Applied Soft Computing Journal 187 (2026) 114259

Contents lists available at ScienceDirect

Applied Soft Computing

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

A transformer-based deep learning approach for detecting online hate speech in Spanish

Jesus M. Sanchez-Gomez [a,*], Fernando Batista [b,c], Miguel A. Vega-Rodríguez [a], Carlos J. Pérez [d]

[a] Universidad de Extremadura, Departamento de Tecnología de Computadores y Comunicaciones, Campus Universitario S/N, 10003 Cáceres, Spain
[b] INESC-ID Lisboa, Rua Alves Redol, 9, 1000-029 Lisboa, Portugal
[c] Iscte - Instituto Universitário de Lisboa, Av. Forças Armadas, 1649-026 Lisboa, Portugal
[d] Universidad de Extremadura, Departamento de Matemáticas, Campus Universitario S/N, 10003 Cáceres, Spain

HIGHLIGHTS

- The task of detecting online hate speech in Spanish language has been addressed.
- A transformer-based deep learning approach (SHS-ALBETO) has been developed.
- Experimentation has been carried out using HatEval dataset.
- SHS-ALBETO has been compared with competing models and with the state-of-the-art.
- The advantages achieved in the performance of SHS-ALBETO have been analyzed.

ARTICLE INFO

Keywords:
Hate speech
Natural language processing
Deep learning
Transformer models

ABSTRACT

The amount of content published on the Internet has grown exponentially in recent times. Social networks have enabled this content to reach an even wider audience. However, the freedom of communication provided by these networks can consequently facilitate the spread of offensive language and hate speech. Although social media platforms have attempted to implement mechanisms for detecting and addressing such content, it remains an ongoing challenge, particularly for languages other than English, such as Spanish. One promising approach to tackle this problem is the application of Natural Language Processing (NLP) tools, which rely on the use of language models and deep learning for text classification. In this work, an approach for detecting Spanish Hate Speech with ALBETO (SHS-ALBETO) is proposed. Experimentation is conducted with HatEval dataset. The performance of SHS-ALBETO is compared with other competing models, such as BERT, BETO, and DistilBETO, along with other proposals from the state-of-the-art. SHS-ALBETO has improved the existing results in the scientific literature, simultaneously providing reduced computing times. Additionally, analyses of the results have revealed its advantages together with challenging aspects that must be addressed to further improve the performance of this kind of approach.

1. Introduction

Offensive language or hate speech has become a popular topic in the last few years mainly due to the exponential growth in the use of social networks (Twitter, Facebook, Instagram, and YouTube, among others), where users can communicate freely [1]. According to [2], hate speech is defined as “public speech that expresses hatred or encourages violence toward a person or group based on race, religion, gender, sexual orientation, or any other diversity feature”.

* Corresponding author.
Email addresses: jmsanchezgomez@unex.es; jesusmsa@inf.uc3m.es (J.M. Sanchez-Gomez), fernando.batista@inesc-id.pt (F. Batista), mavega@unex.es (M.A. Vega-Rodríguez), carper@unex.es (C.J. Pérez).
URL: https://ror.org/0174shg90; https://ror.org/03ths8210 (J.M. Sanchez-Gomez), *URL:* https://ror.org/04mqy3p58; https://ror.org/014837179 (F. Batista), *URL:* https://ror.org/0174shg90 (M.A. Vega-Rodríguez), *URL:* https://ror.org/0174shg90 (C.J. Pérez).

https://doi.org/10.1016/j.asoc.2025.114259
Received 25 July 2024; Received in revised form 15 September 2025; Accepted 10 November 2025
Available online 11 November 2025

Media coverage of this issue is concurrently growing as political attention increases [3]. However, automatic hate speech detection is a very challenging task [4]. Recently, the main social networks agreed with the European Commission on the European Union code of conduct for hate speech, which aims to prevent and avoid the online spread of illegal hate speech [5]. With the aim of improving the quality of service, the social media companies are interested in detecting and removing the existing hate speech on their platforms. Nevertheless, automatic hate speech detection is technically a very difficult task, since the line between appropriate free expression and hate speech is quite blurry [6,7]. In addition, there is a lack of data collections and monitoring about hate speech, making it difficult to research it in an standardized way, especially in languages other than English as it is the case with Spanish. The specific context of this language involves extra challenges compared to others, such as the dialectical variations existing in different geographical regions (European Spanish vs. Latin American Spanish) and the limited resources for training detection systems, which affect their accuracy and robustness [8]. This has resulted in a substantial current research gap in detecting hate speech in Spanish.

In the scientific literature, there are different automatic methods for classifying texts as hate speech [9]. They can be categorized into two main groups: traditional shallow classification methods and deep learning methods. On the one hand, shallow classification methods focus on the use of traditional word representation techniques for encoding words and applying classifiers to carry out the detection. Some popular methods are semantics-based approaches, clustering-based word representation, and classification-based models. On the other hand, deep learning methods utilize deep neural networks. Specifically, deep learning techniques can equally be classified into methods based on static embeddings, which take advantage of distributed representation of words [10], and modern transformer-based methods, which use the self-attention mechanism to capture long-range dependencies and correlations [11]. These transformer-based methods, which were introduced recently, involve pre-trained language models. Among them, one of the most widely used models is the Bidirectional Encoder Representations from Transformers (BERT, [12]), known for its superior performance.

In this work, a transformer-based deep learning approach for Spanish Hate Speech detection with ALBETO, called SHS-ALBETO, is proposed. To evaluate its performance, HatEval dataset [13] is used for carrying out the experiments. For this, the standard evaluation metrics in natural language processing tasks are considered: Accuracy, Precision, Recall, and F1 scores. The main contributions of this work are as follows:

- For the first time, an approach for Spanish Hate Speech detection with ALBETO (SHS-ALBETO) is proposed.
- SHS-ALBETO is compared with other competing models, such as multilingual BERT, BETO (Spanish BERT), and DistilBETO (a distilled version of BETO).
- SHS-ALBETO is compared with the traditional and deep learning methods existing in the scientific literature.
- A comprehensive analysis is carried out to examine the advantages produced by the use of SHS-ALBETO.

The remainder of this paper is structured as follows. Section 2 presents the related work. Section 3 describes the SHS-ALBETO approach, including other competing models, as well as the datasets and evaluation metrics considered. Section 4 reports the results obtained in the conducted experiments, including experimental settings, comparisons, and analyses. Finally, Section 5 presents the conclusions drawn from this study and outlines future work.

2. Related work

This section reviews the approaches that perform the task of detecting hate speech in Spanish, following a chronological order.

The detection of hate speech in Spanish was one of the tasks included in the International Workshop on Semantic Evaluation (SemEval-2019, [13]), where numerous proposals were presented. Firstly, the proposals that employed traditional classification methods are reviewed. The proposed baseline used a linear support vector machine based on term frequency-inverse document frequency (tf-idf) representation. Almatarneh et al. [14] and Ameer et al. [15] utilized a bag-of-words model with tf-idf feature representation values. Furthermore, [14] combined this model with word embeddings to improve the performance of the classifier. In the method proposed by Argota Vega et al. [16], linguistically motivated features and various types of n-grams were considered to train a support vector machine model. An approach combining word embeddings, semantic similarity, tf-idf, and n-grams was employed by Benito et al.[17]. Moreover, [18] proposed a model consisting of a linear classifier based on a support vector machine. This model incorporated n-grams, sentiment analysis, and word embeddings as the main machine learning characteristics. A combination of three prediction algorithms within a voting ensemble classifier model was proposed by Plaza-del-Arco et al. [19]. The algorithms used were logistic regression, decision tree, and support vector machine. In the proposal of [20], three models based on random forest, support vector machine, and logistic regression were applied and combined in an ensemble setting. These traditional classification methods were employed to tackle the task of hate speech detection in Spanish.

