Sage Journals

Journal of Intelligent & Fuzzy Systems
OnlineFirst

https://doi.org/10.3233/JIFS-219349

Research Article

Odio-BERT: Evaluating domain task impact in hate speech detection

Mesay Gemeda Yigezu, Olga Kolesnikova*, Alexander Gelbukh, and Grigori Sidorov

Abstract
The rise of social media and micro-blogging platforms has led to concerns about hate speech, its potential to incite violence, psychological trauma, extremist beliefs, and self-harm. We have proposed a novel model, Odio-BERT for detecting hate speech using a pretrained BERT language model. This specialized model is specifically designed for detecting hate speech in the Spanish language, and when compared to existing models, it consistently outperforms them. The study provides valuable insights into addressing hate speech in the Spanish language and explores the impact of domain tasks.

Keywords
BERT, hate speech, domain task, fine tune, Odio-BERT, Spanish

Centro de Investigación en Computación(CIC), Instituto Politécnico Nacional (IPN), Mexico City, Mexico

Corresponding author(s):

*Corresponding author. Olga Kolesnikova. E-mail: mgemedak2022@cic.ipn.mx.

1 Introduction

Due to the heightened interconnectedness among individuals, the rise in popularity of social media and micro-blogging platforms continues to have uncharted consequences for our daily lives. Even though this connectivity has potential benefits [1, 2], the prevalence of hate speech and associated problems frequently overshadows them. If this is not addressed effectively and continues to attack people and organizations, it could fuel violence on a global scale.

Eventually, this can target communities based on their characteristics or affiliations not only to provoke violence but also to exacerbate the already-existing marginalization and discrimination experienced by minority communities [3, 4].

The troubling aspect of this phenomenon is that vulnerable members of society are more likely to constantly be exposed to such harmful marginalization through online platforms and social media. For instance, one of the most concerning consequences of prolonged exposure to hate speech is the significant psychological trauma it can cause. Individuals who are subjected to this phenomenon may suffer from anxiety, depression, and a sense of hopelessness, which can have a profound impact on their mental well-being [5].

To tackle all the above problems, various researchers have proposed automatic methods to detect hate speech incidents. Some of them use feature-based linear classifiers, which belong to a category of machine learning algorithms. The objective of linear classifiers is to make predictions based on a set of features or attributes associated with input data [6–9]. Beside linear classifiers, deep learning methods, particularly deep neural networks like CNN, LSTM, Bi-LSTM, and RNNs have been employed [10–14]. Also, language models that had already been trained, such as BERT [15], RoBERTa [16], T5 [17], XLNet [18], ELEC-TRA [19], and GPT [20, 21], were used to test how well they could find hate speech.

The outcomes in the realm of natural language processing (NLP) and different approaches to this task vary greatly as they are extremely dependent on a wide diversity of circumstances, prominently encompassing the dataset and the architectural choices made.

This inherent variability in results has led to some intriguing observations. Linear classifiers have consistently demonstrated their competitive performance, often rivaling, and in some cases surpassing, the effectiveness of neural networks across machine learning tasks [22–24].

However, the advent of pretrained language models known as GPT-3, BERT, RoBERTa, and their variants, have achieved unprecedented state-of-the-art results across a wide spectrum of NLP tasks. They achieved this by harnessing the power of massive pretraining on vast and diverse text corpora, enabling them to capture rich semantic and contextual information. Their versatility and capability show that they can generalize to a wide range of language-understanding tasks.

In spite of language models' proficiency in handling general-purpose language understanding tasks, they may falter when confronted with more specialized or domain-specific language varieties. The reason for this limitation lies in their training data, which comprises a diverse array of languages and topics, making them exceptionally well-rounded but potentially less attuned to specific jargon, terminology, or nuances that are prevalent in niche domains. To mitigate this challenge effectively, it becomes necessary to fine-tune these pretrained models on domain-specific data using transfer learning. Toward this concept, many studies have attempted to train and build domain-specific GPT as pretrained language models. These are BioGPT for biomedical tasks [25], DialoGPT for conversation activities [26], EmoDialoGPT: generating response for emotion [27], BERT pretrained language models such as FinBERT for financial domain [28–30], LEGAL-

BERT for legal domain [31], BioBERT for the biomedical task [32, 33], HateBERT for English hate speech domain [34].

These domain-specific models fine-tuned on domain-specific data allow for higher accuracy and better performance when compared to models built for general tasks [32]. They offer several advantages over generic NLP models. Such advantages are often crucial for addressing hate speech detection.

Here are some key advantages of using a domain-specific model:

- Improving the model's understanding of domain-specific language makes it more contextually relevant. For instance, medical domain-specific models can outperform generic models in medical text-understanding tasks [32].
- From a data perspective, it requires less training data compared to training a generic model from scratch. This is because the pretrained base model already has a strong understanding of language, and fine-tuning focuses on adapting this knowledge to a specific domain. This can be a special advantage when labeled data is scarce or expensive to obtain [35, 36].
- Since many domains have unique terminology and vocabulary, generic models may struggle with domain-specific jargon; however, domain-specific models can be tailored to recognize and interpret this specialized language more effectively [37]. For instance, domains like law, medicine, finance, and engineering rely heavily on specialized terminology and jargon.

Therefore, the present study proposes """**Odio-BERT**""", a pretrained BERT model for addressing hate speech in the context of social media with a focus on the Spanish language (odio in Spanish means hate in English).

The following is a summary of this paper's significant contributions:

- Introduction of Odio-BERT, a specialized pretrained BERT model designed for detecting hate speech in Spanish, along with our source code.

- Compilation of all available Spanish hate speech detection datasets.
- Evaluation and comparison of our models with existing pretrained language models.
- Study of the impact of domain tasks and an insightful advice on how to deal with hate speech in Spanish.

For any researchers interested in furthering this research, they can access our source code and a comprehensive hate speech dataset through the following link:[1]. This repository provides access to the tools and resources necessary to build upon our work and delve deeper into the subject. You can also download our models from the following link:[2]. This resource allows easy access to the pretrained models we've developed, facilitating further research and experimentation in the field of hate speech detection.

The remainder of the paper is structured as follows: in Section 2, we explore the body of prior research and offer an insightful summary of it. We provide a thorough description of our data collection procedure in Section 3. In Section 4, we provided a detailed discussion of the proposed model. The experiments and their parameters are described in Section 5, the results are discussed in Section 6, in Section 7, we addressed the limitations and ethical considerations of the study, and the paper concludes with suggestions for further research.

2 Related work

Recently, researchers have achieved promising results by using domain-specific NLP models for tasks like hate speech detection. The utilization of domain-specific models can significantly enhance the performance of NLP applications in specific tasks, as highlighted in various studies we review in what follows.

