Array 28 (2025) 100526

Contents lists available at ScienceDirect

Array

journal homepage: www.elsevier.com/locate/array

From data to detection: Developing a corpus and training language models for the identification of anti-refugee narratives in Spanish

Jacinto Mata [a,*], Estrella Gualda [b], Victoria Pachón [a], Carolina Rebollo [c], Juan L. Domínguez [a]

[a] I2C Research Group, CITES Research Center, Information Technology Department, Universidad de Huelva, Spain
[b] Social Studies and Social Intervention Research Group (ESEIS), Center for Research in Contemporary Thought and Innovation for Social Development (COIDESO), Universidad de Huelva, Spain
[c] Universidad de Granada and Social Studies and Social Intervention Research Group (ESEIS), Center for Research in Contemporary Thought and Innovation for Social Development (COIDESO), Universidad de Huelva, Spain

ARTICLE INFO

Dataset link: https://doi.org/10.5281/zenodo.17259982

Keywords:
Deep learning
Language models
Transformers
Social media
Twitter
Hate speech
Refugees

ABSTRACT

This study addresses the automatic detection of negative anti-refugee messages in Spanish texts, using language models based on pre-trained Transformers models. Despite numerous studies on hate speech detection, few have concentrated on Spanish, particularly regarding hostility towards refugees. To fill this void, we developed *HateRADAR-es*, a new corpus of Spanish-language tweets manually annotated by sociologist and social workers experts to identify the presence or absence of hateful content directed at refugees. This dataset has been made available to the research community to encourage further investigation. A comprehensive experimental framework to tackle this challenge, composed of several stages to achieve language models with a high efficacy in detecting such messages, is presented. To address the class imbalance issue in the data, data augmentation techniques are applied, and extensive experimentation is carried out to find the best values for the hyperparameters of the language models to achieve better performance. In the evaluation process, an ensemble of the fine-tuned models BETO, XLM-RoBERTa, and RoBERTa-large achieved the best results, with an accuracy of 0.891, an F1-measure of 0.860, and an AUC-ROC of 0.892. These findings underscore the effectiveness of combining multiple models into an ensemble to handle the complexity and nuances of hate speech on social media, offering a promising direction for future adaptations and applications of language models in specific hate contexts.

1. Introduction

In recent decades, the diversity of international conflicts, as well as natural disasters, have led to a worldwide increase in the number of refugees, which has been on the rise since 2011 and reached a significant peak in 2015 in the context of the so-called "refugee crisis" [1–4]. The massive arrival of refugees to a territory is commonly accompanied by different changes and the emergence of fears and apprehensions on the part of the population [1]. On the other hand, the increased use of the Internet and the loudspeaker of social media produces the growth of online hate speech over the last years [3,5–8]. Aspects such as anonymity, disinhibition, or the feeling of impunity on the Internet contribute to the spread of hatred. Likewise, political interest and violent extremism propagate and incite hatred directed at individuals and their communities, mobilizing emotions and contributing to the polarization of different sectors of society.

Although there is no universally accepted definition of the term hate speech [9,10], it can be understood as those forms of expression that are disseminated and that incite, promote, or justify hatred or intolerance [5,11,12] directed against a person or a particular group of persons based on their identity (real or perceived), be it related to their religion, ethnicity, nationality, race, color, descent, gender or other identity factor [5,10], as would be the case of hate speech directed at refugees.

It is essential to mention that among the main types of hate speech is Islamophobic speech, as the European Commission, the United Nations, and other organizations continuously warn of the hostility received from the Muslim population [13]. As [1] indicates, a growing xeno-racism of an Islamophobic nature and a manifest hostility towards the Muslim community is detected in Europe, together with an atmosphere of intolerance and exclusion, which is accompanied by different hate crimes in the form of physical and verbal attacks. It is a negative

* Corresponding author.
E-mail address: mata@uhu.es (J. Mata).

https://doi.org/10.1016/j.array.2025.100526
Received 19 March 2025; Received in revised form 4 August 2025; Accepted 27 September 2025
Available online 3 October 2025

discourse strongly disseminated on social platforms around specific cases [3]. Hate speech propagated against refugees is recurrent internationally in different Internet and social media spaces (YouTube, Facebook, Twitter or X, Instagram, and others). The harmful content is diverse. Stereotypes, prejudices, false or distorted information, portraying them as "villains", "criminals", "dangerous", "uninvited visitors", or other negative appellatives are propagated, sometimes subtly [6, 14–17]. These contribute to the construction and deepening of public stigmatization and otherization of individuals or groups of refugees and generate problems for societies such as racism, intolerance, and discrimination [3,18], increased vulnerability or risks of exploitation and marginalization of refugees [19]. Moreover, hate speech or incitement to hatred is close to hate crimes and violence. European-Commission [12] has warned of the dangerous link between hate speech and violence, stating that hate speech is an extreme form of intolerance that contributes to hate crimes.

The context briefly described above justifies the need for tools for the automatic detection of hate speech directed towards refugees. This need is more pressing in the case of languages other than English, where the development of systems to monitor online content and remove abusive, offensive or hateful content has been lower, this detection being vital to fostering less bellicose environments [20].

The interest of automatically identifying hate speech in online texts has emerged as a critical area of research in Natural Language Processing (NLP) and computational linguistics. Various studies have focused on developing machine learning models and algorithms capable of identifying and classifying hate speech content in different languages [21–24]. However, the majority of these studies have been conducted in English, with limited attention given to other languages, such as Spanish. This is especially relevant considering that Spanish is one of the most widely spoken languages in the world, and that anti-refugee narratives are increasingly present among Spanish-speaking populations, yet they remain underrepresented in current research and available resources for hate speech detection. Furthermore, the datasets available for hate speech detection are labeled for a general domain and not for a specific domain such as the refugee population. Hence, the aim of this study is to obtain a customized and effective language models to automatically identify negative messages in Spanish that are aimed at this specific group. While our focus is on tweets in Spanish, the approach outlined in this paper and the resulting findings can be applied to other languages.

For a language model to learn to detect a certain type of content from a specific domain, it is necessary to have a training dataset. To ensure that the model is effective, it is essential to ensure that the training dataset is correctly labeled. Otherwise, the results provided may be biased or incorrect. For this reason, obtaining such resources is not easy. The construction of the corpus used in this study has been carried out by sociologist and social workers experts following a meticulous manual labeling methodology, ensuring the validity of the data. This work addresses a binary classification task, where each document (tweet) is assigned a positive ("1") or negative ("0") label indicating the presence or absence of hateful messages towards refugees. Details on the construction of the corpus are described in the following sections.

The following items summarize the main contributions of our work:

(i) A study on the implication of hate speech directed towards refugees and its treatment from the point of view of NLP.
(ii) A new dataset about refugees for Spanish hate speech detection, created using Twitter (rebranded as X) as a source of information, and labeled by sociologist and social workers experts.
(iii) A comprehensive experimental framework to obtain effective language models based on Transformers.
(iv) An exploration of different computational methods to achieve better model performance: data augmentation, hyperparameter search, and ensemble techniques.
(v) An error analysis, from computational and sociological perspectives, to determine future directions of the study.

The rest of this paper is organized as follows: in Section 2, related works are discussed, exploring prior research conducted in the field. Section 3 presents a new dataset in Spanish for hate speech detection towards refugees, and the annotation process. Section 4 details the different approaches used to achieve the proposed task. The experimental framework developed is described in Section 5. Results and error analysis are presented in Section 6. Finally, Section 7 summarizes the conclusions of this study and shows future work.

2. Related works

The automatic detection of hate speech in online platforms has attracted significant attention across various languages and contexts. While considerable work has been done in English, less focus has been given to other languages, particularly in detecting hate speech against minority groups, such as refugees in the Spanish language. Jahan and Oussalah [22] provide a comprehensive systematic review of hate speech detection using NLP, highlighting the prevalence of English language studies and underscoring the need for research in other linguistic contexts. Similarly, Mansur et al. [24] discuss the breadth of methods used for hate speech detection on Twitter, noting that while many techniques have been developed, applications in languages other than English remain limited, emphasizing the relevance of our focus on Spanish. Advancements in machine learning models, particularly those based on deep learning, have shown promising results in text classification tasks. Li et al. [25] survey these developments, from traditional machine learning methods to modern deep learning techniques. Furthermore, Garrido-Merchan et al. [26] compare the efficacy of BERT models against traditional machine learning models in text classification tasks, providing evidence that supports the use of transformer-based language models for enhanced accuracy and generalization in hate speech detection. In addition to model comparisons, the societal impact and technological challenges of detecting hate speech are critically reviewed by Gudumotu et al. [23], who explore various deep learning models for identifying hate speech and bullying. This work provides insights into the complexities of implementing these models in social media platforms, which align with the challenges we address in our own research.

The detection of hate speech online has been the subject of various investigations, particularly in contexts where minorities such as migrants and refugees are frequently victimized. While significant studies exist, few address this phenomenon in the Spanish language with a specific focus on refugees. Arcila-Calderón et al. [27] developed classifiers using both shallow and deep learning to detect racist and xenophobic discourses directed towards migrants and refugees in Spanish. Unlike our study, their research focused on a broader range of xenophobia without constructing a specific corpus for refugees. In [28], the authors present a study for the automatic detection of racism and xenophobia on Twitter using deep learning techniques. Although the results obtained, especially with transformer-based language models, are good, the authors assert that their research presents preliminary findings that need to be deeply investigated to address some weaknesses such as the size of the dataset used. Moreover, Sánchez-Holgado et al. [18] explored the correlation between online hate speech and the acceptance of immigrants at the provincial level in Spain. This study used social acceptance data to measure the impact of hate speech, a methodology distinct from ours, which focuses on the development and evaluation of NLP models to automatically identify hate speech on social networks.

These works complement the research presented in this study by demonstrating the diversity of approaches in the study of hate speech and highlight the importance of developing specific solutions for linguistic and cultural contexts. Our work distinguishes itself by focusing on the optimization of NLP models for the Spanish and by directly addressing the detection of hate speech towards refugees, adding a layer of specificity in the identification of such discourses on social media platforms.

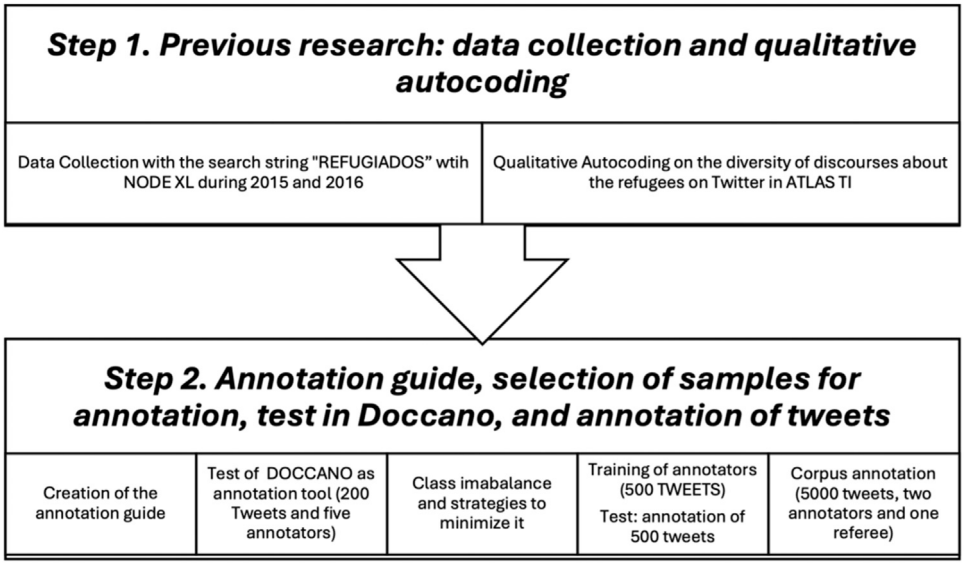

Fig. 1. Process of creation of the Corpus for training: the *HateRADAR-es* dataset.