Continuing with SemEval-2019, the proposals that considered deep learning-based methods are the following. In Benballa et al. [21], a model based on feature-level dynamic meta-embedding was proposed. Two different model architectures, long short-term memory and convolutional neural network with hybrid long short-term memory, were analyzed by Bojkovský and Pikuliak [22]. Different neural transfer learning techniques, combined with word embeddings, were used by Gertner et al. [23]. In the work of [24], a framework based on genetic programming was used to combine predictions from different knowledge sources for text classification. The development of a long short-term memory model with an embedding layer was addressed by Manolescu et al. [25]. A neural classifier utilizing word embeddings and long short-term memory layers was developed by Montejo-Ráez et al. [26]. In the approach of [27], a model that combined n-gram embeddings within a feed-forward neural network was proposed. The design of different neural network architectures for testing word representations and diverse corpora was used by Nina-Alcocer [28]. A model based on recurrent neural networks that learned compositional numerical representations of words based on character sequences was presented by Paetzold et al. [29]. In the work of [30], some approaches to language modeling which contained word-level n-gram and character-level neural language models were presented. An approach that merged a recurrent neural network based on bi-long short-term memory with an attention mechanism was proposed by De la Peña [31]. Several linear classifiers and recurrent neural networks specifically trained using classical and recent features, such as bag-of-words, bag-of-characters, word embeddings, and contextualized word representations, were presented by Pérez and Luque [32]. A single multilingual system architecture for hate speech detection, consisting of a dictionary of unique characters, key values, and transformed binary arrays, was used by Raiyani et al. [33]. A convolutional neural network was developed by Ribeiro and Silva [34] using word embedding models like GloVe and FastText. A system based on word embeddings and convolutional neural networks, considering both dilated and traditional convolution layers, was presented by Winter and Kern [35]. Lastly, in the work of [36], several shallow and deep learning approaches were analyzed, including engineered features such as n-grams and word embeddings, and neural network methods such as multilayer perceptrons, convolutional neural networks, long short-term memory, and BERT. These studies showcased the diverse range of deep learning techniques employed by participants in SemEval-2019 for hate speech detection.

After the SemEval-2019 workshop, several approaches were proposed, most of which continued to utilize deep learning methods. A multi-channel BERT fine-tuning model was proposed by Sohn and Lee

[37]. This model integrated hidden features from separate BERT models trained in different languages. In the work of [38], three approaches were presented: a supervised machine learning approach utilizing techniques such as naive Bayes, support vector machine, logistic regression, decision tree, and ensemble voting classifier; a deep learning approach based on long short-term memory; and a lexicon-based approach. A cross-lingual contextual word embeddings model was applied by Ranasinghe and Zampieri [39] for text classification, performing inter-task and inter-language transfer learning. In the method of [40], a Spanish language model based on average word embeddings and linguistic features was evaluated with three different machine learning classifiers: random forest, SMO (Sequential Minimal Optimization) support vector machine, and linear support vector machine. A comparison of different pre-trained language models was conducted by Plaza-del-Arco et al. [41], including two multilingual models (multilingual BERT and a cross-lingual language model) and BETO, a monolingual model specifically trained on Spanish. A novel zero-shot cross-lingual transfer learning pipeline based on pseudo-label fine-tuning of transformer language models was presented by Zia et al. [42]. This methodology involved using the cross-lingual classifier to obtain pseudo-labels for training the model. Lastly, [43] studied the most effective features for detecting hate speech and how they can be combined. For that, a system based on linguistic features, knowledge integration, and transformers was developed. These approaches showcase the ongoing development and refinement of deep learning methods for hate speech detection beyond the SemEval-2019 workshop.

The chronological review highlights the shift from traditional classification methods to deep learning methods, particularly to transformer-based methods, which have gained popularity in recent years [11]. This is because traditional classification methods do not properly detect the evolution of hate speech patterns, and exhibit a limited ability to capture complex linguistic nuances. These weaknesses are addressed by deep learning methods, which are capable of learning complex context-dependent features such as irony and sarcasm. Furthermore, transformer-based methods go further by capturing long-range dependencies in texts, thus achieving state-of-the-art performance [9]. Additionally, it is worth noting that in addition to the proposals presented in the SemEval-2019 workshop, all the reviewed approaches conducted their experiments using the HatEval dataset. Thus, this dataset has been used as a common benchmark for evaluating hate speech detection, allowing for fair comparisons and assessments of different approaches.

3. Methodology

This section includes the description of the SHS-ALBETO approach and the description of other competing models used for comparison. It presents the datasets and, in addition, the evaluation metrics.

3.1. Description of the SHS-ALBETO approach

The presented approach, Spanish Hate Speech detection with ALBETO (SHS-ALBETO) is based on transformers, which is a deep learning framework capable of learning the contextual relations between words in a text [44]. Transformers consist of an encoder and a decoder mechanism. The encoder processes the input text through multiple layers of self-attention and feed-forward neural networks to generate a representation of the input sequence. The decoder uses the encoder's representation, along with other inputs, to generate an output sequence.

SHS-ALBETO has adopted a language model from the BERT family, a multi-layer bidirectional transformer encoder [12], which is a deep learning approach used for natural language processing tasks. Particularly, the ALBETO model is considered in this approach. ALBETO is a version of ALBERT (a lite version of BERT) exclusively trained on Spanish corpora [45]. It is more efficient regarding the number of total parameters since it uses a weight-tied strategy, which consists of sharing all parameters across all layers of the model. More specifically, ALBETO-base model is used. It has 12 self-attention layers (encoder transformer blocks) with 12 attention-heads each, 768 hidden layers, and 12 million total parameters. The technical details of how this model works are shown in Fig. 1. Three different parts can be distinguished. First, the input preprocessing, where tokenization and transformation to vector representation by word embeddings are performed. Secondly, the transformer model as such, with the 12 encoder blocks. In turn, each encoder block contains 12 attention heads, composed of the multi-head self-attention mechanism and a feed-forward neural network. And third, the output consists of the output tokens in the form of vectors, which are contextualized representations, in addition to the classification layer. This layer consists of a new feed-forward neural network with a softmax layer to predict the probability that the input belongs to the different classes.

Now, the operation of SHS-ALBETO is detailed. The full approach has been implemented using the Python programming language. Both the *Transformers* and the *TensorFlow* libraries have been used for the development [46]. Fig. 2 shows the processing flowchart followed by SHS-ALBETO.

The steps of the flowchart are presented next:

1. Loading the dataset. The training and test sets are properly formatted and loaded.
2. Loading the pre-trained model and tokenizer. The language model is loaded from the state-of-the-art pre-trained *Transformers* API, whose technical details are provided above. In addition, the tokenizer is loaded from this API.
3. Data tokenization and formatting. The tokenizer is used to convert the input text (training and test sets) into tokens. This sequence of tokens is used as the input to the pre-trained transformer model.
4. Model setup, compilation, and training. The learning parameters, such as the drop-out rate, learning rate, loss function, optimizer, and weight decay, are defined in this step. Then, the *TensorFlow* library is used for building the model. Finally, the training parameters (batch size and number of epochs) are determined for the training.
5. Model evaluation. The SHS-ALBETO approach is evaluated by predicting the classes using the test set and comparing the predictions with the ground truth. Additionally, a classification report is generated, containing the adopted evaluation metrics, such as accuracy, precision, recall, and F1 score.

Following these steps, the SHS-ALBETO approach aims to detect hate speech in Spanish.