2.1 Research focus on specific language

RoBERTuito [38]: This research demonstrates the effectiveness of a language model pretrained specifically for user-generated text in the Spanish language, utilizing a vast dataset comprising more than 50 million tweets. The primary emphasis of this model lies in its specialization in Spanish, with its training and fine-tuning being exclusively dedicated to this language. According to the author, the results indicate that this language model surpassed the performance of other pretrained models in Spanish such as ALBETO and DistilBETO [39].

ARBERT and MARBERT [40]: These are the names of two new transformer-based language models specifically designed for the Arabic language processing. These models were meticulously crafted by pretraining them using extensive and varied datasets. The primary goal behind this effort was to enable more effective transfer learning for Arabic dialects. Additionally, a new benchmark called ARLUE was introduced to assess the models' performance thorough evaluations and comparisons with existing models. Impressively, their results demonstrated superior performance across a range of downstream tasks.

BanglaBERT [41]: In this research, the authors introduced a Natural Language Understanding (NLU) model pretrained specifically for the Bengali language, known as BanglaBERT. To develop this model, they gathered a vast dataset comprising 27.5 gigabytes of text in Bengali for pretraining. Furthermore, they created two additional datasets for downstream tasks, focusing on natural language inference and question answering. The results they obtained with BanglaBERT were truly impressive, as it surpassed the performance of both multilingual and monolingual models, including mBERT and XLM-R (base). Specifically, when comparing blue scores, BanglaBERT outperformed these models by 6.8 and 4.3 points, respectively.

The above results indicate that when a pretrained model is further trained for a particular language, it demonstrates superior performance compared to more generalized models that are not fine-tuned for that specific language. This suggests the importance of language-specific fine-tuning for achieving better results in natural language processing tasks.

2.2 Research focus on specific tasks

FinBERT [29]: aims to fulfill a specific requirement within the financial domain by developing BERT models trained on extensive financial communication datasets. Through rigorous evaluation across three distinct financial sentiment classification tasks, to assess its effectiveness, the authors put FinBERT to test in three different financial sentiment classification tasks. The results demonstrate that FinBERT outperforms the generic domain BERT model in these tasks, underscoring its superior suitability and performance in the financial domain.

BioBERT [32]: is designed for biomedical text mining, specifically trained on extensive biomedical datasets. BioBERT maintains a remarkably similar architecture to BERT across various tasks. This model surpasses them significantly in three pivotal biomedical text mining domains: biomedical named entity recognition, biomedical relation extraction, and biomedical question answering. These findings emphasize the profound impact of pretraining BERT on biomedical corpora, equipping it with a deeper comprehension of intricate biomedical texts and enabling it to excel in these crucial biomedical language understanding tasks.

LEGAL-BERT [31]: is situated within the legal domain, it was designed to adapt BERT for various legal tasks. The researchers collected a diverse dataset of 12 gigabytes comprising English legal content from different domains, which was obtained from publicly available resources. Subsequently, they evaluated their model on multiple datasets. Previous to this research, the use of BERT was often pursued without careful consideration and did not consistently deliver satisfactory outcomes in the legal sphere. In simpler terms, the standard procedures did not consistently perform well when applied to legal tasks, but Legal-BERT proved to be effective.

2.3 Hate speech detection

Abuse-XLMR [42]: The authors designed a model to detect abusive content. The model was trained using an extensive dataset of social media comments in over 15 Indic languages. What sets AbuseXLMR apart is its exceptional performance compared to existing models like XLM-R [43] and MuRIL [44] when tested on various Indic datasets. This accomplishment marks a significant breakthrough in the realm of abusive content detection, particularly for Indic languages. AbuseXLMR's superior performance suggests its potential to greatly enhance the effectiveness of moderating online content and promoting a safer online environment across a diverse range of languages and cultures.

COVID-HateBERT [45]: is an advanced language model designed to combat hate speech in English tweets, specifically focusing on those related to COVID-19. To tackle this pressing issue, researchers gathered a vast dataset comprising 200 million English tweets containing hateful keywords and hashtags associated with the pandemic. From this extensive collection, a classifier was employed to identify and isolate approximately 1.27 million potential hateful tweets. Subsequently, this data was used to fine-tune the BERT-base model. The performance of COVID-HateBERT was rigorously evaluated across four benchmark datasets. The results were striking, demonstrating the model's effectiveness. When compared to baseline methods in traditional hate speech detection, COVID-HateBERT exhibited a significant improvement of a macro average F1 score from 14.8% to 23.8%. Moreover, in the realm of COVID-19-related hate speech detection, COVID-HateBERT continued to outperform, boasting a 2.6% to 6.73% higher macro average F1 score compared to classifiers using BERT and BERTweet.

HateBERT [34]: is an advanced BERT-based model specifically developed for the purpose of identifying the abusive language in English text. This model underwent training using RAL-E, a comprehensive dataset compiled from Reddit comments in English originating from communities that were banned due to their offensive, abusive, or hateful content. The study in [34] includes a thorough examination of HateBERT, comparing it with a conventional, general-purpose pretrained language model. This comparison is made using data from three separate English datasets tailored to offensive language, abusive content, and hate speech detection tasks. The results of this comparison reveal that, across all three datasets, HateBERT consistently outperforms the standard BERT model designed for general language understanding. This impressive performance demonstrates HateBERT's prowess in tackling the critical issue of identifying and mitigating abusive language, showcasing its potential as a valuable tool for improving online content moderation and fostering a safer digital environment.

In [46], the researchers conducted an investigation into the most effective attributes for detecting hate speech in the Spanish language. They also explored how these attributes could be integrated to create more precise systems. Furthermore, the researchers analyzed the linguistic characteristics found in different forms of hate speech using explainable linguistic features. They then compared their findings with existing state-of-the-art methods. The results of this research demonstrate that combining linguistic features and transformer models through the integration of knowledge yields better performance than existing solutions when it comes to identifying hate speech in Spanish. Additionally, numerous studies have been conducted to address the problem of detecting hate speech in Spanish. The use of machine learning, [47, 48], deep learning [49–53], and the use of various pretrained language models [54–57] are just a few of the techniques that have been used in these initiatives. Despite these substantial research efforts, there is still a clear void in this area.

Based on the insights gleaned from the above-mentioned study, there are fewer works in Spanish than there are in English. It is crucial to increase the accuracy of hate-speech identification in Spanish. It becomes evident that using domain-specific tasks can significantly enhance performance across various dimensions.

3 Datasets

In this section, we provide an overview of the Spanish hate speech datasets utilized in our study. To conduct our experiments, we employed hate speech datasets sourced from prior research papers that are publicly accessible. These datasets were compiled by researchers from various social media platforms, including YouTube, Facebook, and Twitter[3], and were obtained in text format. Our dataset is obtained from the following datasets: (1) Spanish Miso Corpus 2020, (2) Multi-lingual HateSpeech Dataset, (3) HomoMex, (4) HaSCoSVa-2022, (5) HaterNet, and (6) HatEval 2019 dataset.