3. Corpus design and construction

This section introduces the ***HateRADAR-es*** (*Anti-Hate Refugees Annotated Dataset and Analysis Resource*) dataset, a corpus of Spanish-language Twitter messages meticulously compiled for the purpose of identifying content containing both hateful and negative expressions directed towards refugees. We elaborate on the methods employed for extracting and annotating the documents, providing comprehensive statistical insights into the collection. Currently, there are few datasets in Spanish explicitly designed for identifying adverse messages directed towards refugees. Therefore, we want to offer the scientific community a useful resource for researching and deepening new approaches for the automatic detection of hate speech towards this group. Overall, we are confident that this dataset will be an invaluable resource for researchers who are seeking to understand the impact of hate speech on refugees, especially filling a gap with regard to annotated datasets on refugees in Spanish. By providing access to this data, we hope to contribute to a greater understanding of this issue and to promote positive change in our communities. The *HateRADAR-es* dataset is available on Zenodo[1] for research purposes.

Fig. 1 shows, schematically, the steps followed to build the *HateRADAR-es* dataset.

3.1. Data collection and qualitative autocoding

The tweets in this collection [8] were extracted using the NodeXL tool (https://nodexl.com/), which consists of an Excel template that allows the extraction of data from Twitter and other social networks through a search engine where the keyword or words that function as search criteria are inserted. This tool is mainly oriented to perform social network analysis (relationship between users, retweet networks, hashtags, communities, etc.) and therefore, it collects and orders the data in a structure prepared for this purpose. For example, renaming users as source and target according to their role in the social interaction. One of the fields collected by NodeXL is the "tweet" field, which contains the content of the message and is the one used to build this working collection, together with the "id". NodeXL, at the time the tweets were exported, collected the data through the Rest API, collecting the tweets published retrospectively.

Thousands of tweets were extracted on a daily basis from the search string "refugees" in six different languages: Spanish ("refugiados"), English ("refugees"), German ("fluechtlinge"), French ("réfugiés"), Italian ("rifugiati") and Portuguese ("refugiados"), during a period from December 2015 to December 2016 (one year). During this period, the international refugee situation worsened significantly, reaching unprecedented numbers. The main causes of this crisis were civil wars and violence in Syria, as well as in other places such as Iraq, Yemen, and sub-Saharan Africa. These conflicts forced the inhabitants of these regions to flee their homes and seek refuge in neighboring countries or risk reaching European Union countries through irregular displacement routes. In response to this situation, the UNHCR declared it a refugee crisis [29].

This serious scenario did not go unnoticed on social media, where the migration crisis became a topic of intense debate among numerous and diverse users. Twitter offered significant advantages in contributing to the debate and disseminating information owing to its immediacy and ease of following events in real-time. Users used these tools to inform themselves, socially denounce the situation, criticize government policies, and actively participate in conversations about the refugee crisis. In addition, these social media discussions contributed to increasing the visibility of the crisis, generating greater international awareness and mobilization regarding the situation of refugees.

All daily extractions were merged by language to obtain six different datasets. Once merged, any duplicated tweets were removed by taking into account the "id" of the tweet, which confers a unique identification number to each one. In this work we have used the Spanish dataset composed of 355,810 tweets. Since our interest was not in the dissemination patterns of the tweets, we excluded retweeted messages to avoid repetition of tweets, finally obtaining a sample of 90,144 tweets. As part of the preprocessing applied during the construction of the dataset, emojis and URLs were removed, and user mentions were anonymized by replacing them with @USER. The rest of the text, including hashtags, was preserved in its original form to maintain linguistic context. This process resulted in a clean and noise-free dataset, so the only additional preprocessing performed before models training was tokenization using the tokenizer associated with each transformer-based language model, and the tokenizer provided by the Natural Language Toolkit (NLTK) [30] library for traditional machine learning algorithms.

This collection of tweet messages was previously labeled with the perspective of knowing the diversity of different discourses about refugees in different languages [7,8]. At that time, the qualitative

[1] https://doi.org/10.5281/zenodo.17259982

software *Atlas.ti* (https://atlasti.com/) was used to study the tweets about the refugees through sociological content and thematic analysis applying automatic coding techniques in *Atlas.ti*. An extensive codebook was produced in that moment, which can be found in the annexes of Rebollo [8].

3.2. Annotation guide

While previous research (III-a) used a codebook to label the diversity of discourses on refugees, an annotation guide was explicitly developed to detect online hate speech towards refugees. This guide, designed by experts in the field, would make it easier to tackle the complex and subjective task of manually tagging our dataset. The annotation process consisted of labeling with "1" when "Hate speech towards refugees is detected in the tweet" and "0" when "Hate speech is not detected in the tweet". At annotation, hate speech has been understood, as specified in Section 1, by those forms of hate speech that incite, promote, or justify hatred or intolerance directed against a person or a particular group of persons based on their identity (real or perceived), whether related to their religion, ethnicity, nationality, race, color, descent, gender or other identity factor, as would be the case with hate speech directed at refugees.

3.3. Test of the annotation tool

To carry out a manual labeling process by human experts, we used the *Doccano* tool (https://github.com/doccano/doccano), an open-source text annotation platform designed for human users. *Doccano* offers annotation functions tailored for tasks such as text classification, among others. Five annotators were trained to test this tool. Two hundred tweets were randomly extracted from the global Spanish dataset (90,000 tweets) to do this test. The complete team evaluated all the processes and results of this first test.

3.4. Class imbalance and sampling of tweets for annotation

After the Doccano annotation test was performed by five annotators, a class imbalance was detected in the dataset (with a deficit of "1. Hate" and a predominance of "0. Non-hate/Neutral" tweets). As a strategy to achieve a more balanced collection to be annotated, the previous codebook on the diversity of discourses on refugees (III.a) was used to extract different tokens and regular expressions associated with hate tweets, on the one hand, and non-hate and neutral tweets on the other. Two subsamples of 250 tweets were extracted from the global dataset. One was filtered with codes relating to neutral or non-hateful messages towards refugees, and the second was filtered with negative messages towards refugees. After randomly selecting both the subsamples, 500 tweets were joined and randomly ordered before annotation. This sample of 500 tweets was used to train the two annotators and the referee before labeling 5000 tweets. The same procedure was used to select the final corpus of 5000 tweets to be annotated.

3.5. Corpus annotation

In a first stage, the dataset was labeled by two experts on sociology and social work with the initial aim of distinguishing between tweets containing hate speech, racist or xenophobic discourse towards refugees (label "1"), and tweets that did not contain such an issue (label "0"). Given the subjective nature and the level of difficulty of the task, manual labeling of this dataset was facilitated with the support of the annotation guide specifically designed by domain experts [31]. The overall inter-annotation agreement during this phase reached a satisfactory correlation index of 0.66 of the Cohen's Kappa coefficient [32]. Out of 5000 labeled tweets, annotators reached an 85% agreement. In a subsequent stage, tweets where there was no consensus among annotators or raised uncertainties were meticulously analyzed and evaluated by a third person, also an expert in sociology of migrations and social work, to resolve disagreements (referee). This thorough review and discussion of the most challenging classification cases assisted in addressing certain subtleties in tweets posted on Twitter in this domain. Table 1 presents various examples of annotated tweets along with a brief explanation of the label chosen by the annotators.

3.6. Dataset features

Once the annotation process was completed, a study was conducted on some of the most relevant features of the collection. Despite implementing a first strategy to reduce class imbalance (III.b.3), after the annotation process, there is still a significant imbalanced distribution of tweets. Table 2 shows this distribution, indicating that 76.2% (3810) of the messages do not contain hatred towards refugees, while the remaining 23.8% (1190) contain negative content towards them.

It is important to note that, although a balance was sought in the subsets of messages without and with hateful content towards refugees, after the annotation process, a considerable imbalance has arisen regarding the labels. This is because many of the messages that were selected while searching for words with potentially negative connotations such as *violación* (*rape*), *invasión* (*invasion*), *delito* (*crime*) or *islamización* (*Islamization*), did not reveal hate messages during the annotation process and, therefore, were labeled with the label "0".

Automated text classification becomes a challenging task when dealing with an imbalanced training dataset, necessitating the implementation of balancing techniques before applying machine learning algorithms [33].

A study of the vocabulary used in the messages was also carried out to identify the most frequently occurring terms in this type of content. Analyses were performed on the complete dataset and separately on the subsets corresponding to each label. Fig. 2 shows the most frequent terms for each label. It can be observed that terms such as "musulmanes" ("muslims") and "violaciones" ("rape") frequently appear in messages containing hatred, whereas terms such as "ayuda" ("help") and "derechos" ("rights") are more prevalent in messages that do not contain hate.

Building on the annotated dataset presented in this section, the following part of the paper describes the methodology used to train and evaluate the classification models designed to detect hate speech against refugee individuals.

4. Proposed approach

Once the data was labeled, machine learning algorithms and deep learning models were implemented to identify hate messages targeting refugees on Spanish tweets. To obtain the most effective model, a specific experimental framework described in the Subsection *Experimental Framework* was designed. This work includes an analysis of traditional machine learning algorithms that have demonstrated good performance in text classification in recent years [25,34,35], as well as a detailed study of several state-of-the-art language models [26,36]. Traditional machine learning algorithms investigated include Support Vector Machines (SVM), Random Forest (RF), K-Nearest Neighbors (KNN) and Logistic Regression (LR), all of which are examined in the study by Das et al. [34], in addition to eXtreme Gradient Boosting (XGBoost) [37]. These traditional algorithms also serve as solid baselines against which to compare the performance of deep learning models.

In terms of deep learning algorithms [38], Transformers [39] were chosen as the prominent models for this study. Transformers have emerged as the leading architectures in NLP, widely adopted in various tasks. These models undergo pre-training using vast text corpora, enabling them to efficiently capture complex linguistic relationships. Moreover, they can be fine-tuned with domain-specific data to achieve optimal results in a wide range of applications, such as text classification or offensive message detection. In this study, a comparative

Table 1
Some examples of labeled tweets in Spanish and English, along with a brief explanation of the reason for the annotation.