3.2. Description of other competing models

The SHS-ALBETO approach is based on ALBETO, a transformer-based language model belonging to BERT family. In addition to SHS-ALBETO, other competing models are now introduced, with which it is compared in Section 4.2. They are: Multilingual BERT, BETO (Spanish BERT), and DistilBETO (a distilled version of BETO).

Firstly, the BERT model was trained with English corpora, as it was originally created for only English language. However, a later version was already trained with multilingual corpora: Multilingual BERT (mBERT). mBERT was created by dumping the entire Wikipedia for each one of the top 104 languages. Particularly, the base version is considered, which has 12 self-attention layers with 12 attention-heads each, 768 hidden layers, and 110 million total parameters.

The other competing models, BETO and DistilBETO, have been specifically trained in Spanish language. On the one hand, BETO [47] was trained by collecting text from different sources: the entire Wikipedia for Spanish language and all the sources of OPUS Project,[1]

[1] https://opus.nlpl.eu/

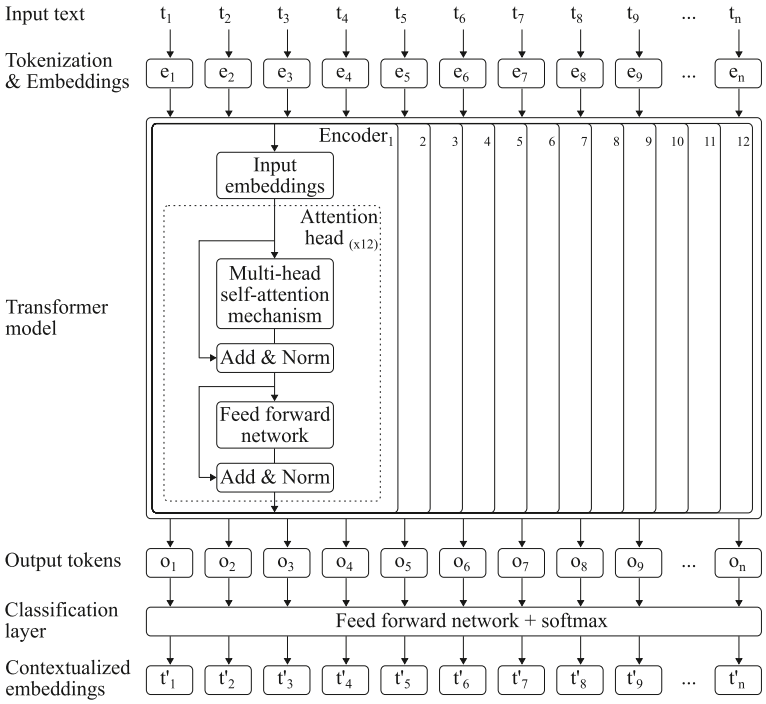


Fig. 1. Architecture of the ALBETO model. Key aspects such as the input, the transformer-based model, and the output are presented.

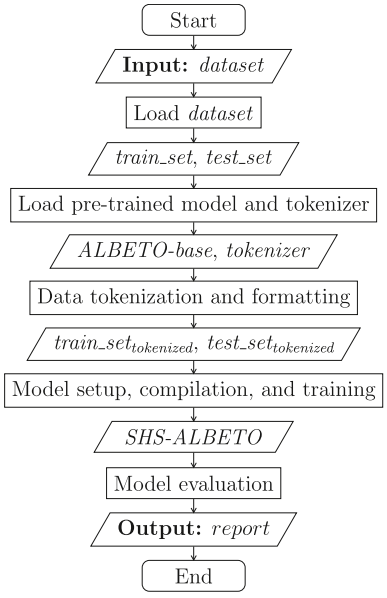


Fig. 2. Flowchart followed by the SHS-ALBETO approach.

additionally for Spanish language. It has a similar size to mBERT, i.e., 12 self-attention layers with 16 attention-heads each, 1024 hidden layers, and 110 million total parameters. On the other hand, DistilBETO [45] is a version of DistilBERT (a distilled version of BERT), which is exclusively trained on Spanish corpora. The architecture of DistilBETO is based on BETO with the exception of removing the token-type embeddings and the pooler layer, and it has 5 self-attention layers, 1024 hidden layers, and 67 million total parameters. Furthermore, the tiny version of ALBETO is considered ($ALBETO_T$). This model has 4 self-attention layers, 312 hidden layers, and 5 million of total parameters.

Table 1
Information about HatEval dataset.

Feature	Description
Language	Spanish
Country	Spain
Source	Twitter
Date	July to September 2018
Size	6600 tweets
Labeled data	4500, 500, and 1600 tweets (training, validation, and test)

These models provide different options for transformer-based language models in Spanish, each with its own architecture and size. In this way, their comparison with SHS-ALBETO in the results section helps to assess their performance and suitability for hate speech detection in Spanish.

3.3. Datasets

The HatEval dataset is used for the experimentation. It was released for SemEval-2019 Task 5 [13], and focuses on detecting hateful content in tweets (posted messages on Twitter) against two targets: immigrants and women. A total of 6600 labeled tweets in Spanish language are provided for training, validation, and testing. Table 1 presents some features of this considered dataset.

Now, since the tweets contained in HatEval dataset are binary classified as hate speech or not, Table 2 shows the counts of the labeled tweets from these two classes. These counts provide an overview of the distribution of hate speech and non-hate speech tweets, which are used for training, validating, and evaluating the performance of the SHS-ALBETO approach. The data for training and testing are distributed following approximately a 3:1 ratio. In addition, in the training set, 10 % of the data is allocated for validation. It is worth noting that the distribution of the two classes in each set is the same (58 % for non-hate speech and 42 % for hate speech). More details about the dataset can be found in Basile et al. [13].

Table 2
Counts of the labeled tweets from non-hate speech and hate speech in HatEval dataset.

Class	Training	Validation	Test
Non-hate speech	2643	278	940
Hate speech	1857	222	660
Total	4,500	500	1,600

3.4. Evaluation metrics

The standard evaluation metrics for natural language processing tasks are considered, i.e., Accuracy, Precision, Recall, and F1 scores [13]. The Accuracy score measures the overall correctness of the model's predictions, and it is computed as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad (1)$$

being TP the number of true positives, TN the number of true negatives, FP the number of false positives, and FN the number of false negatives.

Precision measures the ratio of correctly predicted instances of a particular class to the total number of instances predicted for that class:

$$Precision = \frac{TP}{TP + FP}. \quad (2)$$

Recall calculates the ratio of correctly predicted instances of a particular class to the total number of instances of that class in the gold standard:

$$Recall = \frac{TP}{TP + FN}. \quad (3)$$

Finally, F1 score is the harmonic mean of Precision and Recall:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}. \quad (4)$$

In order to make comparisons with results from the state-of-the-art, macro-averaged versions of Precision, Recall, and F1 are used. Macro-averaged scores are calculated independently for each class, and then the average is taken across all classes. Therefore, from now on, these metrics are referred to as $Precision_M$, $Recall_M$, and $F1_M$, respectively.

4. Experimental results

This section describes the experimental settings, the comparison of SHS-ALBETO with other competing models, an analysis of its advantages, and the comparison of SHS-ALBETO with the state-of-the-art.