Spanish MisoCorpus 2020 [57]: Comprises three distinct splits: SELA, designed to investigate variations in misogynistic messages between Spanish from Spain and Spanish from Latin America; the second is VARS, which focuses on instances of violence directed towards women in politics and public media; and the third is DDSS, containing general characteristics associated with misogyny. In our study, we utilize the entire corpus, which encompasses a total of 8,390 tweets. It's worth noting that this corpus exhibits a slight imbalance, with a greater number of tweets categorized as non-misogynistic. The annotation process for this dataset-corpus involved manual annotations by three human annotators.

Multi-lingual HateSpeech Dataset: We sourced this dataset from Kaggle competitions[4], and it comprises text samples classified into hate speech and non-hate speech categories, where the label "0" signifies non-hate speech and "1" indicates hate speech. Additionally, the dataset includes text samples in various languages that need to be correctly identified and labeled with their corresponding language codes. There are 12 languages present in the dataset, in addition to Spanish.

Furthermore, the dataset provides LASER 1024-Dimensional embeddings for both training and test data, resulting in a total of 12,424 sample texts available for analysis.

HOMO-MEX [58]: This dataset comprises publicly available tweets written in Spanish, specifically from Mexico. The tweets were collected over a span of time from January 1, 2012, to January 10, 2022. In the process, 11,000 tweets were annotated, with an even distribution between those posted by unverified accounts and those by verified accounts, before the implementation of account verification monetization.

The tweets were then subjected to annotation to identify instances of LGBT+ phobia. The dataset was divided into two sets: a training set consisting of 7,000 samples and a test set consisting of 4,000 samples. The primary objective of this task was to detect hate speech, utilizing a multi-class approach that involved categorizing the content into three classes: LGBT+phobic (P), not LGBT+phobic (NP), or not related to LGBT+ (NA).

In our work, we removed 1,777 samples labeled as NA from the training set, resulting in a total of 5,223 data samples used for our experiments.

HaSCoSVa-2022 (Hate Speech Corpus with Spanish Variations) [59]: The assembly and annotation of HaSCoSVa-2022 involved curating a dataset of tweets centered on hate speech directed at immigrants, specifically in the Spanish language. This corpus was enriched with information indicating the specific language variation used. The dataset was further categorized into two subsets based on language variants: (1) Latin American and (2) European. Importantly, this dataset was made available to the research community.

To create HaSCoSVa-2022, the researchers conducted a comprehensive review of publicly accessible data to identify instances of hate speech against immigrants in the Spanish language. It’s noteworthy that, to the best of their knowledge, there were no existing Twitter corpora that accounted for variations in language.

The motivation behind creating HaSCoSVa-2022 was to facilitate experiments in the domain of immigration. Specifically, the researchers focused on two distinct immigration scenarios: one involving immigration from Latin America and certain African countries to Spain, and the other involving immigration from Venezuela to neighboring countries where Spanish is the official language. Both of these cases were associated with prevalent online discourse marked by discrimination, often stemming from religious, stereotypical, and other factors, which affect a portion of the local population.

Hater-Net: is the dataset originated from Twitter and was constructed through a multi-stage process. Initially, a vast set of 2 million tweets was collected and subsequently filtered using both automated and manual methods. Four human annotators then tagged these tweets for classification.

One notable characteristic of the HaterNet dataset is its significant class imbalance, where 1,567 documents were annotated as hateful, while 4,433 were labeled as non-hateful. In their evaluation, the authors of HaterNet focused on assessing the F1 score specifically for the hateful class.

Throughout their research, the creators of the HaterNet dataset introduced a novel approach that combined recurrent neural networks and multilayer perceptrons to integrate embeddings, emojis, and other statistical features. This approach yielded promising results, achieving an area under the curve (AUC) of 0.828.

HatEval 2019 [61]: This dataset was made available as part of the SemEval 2019 shared task. It was specifically designed to assess the ability to detect hate speech directed at immigrants and women. HatEval 2019 introduced two subtasks: (1) identifying hate speech targeting immigrants and women, and (2) classifying aggressive behavior and determining whether the target of the aggression is an individual or a group.

In the context of HatEval 2019, the Spanish subset of the dataset comprised a total of 6,599 tweets, which were further divided into training, validation, and testing sets. In the Spanish binary subtask, the best performance achieved resulted in a macro averaged F1 score of 73% .

Among the datasets mentioned above, we leveraged all datasets for training our new model, with one notable exception being the HatEval 2019 dataset. The HatEval dataset was exclusively reserved for the purpose of evaluating and comparing our novel model against existing models.

Table 1 presents a summary of the benchmark dataset statistics, which encompass the following data sets:

Table 1 Summary of the benchmark dataset for Spanish hate speech (Table view)

Dataset	Label	Size in #	Size in%	Total size
Spanish MisoCorpus 2020	Misogyny	3700	44.1%	8390
	Not-misogyny	4690	55.9%	
Multi-lingual Hate Speech Dataset	Hate	4239	34.1%	12424
	Not-hate	8184	65.9%	
HOMO-MEX	LGBT+phobic	862	16.6%	5223
	Not-LGBT+phobic	4360	83.4%	
HaSCoSVa-2022	Hate	556	13.9%	4000
	Not-hate	3444	86.1%	
Hater-Net	Hate	4433	73.9%	6000
	Not-hate	1567	26.1%	
HatEval 2019	Hate	2739	41.5%	6600
	Not-hate	3861	58.5%	

In total, we gathered 36,037 samples from the sources described in this section except HatEval 2019. After preprocessing the data, we ended up with 34,622 samples, indicating the presence of duplicate entries in the dataset.

4 Odio-BERT

Numerous language models have been developed that are exclusively trained on the Spanish language, such as RoBERTuito [38], BETO [62], and BERTIN [63]. However, it’s important to note that pretraining of these models is not tailored specifically for detecting hate speech in Spanish. This means that these contextual models are primarily trained on extensive multilingual datasets but are not fine-tuned for the hate speech domain. As a result, a domain gap exists, and the performance of these models in hate speech detection may suffer.

To address this issue, we took the initiative to pretrain a BERT model on domain-specific data, specifically, the hate speech dataset, which we extracted from various sources as discussed in Section 3. The choice of BERT for pretraining was deliberate, as it has demonstrated impressive performance and is gaining increasing popularity as a competitive, efficient, and swift solution for adapting pretrained language models to new languages and domains, as exemplified by [34]. We give the name Odio-BERT to our novel, fine-tuned BERT.

BERT models have already acquired substantial knowledge about language and context from extensive general text data. When fine-tuned on domain-specific data, this knowledge becomes tailored to the nuances of a particular domain, often resulting in a significant boost in task performance. Pretraining on the hate speech corpus equips Odio-BERT with an understanding of the intricacies of social media, including common issues like spelling mistakes, grammatical errors, and emoticons. Consequently, it enhances Odio-BERT’s capabilities in comparison to other models like BERT and BETO [62].