Spanish tweet	English version	Label	Explanation
Esta es la mejor prueba de que JAMAS se van a occidentalizar. Islam x encima de todo! FUERA REFUGIADOS DE ESPAÑA!	This is the best proof that they will NEVER westernize. Islam above all else! REFUGEES OUT OF SPAIN!	1	This is a tweet that proposes to expel refugees, rejecting the Islam and its integration possibilities in Western societies
VERGuenza y asco de países europeos q consienten esta pederastia encubierta. STopIslam	SHAME and disgust of European countries that condone this covert paedophilia. STopIslam	1	This is a tweet associating Islam with paedophilia, rejecting European policies
El plan secreto de los refugiados musulmanes: 'islamizar Alemania' vía @user	Muslim refugees' secret plan: 'Islamise Germany' via @user	1	This tweet defends the conspiracy theory of the invasion of Europe by Islam (Eurabia) that rejects refugees
@user aparte como he dicho mientras más refugiados pues más probabilidades hay de atentados violaciones etc... caballo de Troya	@user apart, as I said, the more refugees, the more likely there are to be attacks, rapes etc... Trojan horse	1	This tweet associates refugees with attacks, rape and other negative aspects
Polonia se niega a aceptar refugiados musulmanes tras los atentados de Bruselas	Poland refuses to accept Muslim refugees after Brussels attacks.	0	This is a new about migration and border policies in Poland
#refugioderechoshumanos Vaya habéis tardado hoy en cagar, pero lo de los refugiados ya lo habéis trillado otras veces, ya no cuela piojosos	#humanrightsrefuge You've taken your time today to shit, but you've already done it before with the refugees, it doesn't work any more, you lousy licem	0	This is hate speech but it is not directed specifically at refugees
Manifestación hoy en Berlín de Neonazis y extrema derecha contra política de asilo para refugiados y contra islamización de Alemania	Neo-Nazis and right-wing extremists demonstrate today in Berlin against asylum policy for refugees and the Islamization of Germany	0	This is information about a demonstration, describing its motives
@user ¿Tú qué haces por los refugiados? PD: Lo de 'mosquita muerta' puedes llamárselo a tu madre	@user What do you do for refugees? PS: You can call your mother a 'dead mosquito'	0	It is negative speech towards another person but not specifically towards refugees

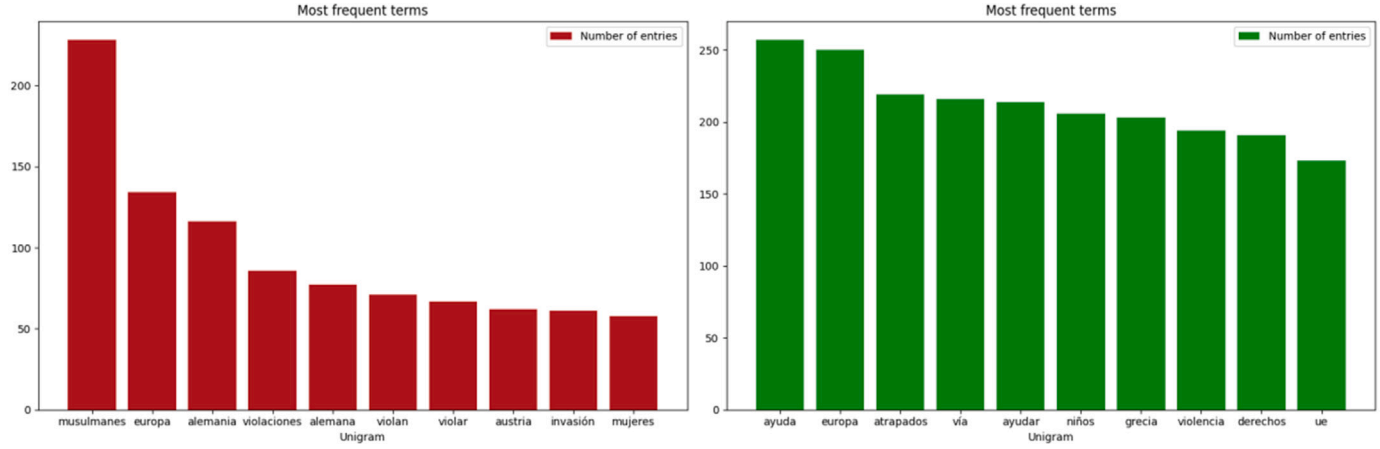

Fig. 2. Most frequent terms (excluding search terms) used in tweets. (left) Label "1". (right) Label "0".

Table 2
Distribution of labels in the dataset.

Dataset	Tweets	Label 0	Label 1
HateRADAR-es	5000	3810 (76.2%)	1190 (23.8%)

evaluation of several pre-trained linguistic models based on BERT [40], a widely recognized type of Transformer, was carried out.

This broad exploration of machine and deep learning approaches allows us to identify the most promising and effective techniques for the detection of hate speech in social networks, with a particular focus on its application in the context of anti-refugee discourse on Twitter in Spanish.

4.1. Traditional machine learning algorithms

For the development of the traditional models, Python, the scikit-learn [41] and NLTK libraries were used. In all cases, the default parameter settings of the scikit-learn library were used. To enhance the robustness of the study, five text representation approaches were used: Count Vectorizer [42], One Hot Encoding [43], TF-IDF [42], Word2Vec [44] and Glove [45].

4.2. Transformer-based language models

Regarding the deep learning models used for our transfer-learning approach, four pre-trained models in Spanish and a multilingual model

Table 3
Examples of tweets translated to increase the number of positive class cases using single translation.

Original tweet	English version	Translated tweet
Los refugiados musulmanes costarán a Alemania 50.000 millones de euros en los próximos dos años	Muslim refugees will cost Germany 50 billion euros in the next two years	Los refugiados musulmanes costarán a Alemania 50.000 millones de euros en los próximos dos años
Ok, tarjeta sanitaria, casa y 400€ a los refugiados. Ahora que ayuden también a los madrileños que están en situaciones penosas.	Ok, health card, house and 400€ to refugees. Now help also the Madriders who are in distress.	Ok, tarjeta de salud, casa y 400€ para los refugiados. Ahora también ayudar a los madrileños que están en apuros.
Solución para la entrada y distribución de refugiados en Europa: musulmanes de vuelta a su país o deportados a países musulmanes.	Solution for the entry and distribution of refugees in Europe: Muslims back to their country or deported to Muslim countries.	Solución para la entrada y distribución de refugiados en Europa: musulmanes de vuelta a su país o deportados a países musulmanes.

were selected. All models were implemented in Python, and the HuggingFace Transformers library [46] was used for the training and validation processes.

- **BETO**. The Spanish-BERT model [47] uses a similar architecture to the BERT-Base model and was trained on a corpus that exclusively contains Spanish texts, including data from Wikipedia and the OPUS Project. Specifically, we have used the *bert-base-spanish-wwm-cased* (https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased) model.
- **RoBERTa**. A robustly optimized BERT pre-training approach is an improved version of BERT where key hyper-parameters are modified [48]. There are several versions of RoBERTa pre-trained in Spanish. In this work, models pre-trained with data from the National Library of Spain (BNE) [49] were used. Specifically, experiments were carried out with the *base* (https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne) and *large* versions (https://huggingface.co/PlanTL-GOB-ES/roberta-large-bne). The architecture of *roberta-base-bne* consists of 12 layers, 768 hidden units, 12 attention heads, and 125 million parameters, whereas that of *roberta-large-bne* comprises 24 layers, 1024 hidden units, 16 attention heads, and 355 million parameters.
- **XLM-RoBERTa**. It is a multilingual version of the RoBERTa model [48]. It was pre-trained on 2.5TB of a filtered CommonCrawl data containing 100 languages [50]. In this work, the *xlm-roberta-base* (https://huggingface.co/xlm-roberta-base) was the selected version.

4.3. Balancing techniques

As mentioned in the Subsection *Dataset Features*, the *HateRADAR-es* dataset exhibits significant class imbalance, a common challenge in machine learning classification tasks. It is necessary to implement approaches that counteract this imbalance to prevent models from prioritizing classification of the majority class. Resampling is the most widely used approach to address class imbalance in text classification, offering a straightforward yet effective solution [51]. Oversampling and undersampling are two fundamental resampling techniques. While oversampling increases the number of samples from minority classes, undersampling decreases the number of samples from majority classes. To overcome this imbalance problem, in this work we use an oversampling approach to balance the number of tweets from both classes. The aim is that no message is lost in this process as all of them can provide valuable information.

Specifically, in this work we have applied the oversampling technique known as Data Augmentation by generating new messages of the minority class using *Round-Trip Translation*. This approach is based on translating the original text into another language and then translating it back into the original language. In this way, new examples are generated with a semantic equivalence to the original text, but with variations in the words used. This technique has proven to be effective in increasing the amount of data in the minority class and improving the performance of machine learning models.

Automatic translation between languages often yields messages that are identical, particularly when dealing with short texts, leading to a dataset of suboptimal quality. To address this challenge, we explored two Data Augmentation techniques. One approach involved translating tweets from Spanish to English and then back to Spanish. Using this method, 106 out of the 952 positive class tweets (11.1%) remained unchanged. The alternative strategy encompassed translating tweets from Spanish to English, then to German, and finally back to Spanish. This approach resulted in only 8 messages (0.8%) being duplicated compared to the originals. These experiments underscore the nuances and challenges inherent in leveraging translation-based augmentation methods to enhance dataset quality.

For the automatic translation process, the models proposed in the OPUS project [52] were used. Specifically, the models used were *opus-mt-es-en* (https://huggingface.co/Helsinki-NLP/opus-mt-es-en) and *opus-mt-en-es* (https://huggingface.co/Helsinki-NLP/opus-mt-en-es) for translating from Spanish to English and from English to Spanish, respectively. We also used *opus-mt-en-de* (https://huggingface.co/Helsinki-NLP/opus-mt-en-de) to translate from English to German, and *opus-mt-de-es* (https://huggingface.co/Helsinki-NLP/opus-mt-de-es) to translate from German to Spanish.

In Table 3, several examples of tweets translated using the single translation approach are shown. It can be observed that backtranslation maintains the final text quite similar to the original message. On the other hand, in Table 4, several examples of tweets translated using a double translation approach are presented. With this strategy, it can be observed that the final texts differ more from the original messages.

4.4. Hyperparameter optimization

Hyperparameters are adjustable parameters that allow optimizing the performance of a pre-trained model during the training phase. Some of the hyperparameters of Transformers strongly influence the validation results and, therefore, their efficiency. In a language model training process, it is crucial to assign the most suitable values to the hyperparameters. Although certain values may perform adequately for most pre-trained models, it is essential to conduct a search for the best values for each model, task, and dataset.

In our experiments, we used the Optuna optimization framework [53] to conduct a systematic search over a predefined hyperparameter space. The specific values explored, the best combinations obtained, and their quantitative impact on model performance are presented and analyzed in *Results* Subsection. This process was key to enhancing the performance of the transformer-based models.

4.5. Ensemble approaches

Ensembles in supervised classification problems leverage the strategy of combining predictions from multiple individual models to enhance predictive accuracy [54,55]. Notable ensemble techniques include hard voting, soft voting, and stacking. Hard voting involves

Table 4
Examples of tweets translated to increase the number of positive class cases using double translation.