4.1. Experimental settings

The SHS-ALBETO approach and the competing models have been trained and fine-tuned with specific hyperparameter values. Moreover, to ensure fair comparisons, all models have been configured with the same training setups and hyperparameters. The learning parameters include the drop-out rate, learning rate, loss function, optimizer, and weight decay, while the training parameters consist of the batch size and the number of epochs. Table 3 shows the tested values for every hyperparameter and the final selected value for each one. The validation data from HatEval dataset has been considered for this experimentation, and the performance has been assessed with $F1_M$ score. First, for each learning parameter, the values indicated in the order listed have been tested (with the exception of the loss function and the optimizer, which are kept constant throughout the experimentation). After that, experiments have been conducted on the training parameters in the same way. In order to assess the model robustness, a sensitivity analysis is conducted.

Table 3
Tested values and selected values of the hyperparameters for SHS-ALBETO and the competing models.

Hyperparameter	Tested values	Selected value
Drop-out rate	0.01, 0.05, 0.1, 0.2, 0.3	0.1
Learning rate	10^{-5}, $2{\cdot}10^{-5}$, $5{\cdot}10^{-5}$, 10^{-4}, 10^{-3}	$2{\cdot}10^{-5}$
Loss function	Binary cross entropy	Binary cross entropy
Optimizer	Adam	Adam
Weight decay	10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}	10^{-2}
Batch size	1, 2, 4, 8, 16, 32, 64	32
Number of epochs	1, 2, 3, …, 18, 19, 20	(see Table 4)

Table 4
$F1_M$ scores obtained by SHS-ALBETO and other competing models for each number of epochs in validation data of HatEval dataset. The best values appear in bold.

No. of epochs	SHS-ALBETO	$ALBETO_T$	mBERT	BETO	DistilBETO
1	0.737	0.690	0.737	0.780	0.809
2	0.806	0.756	0.807	0.829	0.833
3	0.810	0.788	0.825	0.839	0.838
4	0.821	0.790	0.825	0.841	0.835
5	0.831	0.808	0.831	0.855	0.852
6	**0.838**	0.827	**0.841**	0.848	0.851
7	0.819	**0.833**	0.838	0.854	0.842
8	0.832	0.819	0.821	**0.857**	**0.855**
9	0.831	0.809	0.826	0.852	0.852
10	0.830	0.822	0.822	0.844	0.851
11	0.820	0.812	0.820	0.851	0.848
12	0.819	0.796	0.829	0.849	0.846
13	0.829	0.795	0.826	0.840	0.846
14	0.820	0.807	0.820	0.845	0.840
15	0.823	0.818	0.821	0.847	0.840
16	0.826	0.824	0.812	0.850	0.846
17	0.828	0.817	0.834	0.846	0.850
18	0.823	0.806	0.816	0.848	0.836
19	0.830	0.813	0.828	0.843	0.850
20	0.813	0.814	0.823	0.850	0.841

The average percentage of variation in the performance for each hyperparameter studied is reported: 2.48 % for drop-out rate, 1.55 % for learning rate, 2.70 % for weight decay, and 4.88 % for batch size.

Particularly, the number of epochs has been studied for each model, in order to analyze the effect of increasing this number (from 1 to 20). Table 4 presents the $F1_M$ scores obtained with SHS-ALBETO and with the other competing models for every number of epochs in the validation data of HatEval dataset.

It can be observed in Table 4 that the best values for the number of epochs have ranged from 6 to 8. As for the model robustness, the percentage of variation for number of epochs in every model is: 13.70 % for SHS-ALBETO, 20.72 % for $ALBETO_T$, 14.11 % for mBERT, 9.87 % for BETO, and 5.69 % for DistilBETO. Besides, the selection of the best number of epochs is supported by Fig. 3, which displays the curve for the $F1_M$ score and the loss function obtained by the five models analyzed. Each figure shows that the $F1_M$ score attains its maximum when loss function reaches its minimum. Note that no overfitting arises.

Therefore, the subsequent experiments have been performed with the optimal number of epochs for each model. With these configurations, the SHS-ALBETO approach and the other models have been trained and evaluated using the combined training and validation subsets from HatEval dataset, while the test data has been used to assess their performance.

Regarding the experimental platform, the experiments have been performed using Google Colab,[2] which provides a single CUDA device Tesla T4, with 16GB RAM. Version 8302 has been used for the NVIDIA CUDA® Deep Neural Network library (cuDNN), and the models have been developed using Python (version 3.7.14).

[2] https://colab.research.google.com/

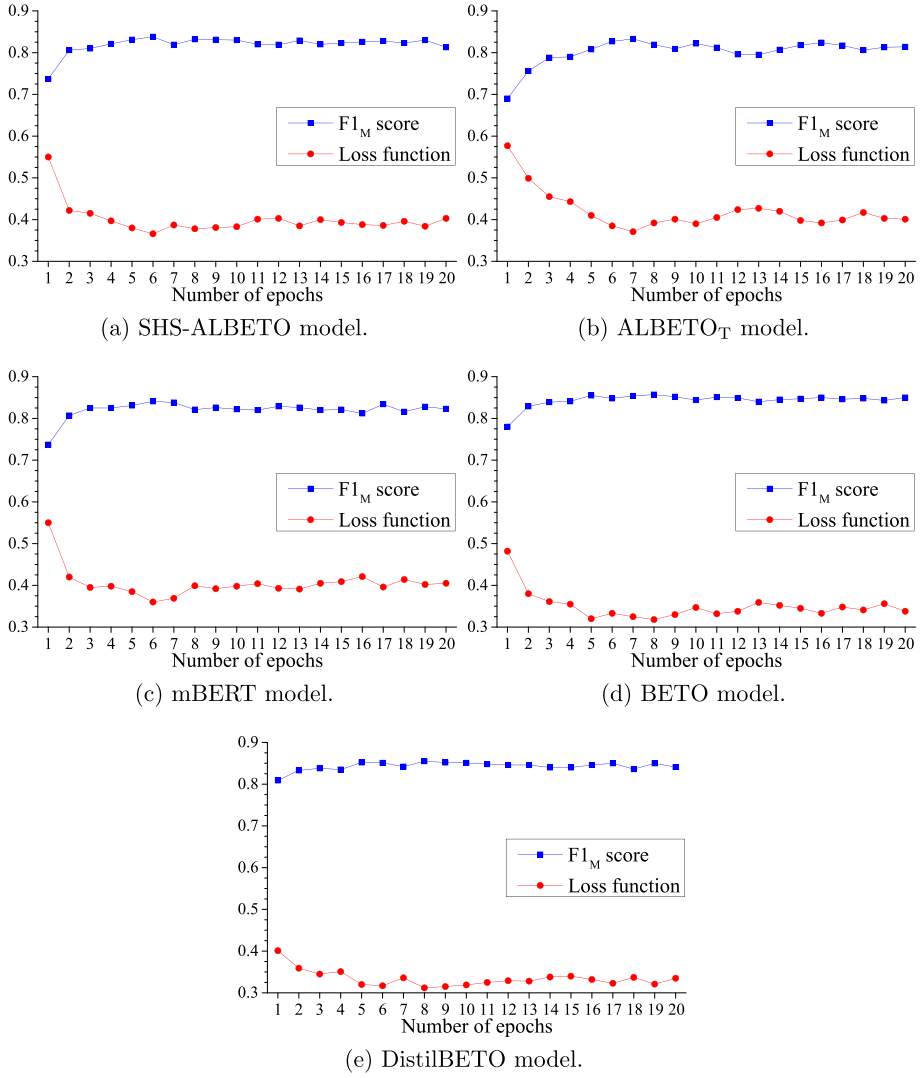


Fig. 3. $F1_M$ score and loss function curves obtained by SHS-ALBETO and other competing models on HatEval dataset.

Table 5
Comparison of the performance and execution time (ET; seconds) for SHS-ALBETO and other competing models on HatEval dataset.