Our experiments provide clear evidence of Odio-BERT’s effectiveness, underscoring the importance of bridging the domain gap across the various datasets we compiled. This work highlights the value of adapting language models for specific domains, ultimately leading to superior performance in the task of hate speech detection.

5 Experiments

For our experiment, we merged all the datasets described in Section 3 and carried out a stratified split into training, development, and test sets in the respective proportions of 32,122, 500, and 2,000 samples. These splits remained consistent across all Odio-BERT experiments, ensuring that the results can be compared between different models and configurations.

To assess the impact of varying the amount of labeled data, we chose a subset from the dataset and repeated this selection process 12 times with different random sets. We then averaged the performance across these repetitions. The subsets we used were stratified samples, with sizes of 10, 20, 30, 100, 200, 300, 1,000, 2,000, 3,000, 10,000, 20,000, and 30,000. This approach enables us to evaluate the impact of utilizing more or less labeled data, both within specific orders of magnitude and across them. The results reveal that there can be substantial variations in performance when fine-tuning on limited data.

To mitigate performance discrepancies among different sets, we employed 5 distinct random seeds for each sample size in our experiments.

Figure 1 shows that our experiments exclusively revolved around Transformer based models. The reason behind this choice is the remarkable performance that Transformer models have exhibited in recent times across various hate speech detection tasks, as evidenced by several prominent studies [64–66]. These models, characterized by their advanced self-attention mechanisms, have consistently demonstrated their proficiency in capturing nuanced contextual information, making them the prime candidates for our investigation.

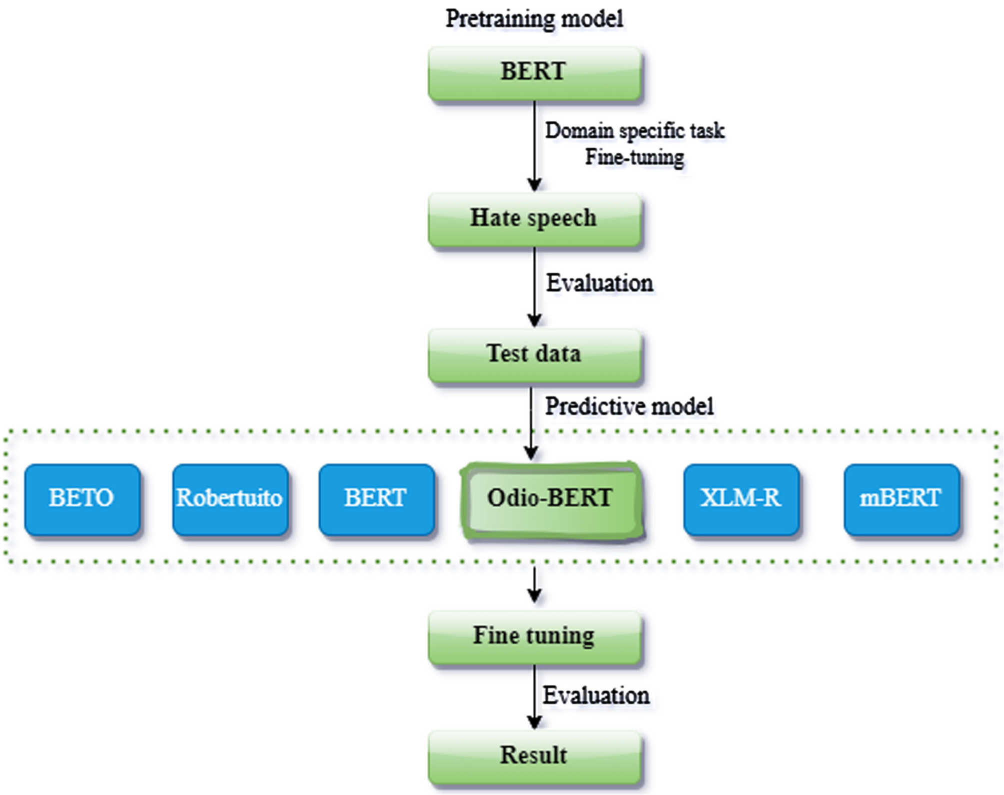

Fig. 1 Proposed framework for Spanish hate speech detection.

As indicated by Table 2, the effectiveness of fine-tuning improves as the size of the training data increases. Drawing insights from our experimental results, we chose a model that had been trained with 20,000 samples as the foundation for our novel model.

Table 2 The macro average F1 scores on the respective test datasets for models fine-tuned using **N** entries, were computed by averaging results from 5 random seeds for each N. The most outstanding performance for a particular N is highlighted in **bold** (Table view)

Odio-BERT												
N	**10**	**20**	**30**	**100**	**200**	**300**	**1,000**	**2,000**	**3,000**	**10,000**	**20,000**	**30,000**
BERT-fine-tuned	0.2	0.46	0.3	0.5	0.52	0.52	0.58	0.63	0.63	0.75	**0.83**	0.83

In addition to introducing our novel model, we aimed to comprehensively assess and combat hate speech in the Spanish. To achieve this, we considered a range of approaches.

The rationale behind our selection of BERT models lies in our desire to examine the distinctions between the standard BERT and our fine-tuned variant, referred to as Odio-BERT. Furthermore, we opted for two models from the pool of multilingual language models, specifically XLM-R and mBERT. These choices were made because both models were trained on extensive datasets encompassing various languages, including Spanish. Our objective is to assess how well these multilingual models perform in downstream tasks when supplied with training data in diverse languages.

In addition to our multilingual focus, we also made deliberate selections of two models that are centered around the Spanish language. These models, BETO and RoBERTuito, were chosen with the aim of conducting a comprehensive analysis and comparison between models designed for language-specific tasks and those tailored for domain-specific tasks, thereby shedding light on their respective performance levels.

Monolingual Models: We explored the utility of monolingual BERT models, which are specifically fine-tuned for Spanish hate speech detection. These models have been customized to provide enhanced performance and accuracy in handling Spanish text data, so we could evaluate and understand the impact of a specific domain task.

Multilingual Models: We extended our evaluation to encompass multilingual models, including mBERT [67] and XLM-R [68]. These models are renowned for their ability to comprehend and process text in multiple languages, making them invaluable in multilingual contexts and scenarios with limited language resources.

mBERT introduced by [67], represents a significant milestone in the field of natural language processing. mBERT is a variant of the BERT model that has been pretrained on a massive corpus of text from 104 languages including Spanish, making it a powerful and versatile tool for multilingual NLP tasks. The primary objective of mBERT is to learn a universal language representation that can capture the nuances of language across various languages and tasks. By pretraining on text from multiple languages, mBERT aims to create a model that can understand and generate text in multiple languages, effectively breaking down language barriers in NLP.

XLM-R [68] is a powerful and versatile natural language processing model that has made significant contributions to multilingual and cross-lingual tasks. One of the notable aspects of XLM-R is its pretraining on a massive corpus of text from a wide range of 100 languages including Spanish. This diverse pretraining enables XLM-R to learn rich representations of text, allowing it to transfer knowledge across languages and perform well on a variety of language-related tasks.