Original tweet	English version	German version	Translated tewet
Han invadido París. Nadie sabe qué hacer. París quiere rápido desalojo de campamento de refugiados	They have invaded Paris. No one knows what to do. Paris wants quick eviction from refugee camp	Sie sind in Paris eingedrungen. Niemand weiß, was zu tun ist. Paris will schnelle Flucht aus dem Flüchtlingslager	Han entrado en París. Nadie sabe qué hacer. París quiere escapar rápidamente del campo de refugiados
De nada sirve que el CNI analice el perfil de los refugiados que vienen a España si sus hijos en 10 o 15 años pueden radicalizarse también.	It is of no use for the CNI to analyze the profile of refugees who come to Spain if their children in 10 or 15 years can also radicalize.	Es nutzt dem CNI nicht, das Profil von Flüchtlingen zu analysieren, die nach Spanien kommen, wenn ihre Kinder in 10 oder 15 Jahren auch radikalisieren können.	No sirve de nada que el CNI analice el perfil de los refugiados que llegan a España, si sus hijos también pueden radicalizar dentro de 10 o 15 años.
Alemania sufre de casos de violaciones a mujeres y niño por parte de migrantes refugiados como nunca antes se ha visto #ANREPORTAJES	Germany suffers from cases of rape of women and children by refugee migrants as never before seen #ANREPORTAJES	Deutschland leidet unter Vergewaltigung von Frauen und Kindern durch Flüchtlingsmigranten wie nie zuvor #ANREPORTAJES	Alemania sufre violaciones de mujeres y niños por inmigrantes como nunca antes #ANREPORTAJES

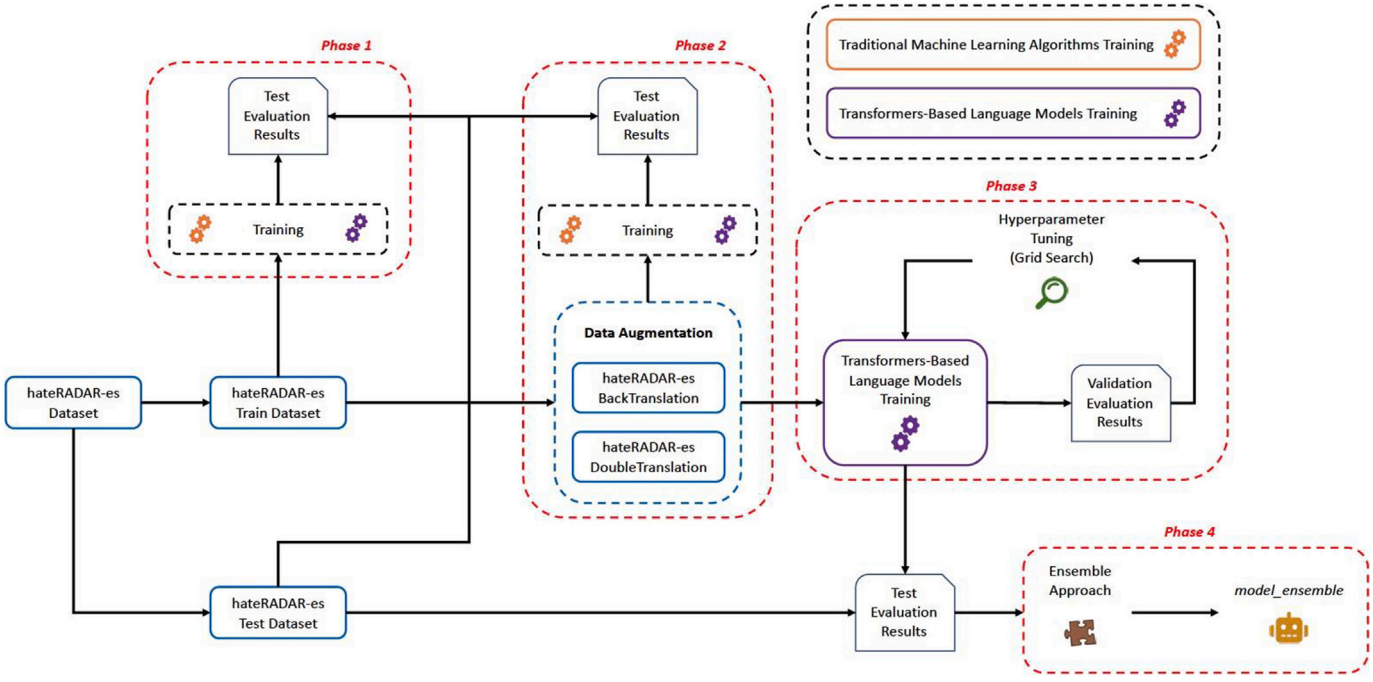

Fig. 3. Experimental Framework.

combining the predictions of individual models and selecting the class with the most votes. Soft voting, on the other hand, considers the probabilities assigned to each class by the individual models and averages them to make the final prediction. Additionally, stacking combines predictions from diverse models using a meta-learner, which learns to combine the base models' outputs effectively. These ensemble strategies have demonstrated effectiveness in improving model performance by harnessing the collective intelligence of multiple models. In this work, ensembles have been constructed using hard and soft voting techniques.

5. Experimental framework

The designed experimental framework, as can be seen in Fig. 3, consisted of a series of successive steps aimed at improving the performance of traditional machine learning algorithms and selected transformer-based language models. To evaluate the performance of the algorithms and models, the original dataset was split into a training dataset and a test dataset, with a distribution of 80% and 20%, respectively, using a stratified approach. Table 5 shows the class distribution in the training and test datasets. To provide greater robustness to the experimental framework, performance evaluation was consistently conducted using the same test dataset. In the first phase of experimentation, a baseline was designed to establish a starting point and assess the initial performance of the algorithms and models. For this purpose, the original training dataset was used, which, as described in the Subsection *III-c. Dataset Features*, is an imbalanced dataset.

Table 5
Label distribution in the training and test datasets.

Dataset	Tweets	Label 0	Label 1
Training	4000	3048	952
Test	1000	762	238

The second phase of experimentation involved the application of two data balancing techniques to the original dataset to improve the metrics reached in the baseline. In order not to reduce the number of messages, oversampling techniques were applied to increase the number of hate speech messages. The applied approaches were based on data augmentation, as described in the Subsection *Balancing Techniques*.

Two new training datasets were constructed, and the algorithms and models were re-evaluated with the test dataset.

Since the results obtained in the second phase outperformed those obtained in the baseline, we proceeded with the optimization process. As will be seen in the *Results* Subsection, the performance achieved by all the transformer-based language models exceeded that achieved by traditional machine learning models. Therefore, for the subsequent experimentation phases, only the optimization of transformer-based language models using the double back translation augmented dataset was considered. This dataset had proven to be highly effective in improving the performance of the language models in previous experiments.

In the third phase, a performance study of the models was carried out by conducting an exhaustive search for the best values of the selected hyperparameters. Finally, in the fourth phase, an ensemble approach was implemented to improve the individual results of the models. To achieve this, the best performing models were selected, and different ensemble methods were applied.

In summary, the experimental framework consisted of progressively applying various optimization approaches to traditional machine learning algorithms and pre-trained language models. The aim was to achieve improved performance in evaluation metrics at successive stages. In each phase, models with the best performance were selected until obtaining the model that achieved the best performance for the task of detecting hate speech towards refugees.

All the traditional machine learning algorithms and transformer-based language models were implemented on an NVIDIA GeForce RTX 4070 12 GB graphics card.

6. Results and analysis

In this section, the results obtained from the various experiments conducted are detailed and analyzed. Additionally, our evaluation methods are outlined, and a thorough error analysis is performed.

6.1. Results

For the evaluation of the generated models, four metrics were used: accuracy, F1-score, area under the ROC curve (AUC-ROC), and area under the PR curve (AUC-PR). Accuracy was calculated using the total sum of correct predictions across all classes; the F1-score was computed as the arithmetic mean of the F1-scores per class, which are the harmonic means of precision and recall metrics. Finally, AUC-ROC and AUC-PR reveal the classifiers' efficiency across all thresholds and are the most useful measures when addressing tasks with imbalanced datasets. For this reason, the AUC-ROC metric has been used to select the best performing models in the different phases of the experimentation.

The first experimentation, as described above, consisted of applying traditional machine learning algorithms and the selected transformer-based language models to the original dataset. For traditional algorithms, parameter values described in Appendix (Table 18) were used, whereas for transformer-based language models, since optimal hyperparameter values cannot be known beforehand, some commonly used values were employed for fine-tuning the pre-trained language models: batch size of 32, learning rate of 3e-5, maximum sequence length of 128 tokens, and weight decay of 0.01. Regarding the number of epochs, an early stopping strategy was used to prevent overfitting. For the remaining transformer-based language models parameters, we retained the default values provided by the HuggingFace Transformers library for all the models, using GELU (*Gaussian Error Linear Unit*) as activation function for hidden layers, AdamW as the optimization function and a dropout rate of 0.1. Table 6 and Table 7 show the results obtained for the different selected metrics. For the traditional machine learning algorithms, it can be observed that the combination of the One-Hot text representation with Logistic Regression yielded the best performance for all measures. Furthermore, Logistic Regression was the best performing algorithm for most of the text representations. As for transformer-based language models, they all perform better than traditional algorithms.

Once a baseline was established, the next step was to apply the two data augmentation approaches described above. For a proper comparison of the results, data augmentation was only performed on the training dataset, while the test dataset remained the same throughout all experimentation phases. In both data augmentation approaches, messages with negative content were duplicated, resulting in a new

Table 6
Results obtained by traditional machine learning algorithms on the test dataset using the original dataset. The asterisk (*) indicates the best performing algorithm for each text representation approach.

Text Representation	Algorithm	Accuracy	F1-score	AUC-ROC	AUC-PR
Count Vectorizer	RF	0.86	0.78	0.76	0.54
	XGBoost	0.86	0.79	0.77	0.55
	SVM	0.86	0.78	0.76	0.55
	KNN	0.78	0.65	0.64	0.36
	LR*	0.87	0.80	0.78	0.57
One-Hot	RF	0.86	0.79	0.77	0.56
	XGBoost	0.86	0.79	0.77	0.54
	SVM	0.86	0.78	0.75	0.54
	KNN	0.81	0.66	0.64	0.39
	LR*	**0.87**	**0.81**	**0.79**	**0.59**
TF-IDF	RF*	0.86	0.78	0.75	0.55
	XGBoost	0.84	0.75	0.72	0.48
	SVM	0.86	0.77	0.74	0.54
	KNN	0.84	0.76	0.74	0.51
	LR	0.83	0.72	0.69	0.46
Word2Vec	RF	0.82	0.65	0.63	0.40
	XGBoost*	0.85	0.77	0.75	0.53
	SVM	0.85	0.77	0.75	0.53
	KNN	0.82	0.76	0.77	0.49
	LR	0.85	0.77	0.74	0.51
Glove	RF	0.81	0.65	0.63	0.39
	XGBoost	0.81	0.71	0.69	0.43
	SVM	0.82	0.70	0.67	0.42
	KNN	0.79	0.68	0.67	0.39
	LR*	0.84	0.77	0.76	0.52

Table 7
Results obtained by transformer-based language models on the test dataset using the original dataset.

Model	Accuracy	F1-score	AUC-ROC	AUC-PR
BETO	0.898	0.858	0.858	0.667
XLM-RoBERTa	**0.902**	0.855	**0.870**	**0.680**
RoBERTa-base	0.897	**0.861**	0.869	0.669
RoBERTa-large	0.901	0.857	0.844	0.672

Table 8
Results obtained by traditional machine learning algorithms on the test dataset using the augmented dataset with a single back-translation.

Text Representation	Algorithm	Accuracy	F1-score	AUC-ROC	AUC-PR
One-Hot	RF	0.86	0.80	0.79	0.56
	XGBoost	0.85	0.79	0.79	0.53
	SVM	0.87	0.82	0.81	0.59
	KNN	0.77	0.70	0.71	0.40
	LR	0.87	0.82	0.81	0.59

Table 9
Results obtained by transformer-based language models on the test dataset using the augmented dataset with a single back-translation.

Model	Accuracy	F1-score	AUC-ROC	AUC-PR
BETO	0.891	0.853	0.862	0.653
XLM-RoBERTa	0.851	0.815	0.854	0.583
RoBERTa-base	0.876	0.837	0.856	0.621
RoBERTa-large	0.902	0.866	0.869	0.680

Table 10
Results obtained by traditional machine learning algorithms on the test dataset using the augmented dataset with a double back-translation.

Text Representation	Algorithm	Accuracy	F1-score	AUC-ROC	AUC-PR
One-Hot	RF	0.85	0.80	0.79	0.55
	XGBoost	0.85	0.79	0.79	0.54
	SVM	0.87	0.82	0.81	0.59
	KNN	0.62	0.60	0.70	0.35
	LR	0.87	0.82	0.81	0.59

Table 11
Results obtained by transformer-based language models on the test dataset using the augmented dataset with a double back-translation.