Model	Accuracy	$Precision_M$	$Recall_M$	$F1_M$	ET (s)
SHS-ALBETO	0.785	0.779	0.785	0.781	241
$ALBETO_T$	0.754	0.747	0.740	0.743	132
mBERT	0.727	0.734	0.740	0.726	1819
BETO	0.769	0.768	0.777	0.767	2248
DistilBETO	0.776	0.769	0.774	0.771	1465

4.2. Comparing SHS-ALBETO with other competing models

This subsection presents the comparison between the SHS-ALBETO approach and the competing models of $ALBETO_T$, mBERT, BETO, and DistilBETO. Table 5 shows the results for the evaluation metrics considered, and the execution times (ET) are additionally presented. Furthermore, Fig. 4 graphically illustrates the results obtained.

The results reported in Table 5 reveal that SHS-ALBETO has achieved the highest performance in all the evaluation metrics. DistilBETO has obtained the second best result in Accuracy, $Precision_M$, and $F1_M$ scores, while BETO has achieved the second best result in $Recall_M$. As for the execution times, SHS-ALBETO has achieved the second fastest execution time, only surpassed by $ALBETO_T$, being one order of magnitude lower than the rest of the models. To further analyze the results, Table 6 presents the percentage improvements achieved by the SHS-ALBETO approach with respect to the compared models.

Based on the results from Table 6, SHS-ALBETO has demonstrated improvements over all the compared models. On average, SHS-ALBETO has achieved improvements of 3.83 % in Accuracy, 3.29 % in $Precision_M$, 3.65 % in $Recall_M$, and 3.95 % in $F1_M$ scores in comparison to the other models. More specifically, the percentage improvement average in the four evaluation metrics studied with respect to $ALBETO_T$ is 4.90 %. As for the other models, these improvements are: 6.94 % regarding mBERT, being 6.55 times faster; 1.59 % regarding BETO, being 8.33 times faster; and 1.29 % regarding DistilBETO, being 5.08 times faster. Overall, the SHS-ALBETO approach demonstrated superior performance compared to the other competing models, with improvements in evaluation metrics while maintaining competitive execution time.

4.3. Analyzing the advantages of SHS-ALBETO

In this subsection, an extensive analysis of the SHS-ALBETO approach is performed in order to examine its advantages when it is applied to detect hate speech in Spanish.

Firstly, Table 7 compares the counts of the number of true positives, true negatives, false positives, and false negatives obtained by SHS-ALBETO and by the competing models in HatEval dataset. The last

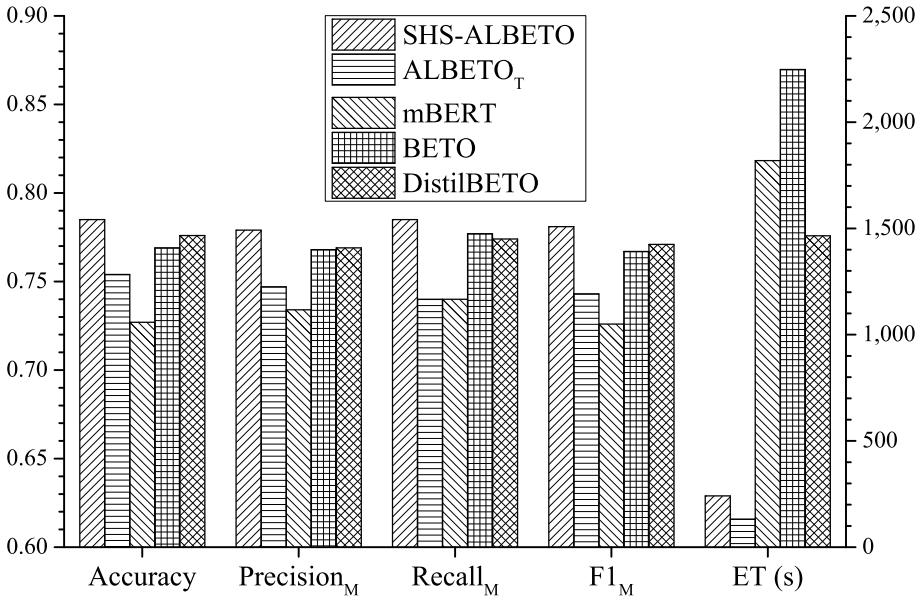


Fig. 4. Comparison of the performance metrics and execution times (ET; seconds) for SHS-ALBETO and for other competing models on HatEval dataset.

Table 6
Percentage improvements obtained from the comparison between SHS-ALBETO and other competing models analyzed on HatEval dataset.

Model	Accuracy	Precision$_M$	Recall$_M$	F1$_M$	ET
ALBETO$_T$	4.11 %	4.28 %	6.08 %	5.11 %	−45.23 %
mBERT	7.98 %	6.13 %	6.08 %	7.58 %	654.77 %
BETO	2.08 %	1.43 %	1.03 %	1.83 %	832.78 %
DistilBETO	1.16 %	1.30 %	1.42 %	1.30 %	507.88 %
Average	3.83 %	3.29 %	3.65 %	3.95 %	487.55 %

Table 7
Number of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) obtained by SHS-ALBETO approach and the competing models in HatEval dataset.

Model	TP	TN	FP	FN
SHS-ALBETO	516	740	200	144
ALBETO$_T$	438	768	172	222
mBERT	539	624	316	121
BETO	543	687	253	117
DistilBETO	505	736	204	155
In common	354	514	73	54

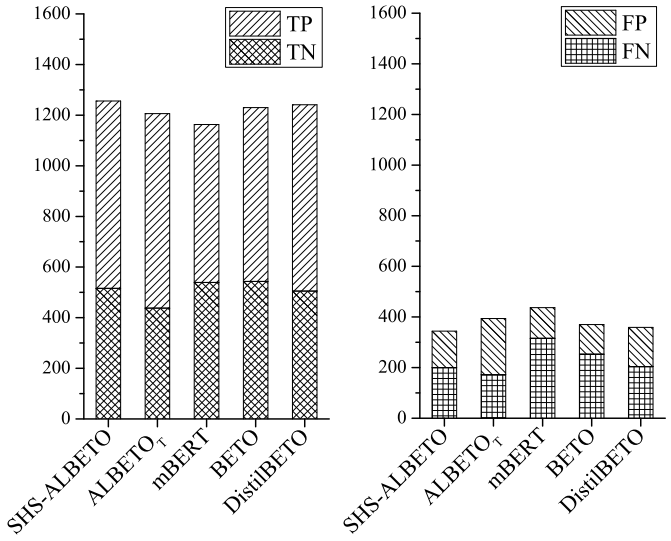


Fig. 5. Counts of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) obtained by SHS-ALBETO approach and the competing models in HatEval dataset.

row presents the number of instances shared by all. These results reveal that more than half of the tweets, 54.25 %, are accurately predicted by all models, indicating a reasonable agreement among them. However, a small percentage of 7.94 % are wrongly predicted by all of them, suggesting some common challenges in identifying hate speech in those instances. In order to provide further visual support, Fig. 5 depicts these counts.

In order to analyze in depth the source of errors in model predictions, Table 8 shows a count of False Positives and False Negatives obtained by all models in common. The source or error type is classified into four categories: context (irony, sarcasm, rhetorical language, etc.), dialectical variations, typographical errors (spelling or writing), and mislabeled tweets by expert annotators of HatEval dataset. As can be seen, the context is the main cause of wrong model predictions, accounting for 85 % of the False Positives and 69 % of the False Negatives in the language models. Then, the second main cause is the dialectical variations (around 12–13 % in False Positives and False Negatives) that the Spanish language has, since it is a language spoken in around twenty countries. The

Table 8
Analysis of error types in tweets classified as False Positives (FP) and False Negatives (FN) by SHS-ALBETO approach and the competing models in HatEval dataset.