Spanish-Centric Models: Our evaluation also involved Spanish-centric models like RoBERTuito [38] and BETO [62]. This model was designed with a primary focus on the nuances and characteristics of the Spanish language, making it particularly effective for tasks specific to Spanish text data.

BETO is pretrained on a massive corpus of Spanish text, allowing it to learn rich linguistic features and nuances of the Spanish language. It leverages the same transformer architecture as BERT, which is based on self-attention mechanisms. This architecture enables BETO to capture dependencies and relationships between words in a bidirectional manner, making it highly effective for various NLP tasks and we discussed about RoBERTuito in subsection 2.1.

To sum it up, the state of the art models we chose for the experiments to compare with our proposal are the following: 1) BERT, 2) BETO, 3) RoBERTuito, 4) XLM-R and 5) mBERT. We used such diverse models to comprehensively assess hate speech detection in Spanish, addressing various linguistic challenges and data scenarios. Results will be discussed in Section 6.

To assess the performance of our novel model in comparison to existing models, we employed the HateEval 2019 data set for evaluation. The primary factor motivating our choice of this dataset is its characteristic balance between classes, which minimizes bias and ensures a fair representation of different categories. This balanced distribution allows for a more reliable and equitable evaluation of model performance. By using this dataset, we aimed to conduct a thorough and unbiased assessment of our new model's capabilities in the context of hate speech detection.

Our experiments were carried out using the NVIDIA System Management Interface (NVIDIA-SMI) with CUDA version 12.1, a powerful framework for GPU acceleration. In our setup, we harnessed the capabilities of an NVIDIA GeForce GTX 1080 graphics processing unit (GPU) equipped with 8 gigabytes (GB) of memory, which is equivalent to 8,192 mebibytes (MiB). This high-performance GPU played a critical role in the execution and optimization of our computational tasks, ensuring efficient and robust results in our experiments.

During the training of a domain-specific task, we utilized a training batch size of 8 while configuring a maximum sequence length of 512. For the evaluation phase, a batch size of 16 was employed, and our training regimen spanned a total of 3 epochs.

When we fine-tuned selected models for analysis with various approaches, we maintained the same settings, except for the epoch count, which was extended to 5. This adjustment allowed us to gather more comprehensive insights during the analysis phase.

6 Results

As we discussed in Section 5, we conducted a series of experiments to assess the effectiveness of our novel model compared to existing models. Table 3 presents the evaluation results for hate speech detection, where we report accuracy and the macro average F1 score to gauge model performance.

Table 3 Experimental results of the models including our model (Table view)

Models	Metrics			
	Precision	Recall	F1 Score	Accuracy
mBERT	0.61	0.6	0.6	0.67
XLM-R	0.65	0.63	0.63	0.67
Robertuito	0.82	0.73	0.77	0.77
BETO	0.81	0.7	0.75	0.76
BERT	0.76	0.69	0.72	0.74
Odio-BERT	0.85	0.76	**0.80**	0.85

The outcomes of our experiments monolingual experiments indicate that the BERT model surpasses multilingual models. This superiority is attributed to its proficiency in capturing language-specific intricacies, nuances, and contextual cues. In cases where the objective revolves around deciphering language-specific hate speech, the language-specific model outperforms its multilingual counterparts.

Notably, our findings reveal that models tailored for the Spanish language demonstrate superior performance compared to mBERT, XLM-R, and BERT models. This enhanced capability is attributed to their adeptness at grasping the linguistic structures of the Spanish language, stemming from their pretraining and fine-tuning using a substantial Spanish corpus.

Conversely, the multilingual model, which encompasses Spanish, exhibited suboptimal performance in hate speech detection when juxtaposed with existing models. The difference is because multilingual models are made to understand a lot of different languages, but they might miss the finer points that are important for certain tasks that are only important for one language. For endeavors such as hate speech detection, where cultural and linguistic contexts play a pivotal role, language-specific models hold a competitive edge [69, 70].

In summary, our model consistently outperformed the spectrum of pretrained models discussed in the previous Section 5. This underscores the value of domain-specific fine-tuning, which enhances a model's capacity to capture task-specific intricacies, as elucidated in Section 1.

7 Limitations and ethical considerations

Limitation: The Spanish language is spoken globally, and addressing hate speech on social media lacks sufficient data. Consequently, we gathered publicly released datasets from various sources, including YouTube, Facebook, Twitter, blogs, and others. However, this collection doesn't represent the entire population, highlighting the need for ongoing collaborative efforts to narrow this gap.

Caution: We kindly ask the community to be aware that the compiled dataset includes comments expressing hate speech toward religion, race, gender, etc., which may be distressing for researchers. We chose not to censor these hateful words or phrases, as it would undermine the study's purpose. Please exercise discretion when engaging with our work.

8 Conclusion

We embarked on a domain-specific task utilizing a pretrained language model to effectively identify hate speech in the Spanish language. The process involved the amalgamation of diverse data sources provided by multiple researchers, which, in turn, served as the foundation for training the model. We thoughtfully curated and created a benchmark dataset tailored to the intricacies of hate speech. Fine-tuning the BERT model with this data was pivotal in comprehending and capturing the nuances of hate speech.

We carefully looked at different model sizes using different parts of the benchmark data and eventually chose a model trained on 20,000 samples as our domain-specific solution. This model is called Odio-BERT. Our proposed model does better than all the other models that have been used before. Tests using fine-tuned monolingual, multilingual, and Spanish-focused models on the HatEval dataset demonstrated this.

As a recommendation for future researchers in this field, we suggest that further enhanced results can be achieved through the fine-tuning of Spanish-centric models. The prospect of conducting a comprehensive study using a large-scale Spanish monolingual model for hate speech detection holds promise; however, we defer this endeavor to future research initiatives.

Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1S-47854 of CONAHCYT, Mexico, and grants 20241816, 20241819, and 20240951 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONAHCYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico, and acknowledge the support of Microsoft through the Microsoft Latin America Ph.D. Award.