Model	Accuracy	F1-score	AUC-ROC	AUC-PR
BETO	0.889	0.851	0.860	0.648
XLM-RoBERTa	0.883	0.845	0.861	0.636
RoBERTa-base	0.902	0.864	0.864	0.679
RoBERTa-large	0.889	0.852	0.865	0.650

Table 12
Accuracy (in percentage) of the minority class achieved by transformer-based language models on the test dataset using the original dataset, the augmented dataset with a single back-translation and the augmented dataset with a double back-translation.

Model	Original	Single back-translation	Double back-translation
BETO	78.2%	80.7%	80.7%
XLM-RoBERTa	80.7%	80.7%	82.0%
RoBERTa-base	81.6%	82.0%	79.5%
RoBERTa-large	73.6%	80.7%	82.0%

Table 13
Hyperparameter space.

Hyperparameter	Values
Batch Size	[8, 16, 32, 64]
Learning Rate	[2e−5, 3e−5, 5e-5]
Weight Decay	[0.001, 0.01, 0.1]
Optimizer	[“adamw_hf“, “adamw_torch”, “adafactor"]

training dataset with 4952 messages and a class distribution of 3048 (61.5%) for class “0” and 1904 (38.5%) for class “1”. The results achieved using the data augmentation approach with a single translation are shown in Table 8 and Table 9, while in Table 10 and Table 11, the results achieved with double back translation are shown. Analyzing the values of the metrics shown in these tables, it is observed that the values of Accuracy, F1-score, AUC-ROC, and AUC-PR are very similar when using the original (imbalanced) dataset compared to the results when using the augmented datasets. This is because the overall accuracy and error rates (both classes combined) remain relatively stable, as expected. However, when data augmentation techniques have been implemented, a higher rate of accuracy has been achieved for the minority class. Specifically in this task, it is more important to have a higher accuracy rate in detecting the minority class in order to detect as many hate speech messages as possible. Table 12 shows the accuracy rates of class “1” in the test dataset for the four language models. It is noteworthy the increase in accuracy rate when employing data augmentation approaches with respect to the results achieved with the original dataset.

From this experimental stage onwards, only transformer-based language models were considered since they achieved better performance than traditional machine learning algorithms. In this third phase, a search for the best values of the language models hyperparameters was carried out. Table 13 shows the hyperparameter space used in the experimentation phase. For the remaining transformer-based language models parameters, we retained the default values provided by the HuggingFace Transformers library for all the models, using *GELU* (Gaussian Error Linear Unit) as activation function for hidden layers,

Table 14
The three best combinations of hyperparameter values for each language model. The search was conducted by optimizing the AUC-ROC measure.

Model		Batch size	Learning rate	Weight decay	Optimizer
BETO	hp_1	8	3e−5	0.1	adamw_hf
	hp_2	8	3e−5	0.001	adamw_torch
	hp_3	16	3e−5	0.1	adafactor
XLM-RoBERTa	hp_1	64	2e−5	0.001	adamw_torch
	hp_2	32	2e−5	0.1	adafactor
	hp_3	32	5e−5	0.001	adafactor
RoBERTa-base	hp_1	32	2e−5	0.01	adafactor
	hp_2	64	3e−5	0.001	adafactor
	hp_3	32	5e−5	0.01	adamw_torch
RoBERTa-large	hp_1	64	2e−5	0.001	adafactor
	hp_2	64	2e−5	0.1	adafactor
	hp_3	64	2e−5	0.01	adamw_hf

Table 15
AUC-ROC values achieved during the hyperparameter optimization process on the test dataset.

Model	Double back-translation	hp_1	hp_2	hp_3
BETO	0.860	0.861	0.867	0.872
XLM-RoBERTa	0.861	0.862	0.870	0.880
RoBERTa-base	0.864	0.866	0.864	0.879
RoBERTa-large	0.865	0.865	0.865	0.879

and a *dropout* rate of 0.1. Due to the nature of the messages in the dataset, a *max_length* of 128 tokens was used. To perform hyperparameter optimization, the double back-translation dataset was used. A grid search was performed by testing all possible combinations of the hyperparameter space. Optuna hyperparameter optimization software framework was employed to find the optimal values of the hyperparameters for each model. In the search process, the number of epochs was set to 20 with an early stopping patience of 3 in all cases. A total of 108 runs were performed for each model, resulting in a total of 432 runs. Approximately 20 h were spent on this search process. The three best combinations of hyperparameter values for each language model are shown in Table 14. The full training dataset was used for the search of these values, and the AUC-ROC measure was selected to optimize the performance of the models.

The results obtained in this phase of experimentation demonstrate that the search for the best hyperparameter values is a very relevant and significant process for improving the performance of transformer-based language models in hate speech detection tasks. Table 15 shows the results reached, for the AUC-ROC metric, by the models fine-tuned with the three best combinations of hyperparameter values. It can be observed that the models trained using the combinations of the best hyperparameter values achieve better performance than those trained with default hyperparameter values. It is important to highlight that, in this optimization process, all evaluation metric values achieved by the models with the three best combinations of hyperparameter values on the validation dataset were significantly higher than the values of the models fine-tuned with the double back-translation approach using default hyperparameter values. However, as can be seen in Table 15, the performance on the test dataset, although also higher, is less pronounced. The main reason is that the number of class “1” tweets in the test dataset is smaller, and furthermore, the models have already achieved optimal performance in previous phases. Nevertheless, it can be concluded that the models perform better with a proper choice of hyperparameter values.

The final stage of the experimental framework consisted of performing an ensemble approach with the predictions of the models that performed best in the previous phases. The aim was to further improve the prediction by leveraging the predictive capabilities of each model. To construct these ensembles, we used the best-performing version of

Table 16
Results obtained by ensembles on the test dataset.

Model	Accuracy	F1-score	AUC-ROC	AUC-PR
BETO (hp_3) + XLM-RoBERTa (hp_3) + RoBERTa-large (hp_3)	0.891	0.860	0.892	0.666
BETO (hp_3) + XLM-RoBERTa (hp_3) + RoBERTa-base (hp_3)	0.888	0.856	0.887	0.658

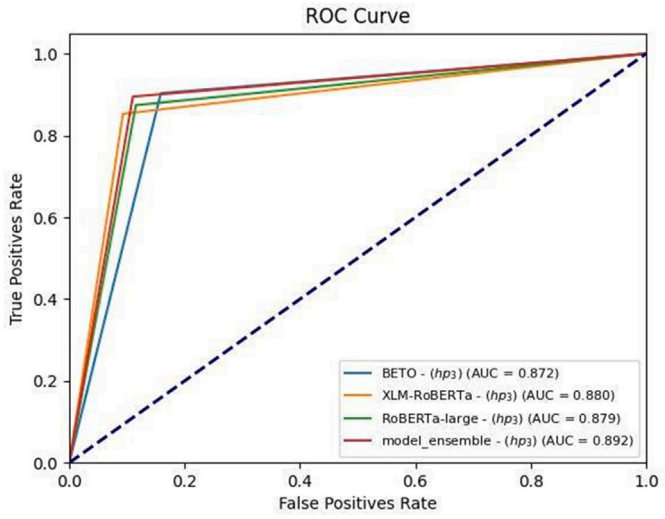

Fig. 4. ROC Curve for the three best models and the ensemble model on the test dataset.

each of the four fine-tuned models, corresponding to their hp_3 hyperparameter configuration (as shown in Table 15). We then evaluated all possible combinations of three models and selected the two ensembles that yielded the highest performance on the test set. Since a hard voting approach was used, ensembles with an odd number of models were constructed to avoid ties. Table 16 shows the results of the two ensembles that achieved the best performance in predicting offensive messages towards refugees. The prediction values of the ensembles were higher than the values obtained by the individual models. It can be concluded that the experimental framework has been successful and that very considerable prediction values have been achieved. Fig. 4 shows the ROC curve of the three best performing models and the ensemble model.

These results are consistent with, and in some cases exceed, the performance levels reported in recent studies on hate speech detection in other languages. Although direct comparisons are difficult due to differences in datasets and label definitions, our ensemble's AUC-ROC of 0.892 and F1-score of 0.860 demonstrate that transformer-based models achieve competitive results in Spanish-language hate speech detection tasks.

6.2. Error analysis

The previous section outlined the results achieved by the models across various metrics. While they have demonstrated commendable performance, it is imperative to acknowledge that these models are not immune to making occasional prediction errors. To delve deeper into the factors contributing to these discrepancies, an exhaustive analysis of the model's mistakes was undertaken. For this study, the model that performed best across all selected metrics was used. Specifically, it was the ensemble composed of the models BETO (hp_3) + XLM-RoBERTa (hp_3) + RoBERTa-large (hp_3). Henceforth, we will refer to this model as *model_ensemble*.

From a quantitative perspective, it can be concluded that the models obtained during the learning process have achieved very good performance in detecting tweets with hateful content towards the refugees community, despite the scarcity of such cases in the training dataset. Fig. 5 shows the confusion matrices of the *model_ensemble* and the models comprising the final ensemble, evaluating the results on the test dataset. All three models achieved over 85% accuracy for the minority class. Even the BETO model correctly classifies 90% of the hate speech tweets (215 out of 238) although, on the contrary, it achieves a lower accuracy rate for the majority class (121 errors in 762 cases). It is noteworthy that the ensemble approach has also been successful in reducing and balancing the error rate for both classes classifying correctly classified 89.5% (213 out of 238) of the tweets from the minority class, and 89% (678 out of 762) from the majority class.

For the qualitative study of error analysis, the results achieved by the *model_ensemble* were used. The expert annotators manually inspected the tweets that were misclassified by the model and Table 17 compiles a selection of them. In tweet 1, the classifier was not able to detect hate speech towards refugees as it is a tweet with very subtle hate content. The hatred is mainly directed at *Macri* (Mauricio Macri, president of Argentina from 2015 to 2019) and not at refugees, but at the same time, it supports a classic discriminatory idea in migration studies that a country should prioritize the needs of nationals over foreigners, ignoring the commitment to international asylum agreements that apply in this case to refugees. The model seems to have missed this subtlety. Tweet 2 is well labeled as it captures the classic stereotype of the immigrant as a profiteer of the system and a burden on nationals. However, the text of the tweet does not contain a hateful message for the model to classify it correctly. Recalling the definition of hate speech used in the annotation, based on [5,10–12], hate speech or communications include expressions such as spreading suspicion of refugees or association with threats of different kinds, such as those of an economic nature or affecting welfare. Therefore, it is not surprising that tweet number 2 in Table 17 is classified as hate speech in the annotation process because it introduces the author's dislike and suspicion that refugees threaten economic security and help them be wasteful for the state [@user 900 euros pension x refugees plus 900 euros rent support, plus employment support, and we will all pay for that]. Among the interpretations generated by the tweet, which makes it difficult to classify, is that it indirectly suggests welfare chauvinism [56], or the preference of nationals over foreigners.

Regarding tweet number 3, it is also well labeled as it subscribes to both the theory that refugees come to invade and the idea that they are not genuine refugees. Since the word "invasion" does not appear in full, the model has not been able to understand this hate speech.

In the case of false positives, tweet number 4 is a difficult tweet to classify because it compares the behavior of refugees with that of other populations. The tweet judges the refugees by generalizing and indicates that they are all rapists. This is hate speech. In contrast, when referring to other sexual aggressors who are different to refugees, it does not generalize and emphasizes that in this case this behavior is an isolated case. Due to the complex comparison between diverse groups, this tweet is very difficult to classify automatically.