Error type	FP	FN
Context	62	37
Dialectical variations	9	7
Typographical errors	1	6
Incorrectly labeled tweets	1	4
Total	73	54

Table 9
Examples of correctly predicted tweets (and their English translation) by SHS-ALBETO approach and incorrectly predicted by all the other models ($ALBETO_T$, mBERT, BETO, and DistilBETO) in HatEval dataset. 0 means non-hate speech, and 1 means hate speech.

Tweet text	True label	Predicted label	
		SHS-ALBETO	Other models
Y una valla de seis metros con concertinas en la salida de todos los vuelos que vengan de UK que ya estoy hasta los cojones de simios. (And a six-meter fence with concertina mesh at the exit of all the flights that come from the UK because I am fucking sick of apes.)	1	1	0
Las mujeres son la puta cagada mi hermano (Women are the fucking shit my brother)	1	1	0
117 inmigrantes subsaharianos dejan a siete guardias heridos en un salto a la valla de Ceuta (117 Sub-Saharan inmigrants leave seven guards injured in a jump over the Ceuta fence)	0	0	1
Calla, que yo también soy "mora de mierda", "gitana asquerosa" o "sudaca", algunos aún no se han enterado de que soy andaluza, hija y nieta de andaluces y parece que tampoco me integro. (Shut up, I am also a "shit Moorish", "disgusting gypsy" or "spic", some have not yet found out that I am Andalusian, daughter and granddaughter of Andalusians and it seems that I do not fit in either.)	0	0	1

Table 10
Examples of incorrectly predicted tweets (and their English translation) by all the models. 0 means non-hate speech, and 1 means hate speech.

Tweet text	True label	Predicted label
Si algún día tengo una hija lo primero que voy a enseñarle es a trabajar, para que no tenga que andar de puta para que le regalen todo. (If one day I have a daughter the first thing I am going to teach her is to work, so she does not have to be a whore to get everything for free.)	1	0
A este le dejaba yo enganchado en las concertinas. (I left this one hooked on the concertinas.)	1	0
Especialmente más ricos y con más inmigrantes son estos 3. Y sí, hay causa efectos. Otra cosa es que solo os guste "la negra" Niurka Montalvo cuando gana medallas como a Le Pen le gusta "el moro" Mbappé cuando le hace ganar un mundial. (Especially richer and with more immigrants are these 3. And yes, there are causes and effects. Another thing is that you only like "the black" Niurka Montalvo when she wins medals as Le Pen likes "the Moorish" Mbappé when he makes her win a World Cup.)	0	1
Q estupides q diga q una mujer es puta por "perrear" capaz hay una q no y es tremenda trola aunque de igual forma no les tendria pq importar (How stupid to say that a woman is a whore for "grind", maybe there is one that is not, and it is tremendous lie although in the same way they would not have to care)	0	1

typographical errors, very common in social networks and even more so in the hurried writing that characterizes this social network, also influence the erroneous predictions of the models. Lastly, several incorrectly labeled tweets have been found in the dataset that led to False Positives and False Negatives. All these types of errors are hereafter analyzed in detail.

Now, Table 9 provides some examples of correctly predicted tweets by SHS-ALBETO and that are wrongly predicted by the competing models. The first two tweets contain hate speech, as both tweets include insults directed at immigrants and women, respectively. The context is clear, but all the other models fail in their predictions. As for the last two tweets, they actually do not contain hate speech. The third tweet does not include insults, but it has been wrongly predicted by all the other models. And the last tweet contains many insults, but in that context they are not offensive. In this case, SHS-ALBETO gets its prediction right, while the other models do not.

Furthermore, several examples of incorrectly predicted tweets by SHS-ALBETO and all the competing models are shown in Table 10. This demonstrates that hate speech detection is a challenging task. As can be observed, these wrongly predicted tweets have in common the importance of the context. In the first tweet, the language models did not label this tweet as hate speech because they did not recognize the misogynistic undertone of the use of the term "whore" as an insult to denigrate women, failing to adequately understand the context of sex discrimination. The second tweet does not contain any insult, but the models ignored the violent implication of the phrase "hooked on the

Table 11
Examples of incorrectly labeled tweets by the annotators of HatEval dataset (and their English translation), with the predictions by SHS-ALBETO approach and other models. 0 means non-hate speech, and 1 means hate speech.

Tweet text	True label	Predicted label
Miriam trasmite lo mismo que Cepeda, una puta mierda. Claro que Mireya no se merecía la nominación, pero Cepeda si. (Miriam transmits the same as Cepeda, a fucking piece of shit. Of course, Mireya did not deserve the nomination, but Cepeda did.)	1	0
Aquí un señor humorista llamando "Mala Puta" a Inés Arrimadas (supuestamente) (Here a comedian gentleman calling Inés Arrimadas "Bad Whore" (supposedly))	1	0
hola, solo te escribo para llamarte puta. Un saludo y espero que mueras a manos de un moro inmigrante. (hi, I am only writing to call you a whore. Greetings and I hope you die at the hands of a Moorish immigrant.)	0	1
Con la cara de monos que tienen así guarros de donde iban a ser? Colombianos, la policia se pudo los guantes y los hizo salir a todos al rellano, al sudaka que es asqueroso, el típico barrigón sucio… (With the monkey face they have like that, where would they be from? Colombians, the police put on their gloves and made them all go out onto the landing, the sudaca who is disgusting, the typical dirty paunchy…)	0	1

concertinas", which suggests harm or injury to another person. In addition, this tweet promotes hostility toward immigrants, as concertinas are placed at borders between countries. As for the third tweet, the models mistakenly classify it as hate speech because it includes several insults related to people of another race or religion ("the black" and "the Moorish"), when it actually criticizes the selective admiration of people based on their achievements. In this case, the models have failed to understand the nuanced criticism and the underlying sociopolitical context. And the error in the prediction of the last tweet is along the lines of the previous one. It seems that the models focused on the use of the word "whore", often considered derogatory, without understanding that the tweet criticizes women being judged for their actions. In this case, the tweet questions harmful stereotypes and defends women's autonomy.

Lastly, some examples of incorrectly labeled tweets by the expert annotators of HatEval dataset are presented in Table 11. These instances represent examples of human errors in data labeling, which induce an adverse impact on the performance of SHS-ALBETO and other models, potentially leading to less accurate results. The first and second tweets have been labeled as hate speech, but it does not appear to be the case, so the predictions seem to be correct. Regarding the third and fourth tweets, the annotators labeled them as non-hate speech, even though there are explicit insults against women and immigrants.

In conclusion, the conducted analyses have highlighted several challenges and issues that language models face when dealing with social media language, where users often do not prioritize careful writing or clear explanations. Sometimes both grammatical and spelling mistakes are present in online content, which negatively impacts in the performance of these language models. Moreover, the models can struggle with other expressions in human language, such as sarcasm and irony, which may be more or less implicit in the context. Furthermore, it is shown that certain content poses challenges to be accurately labeled even by experts in the field, thus demonstrating that it is very difficult to achieve high accuracy rates in models' predictions. All this highlights the value of the better performance obtained by SHS-ALBETO.

4.4. Comparing SHS-ALBETO with the state-of-the-art

Finally, the comparison of the SHS-ALBETO approach with the state-of-the-art proposals is presented in this subsection. Table 12 shows the results in the evaluation metrics considered for SHS-ALBETO and for each one of the proposals in the scientific literature. The symbol "-" is presented when a proposal does not report the corresponding score.

