom *ver todo negro*, which refers to a pessimistic point of view; or the idiom *ponerse ciego de rabia*, which means to get extraordinarily angry. These misclassifications suggest that the attention mechanisms of the LLMs are not capable to understand the meaning of certain figurative expressions.

Model	Precision	Recall	F1-score
LF	73.490	71.650	72.478
BETO	89.264	85.449	87.240
MarIA	87.027	88.462	87.690
BERTIN	85.016	87.437	86.173
mBERT	84.544	85.189	84.842
mDeBERTA	86.919	85.589	86.206
TwHIN	86.522	88.498	87.465
XLM	81.937	83.309	82.468
AlBETO	85.363	89.715	87.440
DistilBETO	86.391	87.348	86.859
KI	88.555	88.923	**88.713**
HIGHEST	67.523	**98.287**	79.690
MEAN	88.822	88.039	88.420
MODE	**90.257**	87.250	88.710
W. MODE	89.200	88.039	88.606

Table 5: Classification report with the test split comparing the linguistic features, each LLM separately, and all features combined using Knowledge Integration (KI) and the four strategies evaluated for ensemble learning.

To evaluate the interpretability of the results, we calculated the Information Gain (IG) of the LFs correlated with each label (see Figure 2). Features concerning with morphological and lexis categories are the most relevant. For example, topics related to colors are the most discriminatory features for the *racist* label. This fact suggests that expressions that refer to skin color are abundant. However, topics related to locations or toponymics are less relevant, which suggests that racist messages do not include specific

countries or cities. Another relevant feature concerning racism is the usage of colloquialisms, proper from informal speech. There are also relevant topics related to the perceptual sense of sight. This is because there are several expressions in Spanish that are used in rhetoric such as for example *Do you see what I'm telling you?*. We also found several features concerning misogyny, including the usage of lexis related to social female groups and family, such as mothers, sisters or grandmothers. It is also relevant for misogyny and racism detection the usage of proper nouns. This fact suggests that a relevant portion of hate-speech is directed towards specific people. Finally, concerning aggressive speech, we found a strong correlation with offensive expressions and negative psycho-linguistic processes. A significant correlation also exists between aggressiveness and the use of hashtags. This latest finding suggests that people get more aggressive and belligerent when discussing certain topics. Hashtags in Twitter are often used to draw attention to a certain topic and get other people to join the discussion.

6 Conclusions and future work

The focus of this work is on detecting hate speech in tweets written in Spanish and associated to the practice of a sport such as football. We have performed an in-depth comparison among different LLMs that can understand Spanish both separated and in combination with LFs using different strategies. For this, we have compiled a novel Spanish multi-label corpus focused on hate-speech in football. This corpus has been manually annotated to identify aggressive, misogynistic and racist comments in social networks. The results of our experiments suggest that the combination of all LLMs along with LFs is beneficial in terms of both performance and the interpretability of the results. Specifically, the best classification model is the one combining all individual features (both LFs and the nine LLMs) by means of the KI strategy, which achieved a macro F1-score of 88.713%.

In general, the results obtained by all LLMs are promising but there is plenty of room for improvement. As for future work, we plan to expand the dataset, with the objective of allowing the classifiers to learn more features concerned with aggressiveness, racism or misogyny by exploring and getting trained on a larger set of samples. In this line, we aim to include more documents for each label expanding the number of queries for each hate-speech trait. We consider that some of the queries employed were quite restrictive, which has enabled the evaluated classifiers to achieve a high performance. Moreover, as long as more datasets concerning hate-speech in football in Spanish are released, we need to test these models to observe how they perform with unseen data in order to identify some potential bias in the dataset. On the other hand, we will improve our pipeline by switching the hyperparameter tuning stage by a nested cross validation, in order to reduce the bias caused by the validation split. Finally, concerning the interpretability of the models, an important limitation of our approach is that it is model agnostic, that is, it does not consider the model but only the correlation between the LFs and the labels. To overcome this limitation, we plan to analyze the results of each model using the SHAP and LIME tools (Mosca et al., 2022).

Acknowledgments

This work is part of the research projects AIInFunds (PDC2021-121112-I00) and LT-SWM (TED2021-131167B-I00) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. This work is also part of the research project LaTe4PSP (PID2019-107652RB-I00/AEI/10.13039/501100011033) funded by MCIN/AEI/10.13039/501100011033.

References

Ali, R., U. Farooq, U. Arshad, W. Shahzad, and M. O. Beg. 2022. Hate speech detection on twitter using transfer learning. *Computer Speech & Language*, 74:101365.

Alkomah, F. and X. Ma. 2022. A literature review of textual hate speech detection methods and datasets. *Information*, 13(6):273.

Arango, A., J. Pérez, and B. Poblete. 2022. Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). *Information Systems*, 105:101584.

Bilal, M., A. Khan, S. Jan, and S. Musa. 2022. Context-aware deep learning model for detection of roman urdu hate speech on social media platform. *IEEE Access*, 10:121133–121151.

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Cañete, J., S. Donoso, F. Bravo-Marquez, A. Carvallo, and V. Araujo. 2022. ALBETO and DistilBETO: Lightweight spanish language models. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 4291–4298. European Language Resources Association.