Through this detailed study of some of the errors made by the models, it can be concluded that transformer-based language models have a great capacity to address NLP tasks such as text categorization. However, specific techniques are sometimes needed to solve problems inherent to human language, such as irony, ambiguity, or lack of context. As the field of NLP continues to evolve, it is important to develop more advanced techniques that can help language models overcome these challenges and process human language more accurately.

7. Conclusions

This study has significantly contributed to the field of NLP by developing models capable of identifying hate speech directed towards

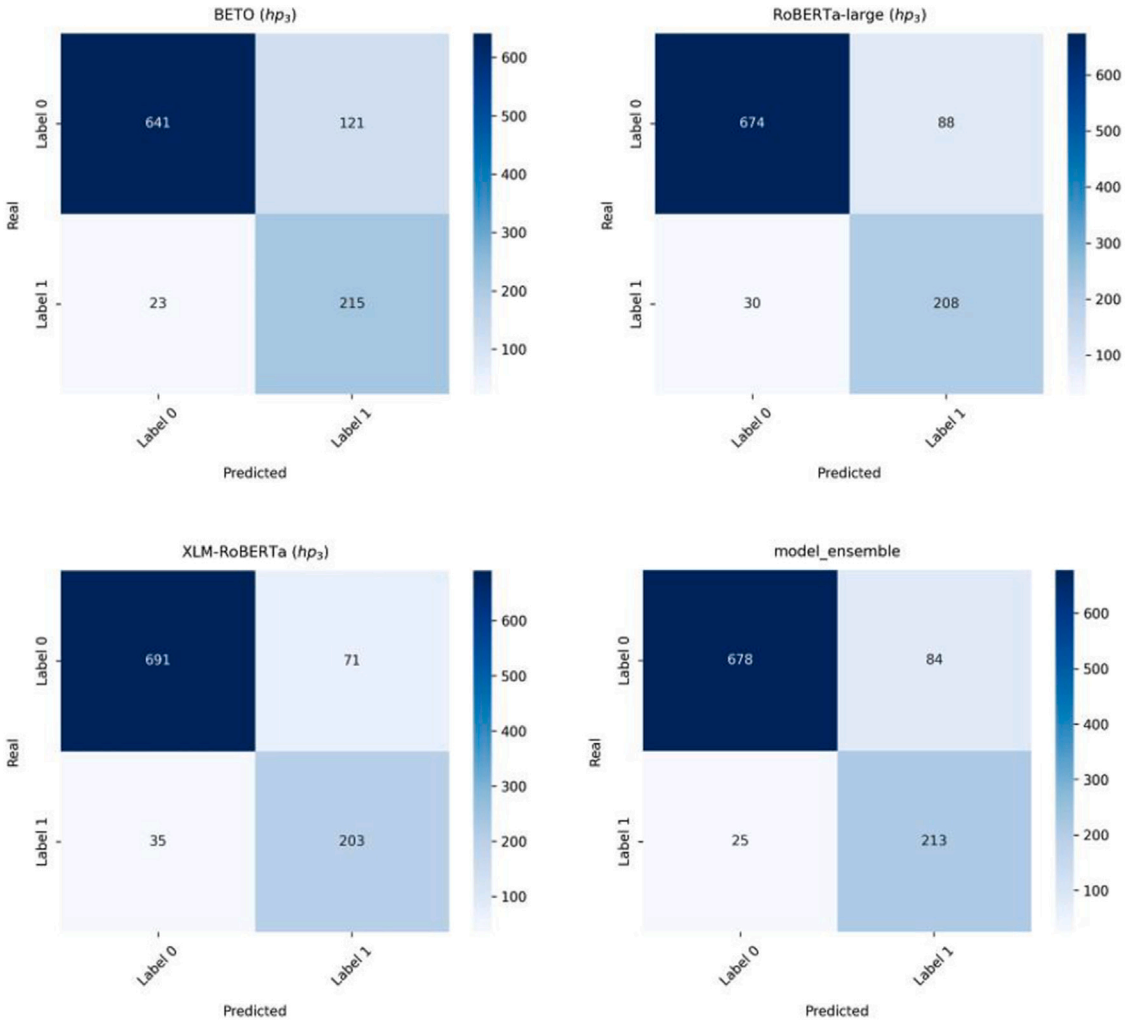

Fig. 5. Confusion matrices of the three best performing models and the ensemble model on the test dataset.

Table 17
Tweets with the manual gold label ('Observed') and the label predicted by the model.

#	Tweet (sp)	Tweet (en)	Observed	Predicted
1	macri no podes mantener un pais y qres recibir refugiados sos un capo!!!!!!	macri you can't maintain a country and you wnt [want] to receive refugees you are a capo!!!!!!	1	0
2	@user 900 euros de pensión x refugiados más 900 de ayuda al alquiler, más ayuda al empleo, y eso lo pagaremos todos	@user 900 euros pension per refugee plus 900 euros rent support, plus employment support, and we will all pay for that	1	0
3	¡calla, demagoga! no es q,algunos, estemos en contra de los refugiados,si lo fueran realmente,es contra los q invad...	shut up, demagogue! it's not t [that], some of us, are against refugees, if they really were, it's against those that invad...	1	0
4	@user enemigo yo? que exageración y en verdad deberias ayudar a los refugiados en ves de decir que te invaden ganarias amistades.	@user me, an enemy? what an exaggeration and you should really help the refugees instead of saying they invade you you would win friends.	0	1
5	entre los refugiados hay violadores –> todos son violadores uno de los agresores sexuales de pamplona era guardia civil –> caso aislado	among refugees there are rapists –> they are all rapists one of the sex offenders in pamplona was a Civil Guard –> isolated case	0	1

refugees in Spanish social media texts. The *HateRADAR-es* dataset, a corpus of 5000 Spanish tweets meticulously labeled by sociologist and social workers, has proven to be a valuable resource for training and evaluating the machine learning and deep learning models proposed. The results obtained indicate that language models based on pre-trained Transformer, especially those based on BERT and RoBERTa, outperform traditional machine learning methods in hate speech detection, providing superior accuracy and generalization ability. We

particularly highlight the performance of an ensemble of transformer-based language models, which achieved an accuracy of 0.891, an F1-score of 0.860, and an AUC-ROC of 0.892. These results not only demonstrate its overall effectiveness but also its ability to correctly identify positive cases of hate speech. These outcomes attest to the potential of ensembles to handle the subtleties and complexities of language in emotionally charged and biased contexts.

The significance of this work extends beyond its technical contributions. By focusing on anti-refugee hate speech in the Spanish language, we address a significant gap in current research, which predominantly concentrates on English-language texts and non-specific generalizations about groups. This approach allows for a better understanding of the dynamics of hate in specific linguistic and cultural contexts, which is crucial for developing more effective interventions against hate speech in society.

According to the problems generated by the spread of hate speech online, it is clear that in order to achieve successful social integration or inclusion of refugees in ethnically and culturally diverse societies, it is critical to stop the spread of hate speech in social networks towards refugees. As [57] point out, the deployment of online hate speech against migrants and refugees is an obstacle to the successful acculturation and well-being of asylum seekers. Likewise, hate speech towards vulnerable people such as migrants or asylum seekers poses a severe challenge and problem for coexistence and democracy in Europe, as well as a risk concerning existing social diversity [6,18], which makes the need to advance in the detection of online hate speech towards refugees and other population segments more pressing.

7.1. Future work and limitations

In future works, our interest lies in deepening the understanding of hate speech directed at refugees by incorporating additional nuances and analyses. To achieve this, we plan to introduce new labels that reflect specific dimensions of hate speech. One of these dimensions will be the target of the hateful message, such as religious identity, nationality, or gender. Another key dimension will be the narrative type expressed in the tweet, distinguishing between forms such as accusation, threat, conspiracy theory, or cultural racism. These additions will not only enhance the granularity of the dataset but also support the development of more nuanced classification models and sociological interpretations.

We also plan to explore the integration of large language models (LLMs), leveraging zero-shot and few-shot prompting strategies as well as fine-tuning to improve hate speech detection. Several recent studies have begun to explore this line of research. In the work by Pan et al. [58], different prompting strategies — including fine-tuning, zero-shot, and few-shot configurations — are systematically compared using LLMs on English hate speech tasks, showing that few-shot prompting can match or come close to the performance of fine-tuned models. In the study by Choudhary et al. [59], GPT-2–based LLMs are evaluated using fine-tuning and multi-shot prompting, with results indicating significant improvements over zero-shot configurations. Finally, Hashir et al. [60] present TARGE, a framework in which LLM-generated rationales support hate speech classification, leading to improvements in both interpretability and accuracy. Exploring how these prompting-based LLM methods perform using our current dataset and its future expansions will allow us to assess potential improvements in generalizability, contextual reasoning, and robustness in detecting hate speech against refugees in Spanish.

Additionally, we intend to conduct a comparative study between our specialized classifiers and established general-purpose classifiers, such as Detoxify [61], to better understand the strengths and limitations of each approach in detecting hate speech in Spanish, particularly towards refugees. This comparison will help to position our work within the broader context of hate speech detection research and may reveal opportunities for hybrid approaches that leverage the advantages of both specialized and general classifiers.

Another direction for future research involves exploring the adaptation of our classifiers to other languages, particularly within the Latin American context. This could involve leveraging transfer learning techniques to adapt models trained on Spanish data to closely related languages, thus contributing to the global effort in combating hate speech online. Given the scarcity of resources in languages other than English, we also plan to explore the adaptation of our classifiers to other languages, particularly within the Latin American context, Portuguese and Italian. This could involve leveraging transfer learning techniques to adapt models trained on Spanish data to closely related languages, thus contributing to the global effort in combating hate speech online. Given the scarcity of these types of resources in languages other than English, this approach is particularly valuable. However, we are fully aware that this adaptation is not a simple process and presents significant challenges that must be addressed. Some key challenges to consider in our future steps are linguistic and grammatical in nature.

While Spanish, Portuguese, and Italian share Latin roots, they are not identical languages. Therefore, we must consider several differences that could affect the performance of transferred models. For example, we need to implement strategies to address false friends and differences or nuances in vocabulary, variations in syntax and grammatical structures, and idiomatic expressions and colloquialisms that operate at a local level and are not easily transferable from one language to another. Regarding cultural and social challenges, we must recognize that while there are internationally common elements in hate speech concerning refugees (for example, narratives using the “invasion metaphor”, or the persistent association of refugees from Arab countries with terrorism), hate speech is a phenomenon deeply rooted in social and cultural contexts. This means that a purely linguistic approach would be incomplete. For this reason, among our future steps, it is essential to examine and study, for instance, in collaboration with other members of our international team, cultural biases and stereotypes (so that our models do not replicate them automatically) as well as social and pragmatic norms.

Finally, this work underscores the need for interdisciplinary collaborations involving both technologists and social science experts to ensure that advances in NLP are applied in ways that respect and promote ethical values and human rights. Together, we can use technology not only to better understand hate speech but also to combat it effectively and sustainably.

Despite the encouraging results and contributions presented in this study, several limitations should be acknowledged.

- **Corpus size and class imbalance.** The size and class distribution of the HateRADAR-es dataset are the result of a deliberate methodological choice to prioritize annotation quality. Tweets were manually labeled by domain experts following a rigorous and controlled process, which inevitably limited the dataset size. Additionally, the observed class imbalance was retained to faithfully reflect the real-world distribution of hate speech towards refugees on social media. While these choices enhance the validity and realism of the dataset, they may also pose challenges for generalization and model performance in less controlled environments. Future work will explore scalable annotation strategies and additional data sources to expand the corpus while preserving label quality.
- **Temporal coverage of data.** The HateRADAR-es dataset consists of tweets collected between December 2015 and December 2016. While this period corresponds to a critical phase of the international refugee crisis, the dataset does not capture more recent developments in public discourse, migration dynamics, or global political events. We acknowledge this temporal limitation and plan to expand the dataset in future work by incorporating more recent content, in order to better reflect current narratives and emerging trends related to refugees.