Table 12
Comparison of the performance for SHS-ALBETO approach and for state-of-the-art proposals on HatEval dataset.

Proposal	Accuracy	Precision_M	Recall_M	F1_M
SHS-ALBETO	0.785	0.779	0.785	0.781
SVC Baseline (Basile et al. [13])	0.705	0.701	0.707	0.701
CiTIUS-COLE (Almatarneh et al. [14])	0.660	–	–	0.640
CIC (Ameer et al. [15])	–	–	–	0.727
MineriaUNAM (Argota Vega et al. [16])	–	–	–	0.730
GSI-UPM (Benito et al. [17])	0.728	0.726	0.733	0.725
UA (Perelló et al. [18])	0.731	–	–	0.722
SINAI (Plaza-del-Arco et al. [19])	0.711	0.707	0.713	0.707
Grunn2019 (Zhang et al. [20])	0.708	–	–	0.701
Saagie (Benballa et al. [21])	–	–	–	0.717
STUFIIT (Bojkovský and Pikuliak [22])	0.710	0.700	0.700	0.700
MITRE (Gertner et al. [23])	–	–	–	0.729
INGEOTEC (Graff et al. [24])	0.710	–	–	0.710
TuEval (Manolescu et al. [25])	0.630	0.618	0.617	0.617
SINAI-DL (Montejo-Ráez et al. [26])	–	–	–	0.686
Tw-StAR (Mulki et al. [27])	0.700	0.700	0.710	0.700
HATERecognizer (Nina-Alcocer [28])	0.735	–	–	0.729
UTFPR (Paetzold et al. [29])	–	–	–	0.664
UNBNLP (Parizi et al. [30])	0.710	0.700	0.690	0.690
GL (De la Peña [31])	0.723	0.717	0.722	0.718
Atalaya (Pérez and Luque [32])	0.731	–	–	0.730
Vista.ue (Raiyani et al. [33])	–	0.596	0.593	0.594
INF-HatEval (Ribeiro and Silva [34])	0.696	0.708	0.712	0.696
Know-Center (Winter and Kern [35])	–	–	–	0.720
ltl-uni-due (Zhang et al. [36])	–	–	–	0.696
MC-BERT4HATE (Sohn and Lee [37])	0.769	–	–	0.766
Vote-uni + bi (Plaza-del-Arco et al. [38])	0.754	0.747	0.739	0.742
XLM-R (Ranasinghe and Zampieri [39])	–	–	–	0.751
AWE + LF (García-Díaz et al. [40])	–	–	–	0.755
BETO (Plaza-del-Arco et al. [41])	–	0.777	0.786	0.776
ZS-CL (Zia et al. [42])	–	–	–	0.730
KI-LF-BF (García-Díaz et al. [43])	0.771	–	–	0.768
Average others	0.716	0.700	0.702	0.711

From the results reported in Table 12, where 32 different proposals from the last five years are compared, it can be concluded that SHS-ALBETO has achieved the best results in Accuracy, Precision_M, and F1_M scores, and the second best one in Recall_M. In particular, the average percentage improvements achieved by SHS-ALBETO regarding the state-of-the-art proposals are 9.93 % for Accuracy, 11.87 % for Precision_M, 12.45 % for Recall_M, and 10.24 % for F1_M scores. Fig. 6 visually summarizes the comparison between the SHS-ALBETO approach

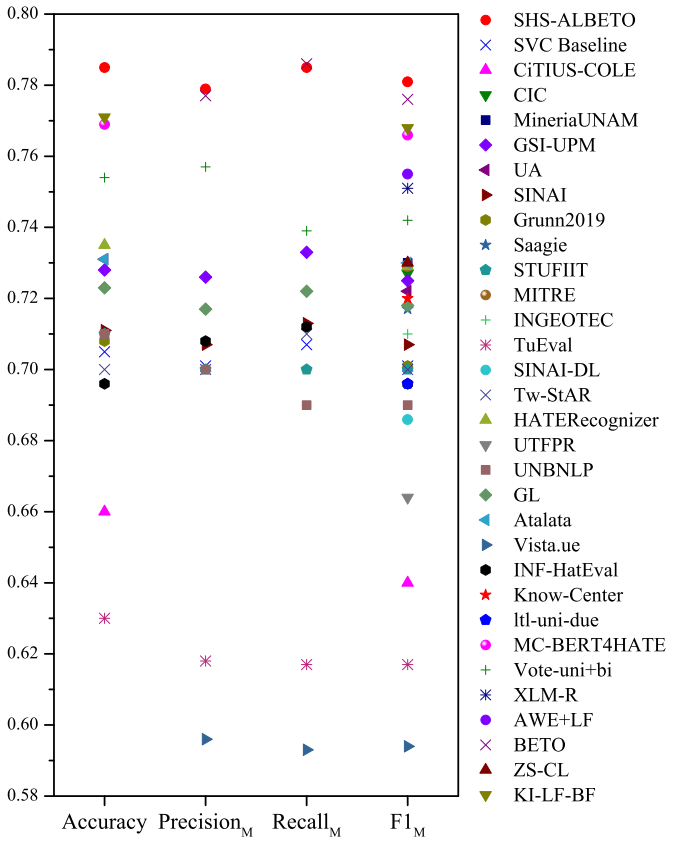


Fig. 6. Comparison of performance metrics for SHS-ALBETO approach and for state-of-the-art proposals on HatEval dataset.

and state-of-the-art proposals, presenting their results across all the evaluation metrics analyzed.

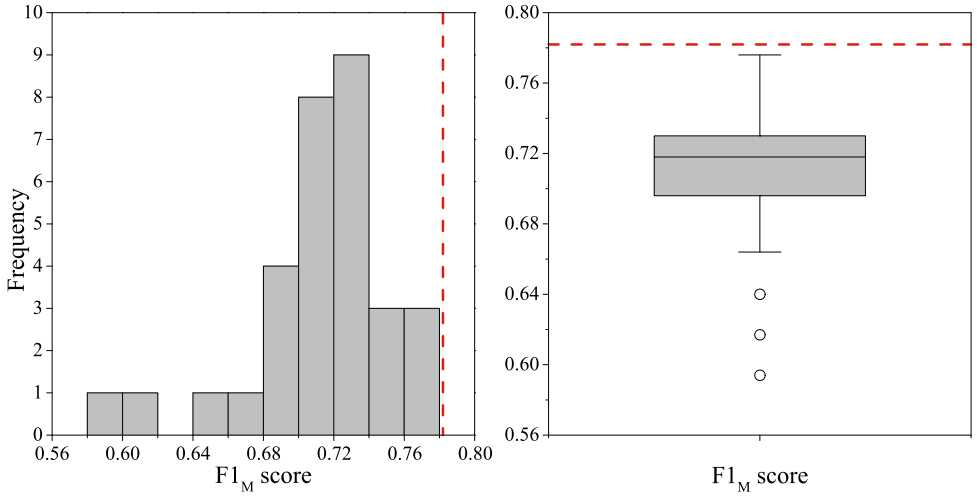
In order to show in a visual way the improvement of SHS-ALBETO, Fig. 7 presents the histogram and the boxplot for the $F1_M$ scores of the state-of-the-art proposals together with a dashed red line marking the $F1_M$ score provided by SHS-ALBETO. As can be seen, SHS-ALBETO has outperformed the state-of-the-art proposals when solving the task of detecting online hate speech in Spanish. It can be observed that the $F1_M$ scores obtained by the state-of-the-art proposals are concentrated in the range of 0.59–0.78. Besides, the distribution is negatively skewed. There are several lower outliers and the average and median values (0.711 and 0.718, respectively) are slightly shifted. Overall, this histogram and the boxplot emphasize the improvement achieved by the SHS-ALBETO approach in terms of $F1_M$ scores, providing a clear visual representation of how SHS-ALBETO outperforms the state-of-the-art proposals, further supporting the conclusion drawn from the obtained results.