References

1. Byman D.L., How hateful rhetoric connects to real-world violence, (2021).
2. Yigezu M.G. Kanta S. Kolesnikova O. Sidorov G. Gelbukh A., Habesha@ DravidianLangTech: Abusive Comment Detection using Deep Learning Approach. In *Proceedings of the Third Workshop onSpeech and Language Technologies for Dravidian Languages* (2023), pp. 244–249.
3. Greenberg J. Pyszczynski T., The effect of an overheard ethnic slur on evaluations of the target: How to spread a social disease, *Journal of Experimental Social Psychology* 21(1) (1985), 61–72. Crossref.
4. Mullen B. Rice D.R., Ethnophaulisms and exclusion: The behavioral consequences of cognitive representation of ethnic immigrant groups, *Personality and Social Psychology Bulletin* 29(8) (2003), 1056–1067. Crossref.
5. Van Geel M. Vedder P. and Tanilon J., Relationship between peer victimization, cyberbullying, and suicide in children and adolescents: A meta-analysis, *JAMA Pediatrics* 168(5) (2014), 435–442. Crossref.
6. Waseem Z. Hovy D., Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop* (2016), pp. 88–

93. Crossref.

7. Ribeiro M. Calais P. Santos Y. Almeida V. Meira W. Jr, Characterizing and detecting hateful users on twitter. In *Proceedings of the International AAAI Conference on Web and SocialMedia* (Vol. 12, No. 1) (2018).
8. Ibrohim M.O. Setiadi M.A. Budi I., Identification of hate speech and abusive language on indonesian Twitter using the Word2vec, part of speech and emoji features. In *Proceedings of the 1st International Conference on Advanced Information Science and System* (2019), pp. 1–5.
9. Akuma S. Lubem T. Adom I.T., Comparing Bag of Words and TF-IDF with different models for hate speech detection from live tweets, *International Journal of Information Technology* 14(7) (2022), 3629–3635. Crossref.
10. Roy P.K. Tripathy A.K. Das T.K. Gao X.Z., A framework for hatespeech detection using deep convolutional neural network, *IEEE Access*. 8 (2020), 204951–204962. Crossref.
11. Badjatiya P. Gupta S. Gupta M. Varma V., Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web companion* (2017), pp.759–760. Crossref.
12. Sutejo T.L. Lestari D.P., Indonesia hate speech detection using deep learning. In *2018 International Conference on AsianLanguage Processing (IALP)* (pp. 39–43). IEEE (2018). Crossref.
13. Kshirsagar R. Cukuvac T. McKeown K. McGregor S., Predictive embeddings for hate speech detection on twitter. *arXiv preprint arXiv:1809.10644* (2018).
14. Pitenis Z. Zampieri M. Ranasinghe T., Offensive language identification in Greek. *arXiv preprint arXiv:2003.07459* (2020).
15. Devlin J. Chang M.W. Lee K. Toutanova K., Bert: Pre-training of deep bidirectional transformers for language understanding. arXivpreprint arXiv:1810.04805, (2018).
16. Liu Y. Ott M. Goyal N. Du J. Joshi M. Chen... D. Stoyanov V., Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
17. Raffel C. Shazeer N. Roberts A. Lee K. Narang S. Matena... M. Liu P.J., Exploring the limits of transfer learning with a unified text-to-text transformer, *The Journal of Machine Learning Research* 21(1) (2020), 5485–5551.
18. Yang Z. Dai Z. Yang Y. Carbonell J. Salakhutdinov R.R., ... Le Q.V., Xlnet: Generalized autoregressive pretraining for language understanding, *Advances in Neural Information Processing Systems* 32 (2019).
19. Clark K. Luong M.T. Le... Q.V. Manning C.D., Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555* (2020).
20. Radford A. Wu J. Child R. Luan D. Amodei... D. Sutskever I., Language models are unsupervised multitask learners, *OpenAI Blog* 1(8) (2019), 9.

21. Brown T. Mann B. Ryder N. Subbiah M. Kaplan J.D. Dhariwaland P. Amodei D., Language models are few-shot learners, *Advancesin Neural Information Processing Systems* 33 (2020), 1877–1901.

22. Zhang C. Bengio S. Hardt M. Recht... B. Vinyals O., Understanding deep learning (still) requires rethinking generalization, *Communications of the ACM* 64(3) (2021), 107–115. Crossref.

23. Yigezu M.G. Mehamed M.A. Kolesnikova O. Guge T.K. Gelbukh... A. Sidorov G., Evaluating the Effectiveness of Hybrid Features in Fake News Detection on Social Media. In *2023 International Conference on Information, ... Communication Technology for Development for Africa (ICT4DA)* (pp. 171–175). IEEE (2023). Crossref.

24. Yigezu M.G. Bade G.Y. Kolesnikova O. Sidorov... G. Gelbukh A., Multilingual Hope Speech Detection using Machine Learning, (2023).

25. Luo R. Sun L. Xia Y. Qin T. Zhang S. Poon... H. Liu T.Y., BioGPT: generative pretrained transformer for biomedical text generation,... mining, *Briefings in Bioinformatics* 23(6) (2022), bbac409. Crossref.

26. Zhang Y. Sun S. Galley M. Chen Y.C. Brockett C. Gao... X. Dolan B., Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536* (2019).

27. Jia Y. Cao S. Niu C. Ma Y. Zan H. Chao... R. Zhang W., EmoDialoGPT: enhancing DialoGPT with emotion. In *Natural Language Processing,... Chinese Computing: 10th CCF International Conference* NLPCC 2021, Qingdao, China, October 13–17, Proceedings, Part II 10 (pp. 219–231). Springer International Publishing, (2021).

28. Liu Z. Huang D. Huang K. Li... Z. Zhao J., Finbert: A pretrained financial language representation model for financial text mining. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence* (2021), pp. 4513–4519.

29. Yang Y. Uy... M.C.S. Huang A., Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097* (2020).

30. Araci D., Finbert: Financial sentiment analysis with pretrained language models. *arXiv preprint arXiv:1908.10063* (2019).

31. Chalkidis I. Fergadiotis M. Malakasiotis P. Aletras... N., I., ... routsopoulos, LEGAL-BERT: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559* (2020).

32. Lee J. Yoon W. Kim S. Kim D. Kim S. So... C.H. Kang J., BioBERT: a pretrained biomedical language representation model for biomedical text mining, *Bioinformatics* 36(4) (2020), 1234–1240. Crossref.

33. Yu X. Hu W. Lu S. Sun... X. Yuan Z., BioBERT based named entity recognition in electronic medical record. In *2019 10th international conference on information technology in medicine,... education (ITME)* (2019) pp. 49–52. IEEE. Crossref.

34. Caselli T. Basile V. Mitrović J., ... Granitzer M., Hatebert: Retraining bert for abusive language

detection in english. *arXiv preprint arXiv:2010.12472* (2020).

35. Howard... J. Ruder S., Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146* (2018).

36. Yigezu M.G. Kebede T. Kolesnikova O. Sidorov... G. Gelbukh A., Habesha@ DravidianLangTech: Utilizing Deep,... Transfer Learning Approaches for Sentiment Analysis. In *Proceedings of the Third Workshop on Speech,... Language Technologies for Dravidian Languages* (2023), pp. 239–243.

37. Amer N.O. Mulhem... P. Géry M., Toward word embedding for personalized information retrieval. In *Neu-IR: The SIGIR 2016 Workshop on Neural Information Retrieval* (2016).

38. Pérez J.M. Furman D.A. Alemany... L.A. Luque F., Robertuito: a pretrained language model for social media text in spanish. *arXiv preprint arXiv:2111.09453* (2021).

39. Cañete J. Donoso S. Bravo-Marquez F. Carvallo... A. Araujo V. distilbeto: Albeto..., Lightweight spanish language models. *arXiv preprint arXiv:2204.09145* (2022).