Chiril, P., E. W. Pamungkas, F. Benamara, V. Moriceau, and V. Patti. 2022. Emotionally informed hate speech detection: A multi-target perspective. *Cognitive Computation*, 14(1):322–352, Jan.

Cleland, J. 2014. Racism, football fans, and online message boards: How social media has added a new dimension to racist discourse in english football. *Journal of Sport and Social Issues*, 38(5):415–431.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

de la Rosa, J., E. G. Ponferrada, M. Romero, P. Villegas, P. González de Prado Salas, and M. Grandury. 2022. BERTIN: efficient pre-training of a spanish language model using perplexity sampling. *Proces. del Leng. Natural*, 68:13–23.

Devlin, J., M. Chang, K. Lee, and K. Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

García-Díaz, J. A., F. García-Sánchez, and R. Valencia-García. 2023. Smart analysis of economics sentiment in spanish based on linguistic features and transformers. *IEEE Access*, 11:14211–14224.

García-Díaz, J. A., P. J. Vivancos-Vicente, Á. Almela, and R. Valencia-García. 2022. UMUTextStats: A linguistic feature extraction tool for spanish. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 6035–6044. European Language Resources Association.

García-Díaz, J. A., R. Colomo-Palacios, and R. Valencia-García. 2022. Psychographic traits identification based on political ideology: An author analysis study on spanish politicians' tweets posted in 2020. *Future Generation Computer Systems*, 130:59–74.

García-Díaz, J. A., M. Cánovas-García, R. Colomo-Palacios, and R. Valencia-García. 2021. Detecting misogyny in spanish tweets. An approach based on linguistics features and word embeddings. *Future Generation Computer Systems*, 114:506–518.

Gutiérrez-Fandiño, A., J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. R. Penagos, A. Gonzalez-Agirre, and M. Villegas. 2022. MarIA: Spanish language models. *Proces. del Leng. Natural*, 68:39–60.

He, P., J. Gao, and W. Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with

gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543.

Husain, F. and O. Uzuner. 2022. Investigating the effect of preprocessing arabic text on offensive language and hate speech detection. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(4), jan.

Mansur, Z., N. Omar, and S. Tiun. 2023. Twitter hate speech detection: A systematic review of methods, taxonomy analysis, challenges, and opportunities. *IEEE Access*, 11:16226–16249.

Mathew, B., R. Dutt, P. Goyal, and A. Mukherjee. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, page 173–182, New York, NY, USA. Association for Computing Machinery.

Mehta, H. and K. Passi. 2022. Social media hate speech detection using explainable artificial intelligence (XAI). *Algorithms*, 15(8):291.

Min, B., H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, and D. Roth. 2021. Recent advances in natural language processing via large pre-trained language models: A survey. *CoRR*, abs/2111.01243.

Mosca, E., F. Szigeti, S. Tragianni, D. Gallagher, and G. Groh. 2022. Shap-based explanation methods: A review for NLP interpretability. In N. Calzolari, C. Huang, H. Kim, J. Pustejovsky, L. Wanner, K. Choi, P. Ryu, H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, and S. Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 4593–4603. International Committee on Computational Linguistics.

Mozafari, M., R. Farahbakhsh, and N. Crespi. 2022. Cross-lingual few-shot hate speech and offensive language detection using meta learning. *IEEE Access*, 10:14880–14896.

Oliveira, L. and J. Azevedo. 2022. Using social media categorical reactions as a gateway to identify hate speech in covid-19 news. *SN Computer Science*, 4(1):11, Oct.

Omar, M., S. Choi, D. Nyang, and D. Mohaisen. 2022. Robust natural language processing: Recent advances, challenges, and future directions. *IEEE Access*, 10:86038–86056.

Paz, M. A., J. Montero-Díaz, and A. Moreno-Delgado. 2020. Hate speech: A systematized review. *SAGE Open*, 10(4):2158244020973022.

Plaza del Arco, F. M., M. D. Molina-González, L. A. Ureña López, and M. T. Martín Valdivia. 2021. Comparing pretrained language models for spanish hate speech detection. *Expert Syst. Appl.*, 166:114120.

Poletto, F., V. Basile, M. Sanguinetti, C. Bosco, and V. Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523.

Reimers, N. and I. Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Roy, P. K., S. Bhawal, and C. N. Subalalitha. 2022. Hate speech and offensive language detection in dravidian languages using deep ensemble framework. *Computer Speech & Language*, 75:101386.

Tausczik, Y. R. and J. W. Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Vasconcelos, M., J. Almeida, P. Cavalin, and C. Pinhanez. 2019. Live it up: Analyzing emotions and language use in tweets during the soccer world cup finals. In *Proceedings of the 10th ACM Conference on Web Science*, pages 293–294.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Wullach, T., A. Adler, and E. Minkov. 2022. Character-level hypernetworks for hate speech detection. *Expert Syst. Appl.*, 205:117571.

Xue, L., N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 483–498. Association for Computational Linguistics.

Zhang, X., Y. Malkov, O. Florez, S. Park, B. McWilliams, J. Han, and A. El-Kishky. 2022. TwHIN-BERT: A socially-enriched pre-trained language model for multilingual tweet representations. *CoRR*, abs/2209.07562.