- **Annotation scalability.** The annotation process in this study was conducted entirely by domain experts to ensure conceptual accuracy and reliability. While this guarantees high-quality labels, it limits the scalability of the dataset. In future work, hybrid strategies combining crowdsourcing with expert validation or multi-level annotation protocols may be considered to expand the dataset while maintaining annotation consistency.

7.2. Ethical implications

The development and deployment of hate speech detection models in sensitive domains such as anti-refugee discourse carry important ethical considerations. The creation of any artificial intelligence tool in this area is not a neutral process; design decisions, from data selection to results evaluation, have profound social and human implications.

7.2.1. Mitigating bias in training data

One of the primary ethical concerns in the development of natural language processing models is bias in the training data. Data can reflect and amplify existing social prejudices, potentially leading to systematic discrimination by the model. To proactively address this risk, several strategies were implemented during the creation of the HateRADAR-es dataset:

- **Rigorous Conceptual Foundation**. The annotation guide was built upon international definitions of hate speech, provided by organizations such as the European Commission and the United Nations. This ensured the conceptualization of hate was consistent and not based on subjective interpretations.
- **Expert Annotators.** Our annotators were not inexperienced personnel; they were experts in sociology and social work. Their background provided a deep understanding of social context, jargon, and the dynamics of discrimination and social exclusion, which was fundamental to discerning between legitimate criticism and genuine hate speech, thereby minimizing the risk of inadvertent biases.
- **Specialized Training.** Training was key for achieving high-quality annotations. It focused on providing clear information on how to consistently identify hate speech, how to use the annotation guide, and how to use the Doccano tool. The process included practical exercises to classify tweets and discuss doubts before the pre-test. This stage was crucial for standardizing the process and ensuring all annotators applied the same criteria and understood the nuances of hate speech towards refugees.
- **Tool and Guideline Validation.** Before beginning the final annotation, a pre-test was conducted on a sample of tweets to validate that the Doccano tool worked correctly. As a result of this test, minor adjustments were made to the annotation guidelines to ensure all rules were robust and complete.

  Despite these measures, we acknowledge that no dataset is entirely free from cultural or contextual bias. Transparency about the methodology used to create the dataset is, therefore, a crucial first step towards accountability.

7.2.2. Consequences of misclassification

Misclassification of messages is another significant ethical challenge. Model errors, whether false positives or false negatives, can have a direct and harmful impact.

- **Risk of False Positives.** A false positive (incorrectly labeling a message as hate speech) could unfairly silence users or restrict legitimate discourse. In the context of refugees, this could lead to the censorship of individuals sharing personal experiences or expressing critical opinions on immigration policies without inciting hatred.
- **Risk of False Negatives.** Conversely, a false negative (failing to detect hate speech) allows harmful content to circulate unchecked. This is especially dangerous for vulnerable populations, as hate speech can incite violence, harassment, and discrimination, severely impacting their psychological well-being and safety.

While our model demonstrated solid performance, we are aware that no system is infallible. These challenges highlight the need for future research to focus on continuous evaluation of model performance, especially in specific contexts and for different demographic subgroups, to mitigate these risks.

7.2.3. Future directions for ethical and responsible deployment

Hate speech detection technology should be viewed as part of a broader, human-centered strategy, not as a standalone solution. Addressing the ethical implications requires ongoing commitment to the following areas:

- **Transparency and Auditability.** Models should be transparent so that their decisions can be audited and understood. Decisions about how classification thresholds are implemented and how errors are handled are not merely technical choices but ethical ones.
- **Interdisciplinary Collaboration.** The development and deployment of these tools must foster collaboration between engineers, data scientists, sociologists, social workers, legal experts, and human rights experts. This synergy ensures that technology is built and deployed in an informed manner, balancing protection against harm with respect for rights and freedom of expression.
- **Accountability.** The entities that implement these models must be accountable for their results and their impact on society. The ultimate goal must be to promote inclusion and protect human rights, using technology as a tool for that purpose, and not as a substitute for human intervention or effective social policies.

CRediT authorship contribution statement

Jacinto Mata: Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Estrella Gualda:** Writing – review & editing, Writing – original draft, Validation, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Victoria Pachón:** Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **Carolina Rebollo:** Writing – review & editing, Writing – original draft, Resources, Methodology, Investigation, Formal analysis, Data curation. **Juan L. Domínguez:** Writing – review & editing, Writing – original draft, Validation, Software, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This paper is part of the I+D+i Project titled “Conspiracy Theories and Hate Speech Online: Comparison of patterns in narratives and social networks about COVID 19, immigrants, refugees and LGBTI people [NON CONSPIRA HATE!]", PID2021 123983OB I00, funded by MCIN/AEI/10.13039/501100011033/ and by FEDER/EU.

The publication is part of grant JDC2022-048239-I, funded by MCIN/AEI/10.13039/501100011033 and by the European Union “NextGenerationEU”/PRTR.

We also thank for the support of the research centers at the University of Huelva "Estudios Sociales E Intervención Social, ESEIS", "Pensamiento Contemporáneo e Innovación para el Desarrollo Social, COIDESO" and "Centro de Investigación en Tecnología, Energía *y* Sostenibilidad, CITES".

Appendix

See Table 18.

Table 18
Parameter values used for traditional algorithms.

Algorithm	Parameter values
RF	bootstrap: True; ccp_alpha: 0.0; class_weight: None criterion: gini; max_depth: None; max_features: sqrt max_leaf_nodes: None; max_samples: None; min_impurity_decrease: 0.0 min_samples_leaf: 1; min_samples_split: 2; min_weight_fraction_leaf: 0.0 n_estimators: 100; n_jobs: None; oob_score: False random_state: None; warm_start: False
XGBoost	objective: binary:logistic; callbacks: None; early_stopping_rounds: None gamma: None; grow_policy: None; importance_type: None interaction_constraints: None; learning_rate: None; max_bin: None max_delta_step: None; max_depth: None; max_leaves: None min_child_weight: None; n_estimators: None; n_jobs: None reg_alpha: None; reg_lambda: None; sampling_method: None scale_pos_weight: None; tree_method: None
SVM	C: 1.0; break_ties: False; cache_size: 200 class_weight: None; coef0: 0.0; decision_function_shape: ovr degree: 3; gamma: scale; kernel: rbf max_iter: −1; probability: False; random_state: None shrinking: True; tol: 0.001
KNN	algorithm: auto; leaf_size: 30; metric: minkowski metric_params: None; n_jobs: None; n_neighbors: 5 p: 2; weights: uniform
LR	C: 1.0; class_weight: None; dual: False fit_intercept: True; intercept_scaling: 1; l1_ratio: None max_iter: 100; multi_class: auto; n_jobs: None penalty: l2; random_state: None; solver: lbfgs tol: 0.0001; warm_start: False

Data availability

https://doi.org/10.5281/zenodo.17259982.

References

[1] Ghosh D. The European union's response to rising xeno-racism in Europe: An assessment. Can J Eur Russ Stud 2022;15(1):1–23. http://dx.doi.org/10.22215/cjers.v15i1.2815.

[2] Hinz T, Walzenbach S, Laufer J, Weeber F. Media coverage, fake news, and the diffusion of xenophobic violence: A fine-grained county-level analysis of the geographic and temporal patterns of arson attacks during the german refugee crisis 2015–2017. PLoS One 2023;18(7):1–23. http://dx.doi.org/10.1371/journal.pone.0288645.

[3] González-Baquero W, Amores JJ, Arcila-Calderón C. The conversation around islam on Twitter: Topic modeling and sentiment analysis of tweets about the muslim community in Spain since 2015. Religions 2023;14(6). http://dx.doi.org/10.3390/rel14060724, URL: https://www.mdpi.com/2077-1444/14/6/724.

[4] UNHCR. Global trends report 2022. 2022, URL: https://www.unhcr.org/global-trends-report-2022. [Accessed 25 July 2023].

[5] European-Commission. Annual report on ECRI's activities covering the period from 1 January to 31 December 2020. Technical report, European Commission against Racism and Intolerance; 2021, URL: https://rm.coe.int/annual-report-on-ecri-s-activities-for-2020/1680a1cd59.

[6] Arcila-Calderón C, Sánchez-Holgado P, Quintana-Moreno C, Amores JJ, Blanco-Herrero D. Hate speech and social acceptance of migrants in Europe: analysis of tweets with geolocation. Comunicar 2022;30(71):21–35. http://dx.doi.org/10.3916/C71-2022-02.

[7] Gualda E, Rebollo C. The refugee crisis on Twitter: A diversity of discourses at a European crossroads. J Spat Organ Dyn 2016;4(3):199–212, URL: http://hdl.handle.net/10272/13624.

[8] Rebollo C. Tuiteando sobre refugiados : una comparación internacional de discursos, imaginarios y representaciones sociales [Ph.D. thesis], Universidad de Huelva; 2021, URL: http://hdl.handle.net/10272/20113.

[9] Weber A. Manual on hate speech. Council of Europe; 2009, URL: https://book.coe.int/en/human-rights-and-democracy/4197-manual-on-hate-speech.html.

[10] United-Nations. Understanding hate speech. 2021, URL: https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech.

[11] European-Commission. ECRI general policy recommendation n° 15 on combating hate speech, adopted on 8 December 2015. Technical report, European Commission against Racism and Intolerance; 2016, URL: https://www.coe.int/en/web/european-commission-against-racism-and-intolerance/hate-speech-and-violence.

[12] European-Commission. Hate speech and violence. Technical report, European Commission against Racism and Intolerance; 2023, URL: https://www.coe.int/en/web/european-commission-against-racism-and-intolerance/hate-speech-and-violence.

[13] European-Commission. Combating anti-Muslim hatred. Activities of the European commission's coordinator on combating anti-Muslim hatred. Technical report, European Commission; 2024, URL: https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/combating-anti-muslim-hatred_en.

[14] Velikonja M. "New others": Ideological images of refugees in present-day Slovenia. Soc Text 2022;40(2 (151)):49–68. http://dx.doi.org/10.1215/01642472-9631131, arXiv:https://read.dukeupress.edu/social-text/article-pdf/40/2(151)/49/1587733/49velikonja.pdf.

[15] Lotero-Echeverri G, Romero-Rodriguez L. Journalism on forced migration in latin america: Recommendations from experts and international journalism guides from a qualitative study. Qual Rep 2023;28(5). http://dx.doi.org/10.46743/2160-3715/2023.6061.

[16] Nguyen TT, Yu W, Merchant JS, Criss S, Kennedy CJ, Mane H, Gowda KN, Kim M, Belani R, Blanco CF, Kalachagari M, Yue X. Examining exposure to messaging, content, and hate speech from partisan news social media posts on racial and ethnic health disparities. Int J Environ Res Public Heal 2023;20(4):1–13. http://dx.doi.org/10.3390/ijerph20043230, URL: https://www.mdpi.com/1660-4601/20/4/3230.

[17] Chimkhlai P, Panyametheekul S. Hate speech in YouTube comments on rohingya refugees in Thailand and Syrian refugees in Europe. LEARN J: Lang Educ Acquis Res Netw 2024;17(1):133–61, URL: https://so04.tci-thaijo.org/index.php/LEARN/article/view/270379.