40. Abdul-Mageed M. Elmadany A.,... Nagoudi E.M.B., ARBERT & MARBERT: deep bidirectional transformers for Arabic. *arXiv preprint arXiv:2101.01785* (2020).

41. Bhattacharjee A. Hasan T. Ahmad W.U. Samin K. Islam M.S. Iqbal... A. Shahriyar R., BanglaBERT: Language model pretraining,... benchmarks for low-resource language understanding evaluation in Bangla. *arXiv preprint arXiv:2101.00204* (2021).

42. Gupta V. Roychowdhury S. Das M. Banerjee S. Saha P. Mathewand B. Mukherjee A., Multilingual abusive comment detection at scale for indic languages, *Advances in Neural Information Processing Systems*. 35 (2022), 26176–26191.

43. Conneau A. Khandelwal K. Goyal N. Chaudhary V. Wenzek G. Guzmán F.,... Stoyanov V., Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116* (2019).

44. Khanuja S. Bansal D. Mehtani S. Khosla S. Dey A. Gopalan... B. Talukdar P., Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730* (2021).

45. Li M. Liao S. Okpala E. Tong M. Costello M. Cheng... L. Luo F., Covid-hatebert: a pretrained language model for covid-19 related hate speech detection. In *2021 20th IEEE International Conference on Machine Learning,... Applications (ICMLA)* (pp. 233–238). IEEE (2021). Crossref.

46. Garcıa-Dıaz J.A. Jiménez-Zafra S.M. Garcıa-Cumbreras... M.A. Valencia-Garcıa R., Evaluating feature combination strategies for hate-speech detection in Spanish using linguistic features,... transformers, Complex & Intelligent Systems, (2023).

47. Vega L.E.A. Reyes-Magaña J.C. Gómez-Adorno... H. Bel-Enguix G., MineriaUNAM at SemEval-task 5: Detecting hate speech in Twitter using multiple features in a combinatorial framework. In *Proceedings of the 13th international workshop on semantic evaluation* (2019), pp. 447–452. Crossref.

48. Almatarneh S. Gamallo P. Pena... F.J.R. Alexeev A., Supervised classifiers to identify hate speech on English,... Spanish tweets. In *Digital Libraries at the Crossroads of Digital Information for the Future: 21st International Conference on Asia-Pacific Digital Libraries*, ICADL Kuala Lumpur, Malaysia, November 4–7, 2019, Proceedings 21 (2019) (pp. 23–30). Springer International Publishing. Crossref.

49. Arcila-Calderón C. Amores J.J. Sánchez-Holgado P. Vrysis L. Vryzas... N. Oller M., Alonso, How to detect online hate towards migrants,... refugees? Developing,... evaluating a classifier of racist,... xenophobic hate speech using shallow,... deep learning, *Sustainability* 14(20) (2022), 13094. Crossref.

50. Ashraf N. Zubiaga... A. Gelbukh A., Abusive language detection in youtube comments leveraging replies as conversational context, *PeerJ Computer Science*, 7 (2021), e742. Crossref.

51. Plaza-del-Arco F.M. Molina-González M.D. Urena-Lópezand L.A. Martın-Valdivia M.T., Comparing pretrained language models for Spanish hate speech detection, *Expert Systems with Applications* 166 (2021), 114120. Crossref.

52. Silva S.C.D. Ferreira T.C. Ramos... R.M.S. Paraboni I., Data-driven, ... psycholinguistics-motivated approaches to hate speech detection, *Computing,... Systems* 24(3) (2020), 1179–1188.

53. Pérez J.M. Luque F.M. Zayat D. Kondratzky M. Moro A. Serrati... P.S. Cotik V., Assessing the impact of contextual information in hate speech detection, *IEEE Access*. 11 (2023), 30575–30590. Crossref.

54. Plaza-Del-Arco F.M. Molina-González M.D. Ureña-López... L.A. Martın-Valdivia M.T., A multi-task learning approach to hate speech detection leveraging sentiment analysis, *IEEE Access* 9 (2021), 112478–112489. Crossref.

55. Shahiki-Tash M. Armenta-Segura J. Ahani Z. Kolesnikova O. Sidorov... G. Gelbukh A., Lidoma at homomex@ iberlef: Hate speech detection towards the mexican spanish-speaking lgbt+ population. the importance of preprocessing before using bert-based models. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) (2023).

56. Yigezu M.G. Kolesnikova O. Sidorov... G. Gelbukh A., Transformer-Based Hate Speech Detection for Multi-Class,... Multi-Label Classification, (2023).

57. Garcıa-Dıaz J.A. Cánovas-Garcıa M. Colomo-Palacios R. Valencia-Garcıa R., Detecting misogyny in Spanish tweets. An approach based on linguistics features,... word embeddings, *Future Generation Computer Systems*. 114 (2021), 506–518. Crossref.

58. Bel-Enguix G. Gómez-Adorno H. Sierra G. Vásquez J. Andersen... S.T. Ojeda-Trueba S., Overview of HOMO-MEX at Iberlef: Hate speech detection in Online Messages directed Towards the MEXican Spanish speaking LGBTQ+ population, *Natural Language Processing*. 71 (2023), 361–370.

59. Castillo-López G. Riabi... A. Seddah D., Analyzing Zero-Shot transfer Scenarios across Spanish variants for Hate Speech Detection. In Tenth Workshop on NLP for Similar Languages,

Varieties,... Dialects (VarDial 2023) (2023), pp. 1–13.

60. Pereira-Kohatsu J.C. Quijano-Sánchez L. Liberatore... F. Camacho-Collados M., Detecting, ... monitoring hate speech in Twitter, *Sensors* 19(21) (2019), 4654. Crossref.
61. Basile V. Bosco C. Fersini E. Nozza D. Patti V. Pardo F.M.R. Sanguinetti M., Semeval-task 5: Multilingual detection of hate speech against immigrants,... women on twitter. In *Proceedings of the 13th international workshop on semantic evaluation* (2019), pp. 54–63. Crossref.
62. Cañete J. Chaperon G. Fuentes R. Ho J.H. Kang... H. Pérez J., Spanish pretrained bert model and evaluation data. *arXiv preprint arXiv:2308.02976* (2023).
63. De la Rosa J. Ponferrada E.G. Villegas P. Salas P.G.D.P. Romero M. and Grandury M., Bertin Efficientre-training of a spanish language model using perplexity sampling. arXiv preprintarXiv:2207.06814 (2022).
64. Bansal M. Villavicencio A., Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL). In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)* (2019).
65. Liu P. Li W. Zou L., NULI at SemEval-task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th international workshop on semantic evaluation* (2019), pp. 87–91. Crossref.
66. Mathew B. Saha P. Yimam S.M. Biemann C. Goyal P. and Mukherjee A., Hatexplain: A benchmark dataset for exlainable hate speech detection, In *Proceedings of the AAAI conference on artificial intelligence* 35(17) (2021), 14867–14875. Crossref.
67. Devlin J. Chang M.W. Lee K. Toutanova K., Bert: Pre-training of deep bidirectional transformers for language understanding. arXivpreprint arXiv:1810.04805 (2018).
68. Conneau A. Khandelwal K. Goyal N. Chaudhary V. Wenzek G. Guzmán F. and Stoyanov V., Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116* (2019).
69. Eisenschlos J.M. Ruder S. Czapla P. Kardas M. Gugger S. and Howard J., MultiFiT: Efficient multi-lingual language model fine-tuning. *arXiv preprint arXiv:1909.04761* (2019).
70. Chung H.W. Garrette D. Tan K.C. Riesa J., Improving multilingual models with language-clustered vocabularies. arXivpreprint arXiv:2010.12777, (2020).

Notes

[1] https://github.com/Mesay-Gemeda/Odio-BERT

[2] https://huggingface.co/Mesay/Odio-BERT

[3] **Twitter was renamed to X. In this paper, we use the previous name Twitter and refer to messages on Twitter as tweets.**

[4] https://www.kaggle.com/datasets/wajidhassanmoosa/multilingual-hatespeech-dataset/data

Sections

List of Illustrations

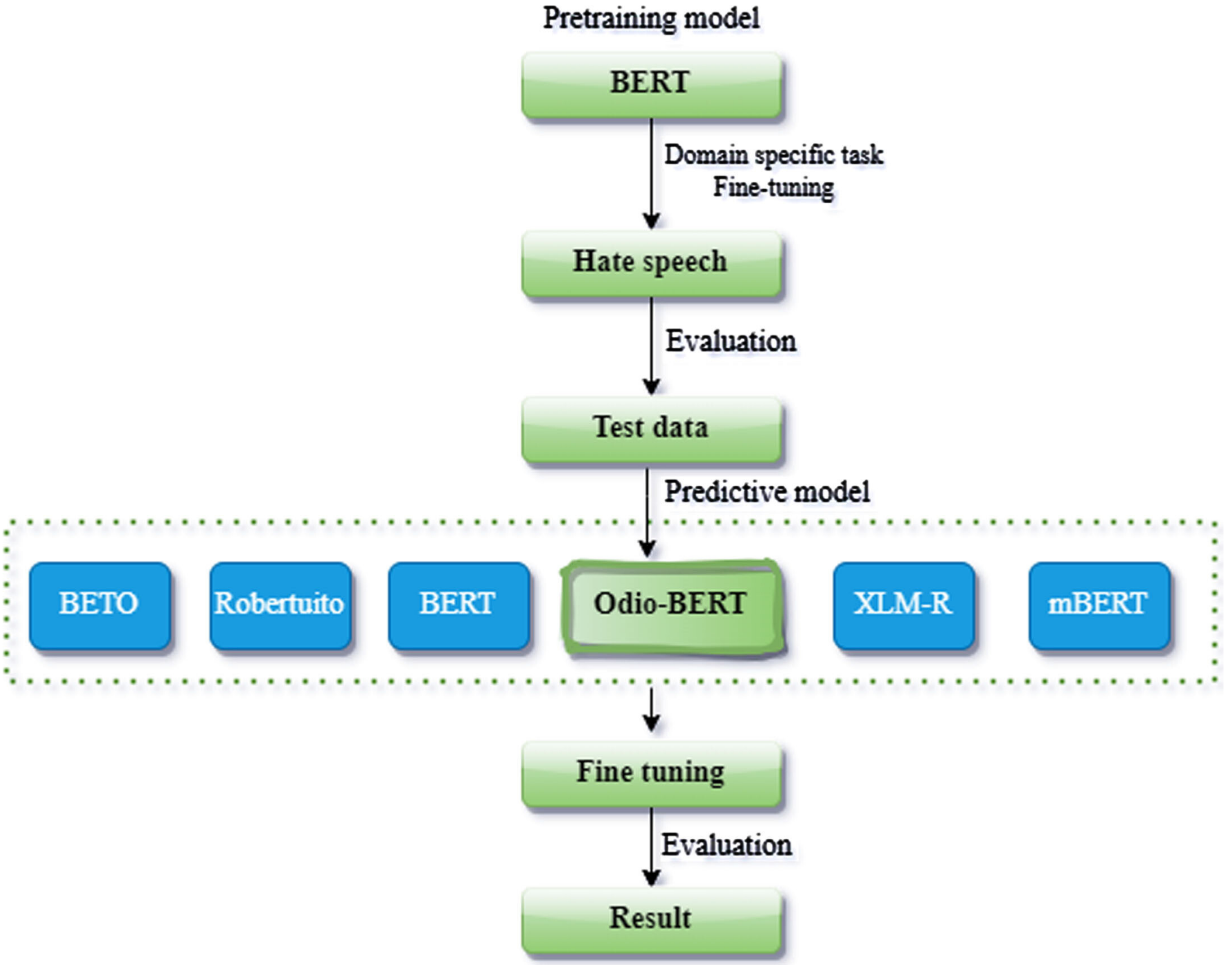

Fig. 1 Proposed framework for Spanish hate speech detection.

Table 1 Summary of the benchmark dataset for Spanish hate speech

Dataset	Label	Size in #	Size in%	Total size
Spanish MisoCorpus 2020	Misogyny	3700	44.1%	8390
	Not-misogyny	4690	55.9%	
Multi-lingual Hate Speech Dataset	Hate	4239	34.1%	12424
	Not-hate	8184	65.9%	
HOMO-MEX	LGBT+phobic	862	16.6%	5223
	Not-LGBT+phobic	4360	83.4%	
HaSCoSVa-2022	Hate	556	13.9%	4000
	Not-hate	3444	86.1%	
Hater-Net	Hate	4433	73.9%	6000
	Not-hate	1567	26.1%	
HatEval 2019	Hate	2739	41.5%	6600
	Not-hate	3861	58.5%	

Table 2 The macro average F1 scores on the respective test datasets for models fine-tuned using **N** entries, were computed by averaging results from 5 random seeds for each N. The most outstanding performance for a particular N is highlighted in **bold**

Odio-BERT												
N	**10**	**20**	**30**	**100**	**200**	**300**	**1,000**	**2,000**	**3,000**	**10,000**	**20,000**	**30,000**
BERT-fine-tuned	0.2	0.46	0.3	0.5	0.52	0.52	0.58	0.63	0.63	0.75	**0.83**	0.83

Table 3 Experimental results of the models including our model

Models	Metrics			
	Precision	Recall	F1 Score	Accuracy
mBERT	0.61	0.6	0.6	0.67
XLM-R	0.65	0.63	0.63	0.67
Robertuito	0.82	0.73	0.77	0.77
BETO	0.81	0.7	0.75	0.76
BERT	0.76	0.69	0.72	0.74
Odio-BERT	0.85	0.76	**0.80**	0.85