[18] Sánchez-Holgado P, Amores JJ, Blanco-Herrero D. Online hate speech and immigration acceptance: A study of Spanish provinces. Soc Sci 2022;11(11). http://dx.doi.org/10.3390/socsci11110515, URL: https://www.mdpi.com/2076-0760/11/11/515.

[19] Camargo J, Cogo D, Alencar A. Venezuelan refugees in Brazil: Communication rights and digital inequalities during the Covid-19 pandemic. Media Commun 2022;10(2):230–340. http://dx.doi.org/10.17645/mac.v10i2.5051, URL: https://www.cogitatiopress.com/mediaandcommunication/article/view/5051.

[20] García-Baena D, García-Cumbreras MÁ, Jiménez-Zafra SM, García-Díaz JA, Valencia-García R. Hope speech detection in spanish: The LGBT case. Lang Resour Eval 2023;57(4):1487–514. http://dx.doi.org/10.1007/s10579-023-09638-3.

[21] Mohdeb D, Laifa M, Zerargui F, Benzaoui O. Evaluating transfer learning approach for detecting Arabic anti-refugee/migrant speech on social media. Aslib J Inf Manag 2022;74(6):1070–88.

[22] Jahan MS, Oussalah M. A systematic review of hate speech automatic detection using natural language processing. Neurocomputing 2023;546:126232. http://dx.doi.org/10.1016/j.neucom.2023.126232, URL: https://www.sciencedirect.com/science/article/pii/S0925231223003557.

[23] Gudumotu CE, Nukala SR, Reddy K, Konduri A, Gireesh C. A survey on deep learning models to detect hate speech and bullying in social media. In: Biswas A, Semwal VB, Singh D, editors. Artificial intelligence for societal issues. Artificial Intelligence for Societal Issues; 2023, p. 27–44. http://dx.doi.org/10.1007/978-3-031-12419-8_2.

[24] Mansur Z, Omar N, Tiun S. Twitter hate speech detection: A systematic review of methods, taxonomy analysis, challenges, and opportunities. IEEE Access 2023;11:16226–49. http://dx.doi.org/10.1109/ACCESS.2023.3239375.

[25] Li Q, Peng H, Li J, Xia C, Yang R, Sun L, Yu PS, He L. A survey on text classification: From traditional to deep learning. ACM Trans Intell Syst Technol 2022;13(2). http://dx.doi.org/10.1145/3495162.

[26] Garrido-Merchan EC, Gozalo-Brizuela R, Gonzalez-Carvajal S. Comparing BERT against traditional machine learning models in text classification. J Comput Cogn Eng 2023;2(4):352–6. http://dx.doi.org/10.47852/bonviewJCCE3202838, URL: https://ojs.bonviewpress.com/index.php/JCCE/article/view/838.

[27] Arcila-Calderón C, Amores JJ, Sánchez-Holgado P, Vrysis L, Vryzas N, Oller Alonso M. How to detect online hate towards migrants and refugees? Developing and evaluating a classifier of racist and xenophobic hate speech using shallow and deep learning. Sustainability 2022;14(20). http://dx.doi.org/10.3390/su142013094, URL: https://www.mdpi.com/2071-1050/14/20/13094.

[28] Benítez-Andrades JA, González-Jiménez Á, López-Brea Á, Aveleira-Mata J, Alija-Pérez JM, García-Ordás MT. Detecting racism and xenophobia using deep learning models on Twitter data: CNN, LSTM and BERT. PeerJ Comput Sci 2022;8:e906.

[29] Fleming M. Declaración del Alto Comisionado para los Refugiados sobre la crisis de refugiados en Europa. Technical report, UNHCR/ACNUR España; 2015, URL: https://www.acnur.org/es-es/noticias/comunicados-de-prensa/declaracion-del-alto-comisionado-para-los-refugiados-sobre-la-crisis.
[30] Loper E, Bird S. NLTK: the natural language toolkit. 2002, CoRR, cs.CL/0205028. URL: https://arxiv.org/abs/cs/0205028.
[31] Gualda E, Rebollo Díaz C. Guía de anotación – Libro de códigos para la detección del discurso de odio hacia inmigrantes y refugiados, versión 3. 2024, Universidad de Huelva. Proyecto PID2021-123983OB-I00 [NONCONSPIRAHATE!], URL: https://rabida.uhu.es/dspace/handle/10272/23340.
[32] Artstein R. Inter-annotator agreement. In: Ide N, Pustejovsky J, editors. Handbook of linguistic annotation. Springer Netherlands; 2017, p. 297–313. http://dx.doi.org/10.1007/978-94-024-0881-2_11.
[33] Padurariu C, Breaban ME. Dealing with data imbalance in text classification. Procedia Comput Sci 2019;159:736–45. http://dx.doi.org/10.1016/j.procs.2019.09.229, URL: https://www.sciencedirect.com/science/article/pii/S1877050919314152.
[34] Das S, Bhattacharyya K, Sarkar S. Performance analysis of logistic regression, naive Bayes, KNN, decision tree, random forest and SVM on hate speech detection from Twitter. Int Res J Innov Eng Technol 2023;7(3):24. http://dx.doi.org/10.47001/irjiet%2F2023.703004.
[35] Tiwari A, Agrawal A. Comparative analysis of different machine learning methods for hate speech recognition in Twitter text data. In: 2022 third international conference on intelligent computing instrumentation and control technologies. ICICICT, 2022, p. 1016–20. http://dx.doi.org/10.1109/ICICICT54557.2022.9917752.
[36] Chae Y, Davidson T. Large language models for text classification: From zero-shot learning to fine-tuning. Open Sci Found 2023. http://dx.doi.org/10.31235/osf.io/sthwk.
[37] Babaeianjelodar M, Prudhvi GP, Lorenz S, Chen K, Mondal S, Dey S, Kumar N. Explainable and high-performance hate and offensive speech detection. 2022, http://dx.doi.org/10.48550/arXiv.2206.12983, CoRR, abs/2206.12983.
[38] Minaee S, Kalchbrenner N, Cambria E, Nikzad N, Chenaghlu M, Gao J. Deep learning–based text classification: A comprehensive review. ACM Comput Surv 2021;54(3). http://dx.doi.org/10.1145/3439726.
[39] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Lu, Polosukhin I. Attention is all you need. In: Advances in neural information processing systems, vol. 30, Curran Associates, Inc.; 2017, URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
[40] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018, http://dx.doi.org/10.48550/arXiv.1810.04805, arXiv preprint arXiv:1810.04805.
[41] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. Scikit-learn: Machine learning in python. J Mach Learn Res 2011;12(85):2825–30, URL: http://jmlr.org/papers/v12/pedregosa11a.html.
[42] Suryaningrum KM. Comparison of the TF-IDF method with the count vectorizer to classify hate speech. Eng MAthematics Comput Sci (EMACS) J 2023;5(2):79–83. http://dx.doi.org/10.21512/emacsjournal.v5i2.9978.
[43] Hackeling G. Mastering Machine Learning with scikit-learn. Packt Publishing Ltd; 2017, URL: https://github.com/PacktPublishing/Mastering-Machine-Learning-with-scikit-learn-Second-Edition.
[44] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. In: 1st international conference on learning representations, ICLR 2013, scottsdale, arizona, USA, May 2-4, 2013, workshop track proceedings. 2013, URL: http://arxiv.org/abs/1301.3781.
[45] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing. EMNLP, Association for Computational Linguistics; 2014, p. 1532–43, URL: https://aclanthology.org/D14-1162.
[46] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Davison J, Shleifer S, von Platen P, Ma C, Jernite Y, Plu J, Xu C, Le Scao T, Gugger S, Drame M, Lhoest Q, Rush A. Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations. Online: Association for Computational Linguistics; 2020, p. 38–45, URL: https://aclanthology.org/2020.emnlp-demos.6.
[47] Cañete J, Chaperon G, Fuentes R, Ho JH, Kang H, Pérez J. Spanish pre-trained bert model and evaluation data. 2023, http://dx.doi.org/10.48550/arXiv.2308.02976, arXiv preprint arXiv:2308.02976.
[48] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zetlemoyer L, Stoyanov V. Roberta: A robustly optimized bert pretraining approach. 2019, http://dx.doi.org/10.48550/arXiv.1907.11692, arXiv preprint arXiv:1907.11692.
[49] Fandiño AG, Estapé JA, Pàmies M, Palao JL, Ocampo JS, Carrino CP, Oller CA, Penagos CR, Agirre AG, Villegas M. MarIA: Spanish language models. Proces Leng Nat 2022;68. http://dx.doi.org/10.26342/2022-68-3, URL: https://upcommons.upc.edu/handle/2117/367156#.YyMTB4X9A-0.mendeley.
[50] Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave E, Ott M, Zettlemoyer L, Stoyanov V. Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th annual meeting of the association for computational linguistics. 2020, p. 8440–51, URL: https://aclanthology.org/2020.acl-main.747.
[51] Branco P, Torgo L, Ribeiro RP. A survey of predictive modeling on imbalanced domains. ACM Comput Surv 2016;49(2). http://dx.doi.org/10.1145/2907070.
[52] Tiedemann J, Thottingal S. OPUS-MT – building open translation services for the world. In: Proceedings of the 22nd annual conference of the European association for machine translation. Lisboa, Portugal: European Association for Machine Translation; 2020, p. 479–80, URL: https://aclanthology.org/2020.eamt-1.61.
[53] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. New York, NY, USA: Association for Computing Machinery; 2019, p. 2623–31. http://dx.doi.org/10.1145/3292500.3330701.
[54] Sagi O, Rokach L. Ensemble learning: A survey. WIREs Data Min Knowl Discov 2018;8(4):e1249. http://dx.doi.org/10.1002/widm.1249, URL: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1249.
[55] Mohammed A, Kora R. An effective ensemble deep learning framework for text classification. J King Saud Univ - Comput Inf Sci 2022;34(10, Part A):8825–37. http://dx.doi.org/10.1016/j.jksuci.2021.11.001, URL: https://www.sciencedirect.com/science/article/pii/S1319157821003013.
[56] Careja R, Harris E. Thirty years of welfare chauvinism research: Findings and challenges. J Eur Soc Policy 2022;32(2):212–24. http://dx.doi.org/10.1177/09589287211068796.
[57] Soral W, Malinowska K, Bilewicz M. The role of empathy in reducing hate speech proliferation. Two contact-based interventions in online and off-line settings. Peace Confl: J Peace Psychol 2022;28(3):361–71. http://dx.doi.org/10.1037/pac0000602.
[58] Pan R, Antonio García-Díaz J, Valencia-García R. Comparing fine-tuning, zero and few-shot strategies with large language models in hate speech detection in english. CMES - Comput Model Eng Sci 2024;140(3):2849–68. http://dx.doi.org/10.32604/cmes.2024.049631, URL: https://www.sciencedirect.com/science/article/pii/S1526149224000493.
[59] Choudhary M, Agarwal B, Goyal V. Hate speech detection: Leveraging LLM-GPT2 with fine-tuning and multi-shot techniques. Procedia Comput Sci 2025;258:2817–25. http://dx.doi.org/10.1016/j.procs.2025.04.542, International Conference on Machine Learning and Data Engineering, URL: https://www.sciencedirect.com/science/article/pii/S1877050925016461.
[60] Hashir MH, Memoona, Kim SW. TARGE: large language model-powered explainable hate speech detection. PeerJ Comput Sci 2025;11. http://dx.doi.org/10.7717/peerj-cs.2911.
[61] Hanu L, Unitary team. Detoxify. 2020, Github. https://github.com/unitaryai/detoxify.