5. Conclusions

The proliferation of social media content has led to an increase in hate speech by users. The task of automatic detection of hate speech is very difficult and challenging, particularly in the Spanish language, which has still received little attention so far.

In this work, an approach based on a transformer-based deep learning model, Spanish Hate Speech detection with ALBETO (SHS-ALBETO), is developed for detecting online hate speech in Spanish online content. SHS-ALBETO is trained, analyzed, and evaluated by using Spanish tweets from HatEval dataset, and its performance is assessed using common natural language processing metrics. For this purpose, SHS-ALBETO is compared with other competing models, such as multilingual BERT, BETO, and DistilBETO, and the obtained results have shown that SHS-ALBETO has improved upon them. The comparison with the existing state-of-the-art proposals is conducted, and SHS-ALBETO has outperformed them obtaining average percentage improvements of 9.93 % for Accuracy, 11.87 % for Precision$_M$, 12.45 % for Recall$_M$, and 10.24 % for $F1_M$ scores. Moreover, an exhaustive analysis of the advantages of SHS-ALBETO has identified several challenges when addressing hate speech detection: special features of the context of human language, such as sarcasm or irony, the presence of grammatical and spelling errors, and even the difficulties in categorizing certain content as hate speech or not by experts. Overall, this work has highlighted the potential of the SHS-ALBETO approach to automatically recognize hate speech in Spanish content despite its technical difficulties.

Regarding future research, the SHS-ALBETO approach will be validated on new datasets that capture possible new trends in online hate speech in the Spanish language. More specifically, the dataset developed by the Hatemedia Project [48] (publicly available during 2026) will be used. Moreover, another plan is to expand this analysis to include other datasets from different real-world scenarios, such as online forums and



(a) Histogram for $F1_M$ scores. (b) Boxplot for $F1_M$ scores.

Fig. 7. Histogram and boxplot for $F1_M$ scores obtained by the state-of-the-art proposals. $F1_M$ score of SHS-ALBETO approach is represented as a dashed red line.

other social networks, which may present unique linguistic and sociolinguistic phenomena related to the production and perception of hate speech. Furthermore, it is intended to incorporate the latest GPT-based models, which have demonstrated superior performance in various natural language tasks. The use of these models could enhance the detection and analysis of hate speech in online contexts. Additionally, the use of generative adversarial networks will be explored to assess the models' noise-handling capabilities and overall robustness.

CRediT authorship contribution statement

Jesus M. Sanchez-Gomez: Writing – original draft, Visualization, Validation, Software, Investigation, Formal analysis, Data curation. **Fernando Batista:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Miguel A. Vega-Rodríguez:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Carlos J. Pérez:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization.

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 research was supported by Ministry of Science, Innovation and Universities - Spain and State Research Agency - Spain (Projects PID2022-137275NA-I00, PID2021-122209OB-C32, PID2023-148577OB-C21, and RED2022-134540-T funded by MICIU/AEI/10.13039/501100011033), Junta de Extremadura - Spain (Projects GR24017 and GR24013), Fundação para a Ciência e a Tecnologia - Portugal (Project UIDB/50021/2020), and European Union (European Regional Development Fund). Jesus M. Sanchez-Gomez was supported by Junta de Extremadura and European Union (European Social Fund) under the doctoral fellowship PD18057.

Data availability

The authors do not have permission to share data.

References

[1] A. Rahul Katarya, Hate speech, toxicity detection in online social media: a recent survey of state of the art and opportunities, Int. J. Inf. Secur. 23 (2024) 577–608.
[2] Cambridge English Dictionary, Definition of hate speech (2023). https://dictionary.cambridge.org/us/dictionary/english/hate-speech Last accessed: 15 September 2025.
[3] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Comput. Surv. 51 (2018) 85.
[4] A. Gandhi, P. Ahir, K. Adhvaryu, P. Shah, R. Lohiya, E. Cambria, S. Poria, A. Hussain, Hate speech detection: a comprehensive review of recent works, Expert Syst. 41 (2024) e13562.
[5] European Union, The EU code of conduct on countering illegal hate speech online (2022). https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en Last accessed: 15 September 2025.
[6] A. Rawat, S. Kumar, S.S. Samant, Hate speech detection in social media: techniques, recent trends, and future challenges, Wiley Interdiscip. Rev. Comput. Stat. 16 (2024) e1648.
[7] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, O. Frieder, Hate speech detection: challenges and solutions, PLoS One 14 (2019) e0221152.
[8] G. Castillo-López, A. Riabi, D. Seddah, Analyzing zero-shot transfer scenarios across spanish variants for hate speech detection, in: Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023), Association for Computational Linguistics, 2023, pp. 1–13, https://doi.org/10.18653/v1/2023.vardial-1.1
[9] J.S. Malik, H. Qiao, G. Pang, A. van den Hengel, Deep learning for hate speech detection: a comparative study, Int. J. Data Sci. Anal. 20 (2025) 3053–3068.
[10] S.S. Birunda, R.K. Devi, A review on word embedding techniques for text classification, in: Innovative Data Communication Technologies and Application, vol. 59, Springer, 2021, pp. 267–281, https://doi.org/10.1007/978-981-15-9651-3_23
[11] W. Cunha, F. Viegas, C. França, T. Rosa, L. Rocha, M.A. Gonçalves, A comparative survey of instance selection methods applied to non-neural and transformer-based text classification, ACM Comput. Surv. 55 (2023) 265.
[12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, Association for Computational Linguistics, 2019, pp. 4171–4186, https://doi.org/10.18653/v1/N19-1423
[13] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F.M.R. Pardo, P. Rosso, M. Sanguinetti, Semeval-2019 task 5: multilingual detection of hate speech against immigrants and women in Twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 54–63, https://doi.org/10.18653/v1/S19-2007
[14] S. Almatarneh, P. Gamallo, F.J.R. Pena, Citius-COLE at semeval-2019 task 5: combining linguistic features to identify hate speech against immigrants and women on multilingual tweets, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 387–390, https://doi.org/10.18653/v1/S19-2068
[15] I. Ameer, M.H.F. Siddiqui, G. Sidorov, A. Gelbukh, Cic at Semeval-2019 task 5: simple yet very efficient approach to hate speech detection, aggressive behavior detection, and target classification in Twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 382–386, https://doi.org/10.18653/v1/S19-2067
[16] L.E.A. Vega, J.C. Reyes-Magaña, H. Gómez-Adorno, G. Bel-Enguix, Mineriaunam at Semeval-2019 task 5: detecting hate speech in Twitter using multiple features in a combinatorial framework, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 447–452, https://doi.org/10.18653/v1/S19-2079
[17] D. Benito, O. Araque, C.A. Iglesias, GSI-UPM at semeval-2019 task 5: semantic similarity and word embeddings for multilingual detection of hate speech against immigrants and women on Twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 396–403, https://doi.org/10.18653/v1/S19-2070
[18] C. Perelló, D. Tomás, A. Garcia-Garcia, J. Garcia-Rodriguez, J. Camacho-Collados, UA at Semeval-2019 task 5: setting a strong linear baseline for hate speech detection, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 508–513, https://doi.org/10.18653/v1/S19-2091
[19] F.M. Plaza-Del-Arco, M.D. Molina-González, M.T. Martín-Valdivia, L.A. Ureña-López, Sinai at Semeval-2019 task 5: ensemble learning to detect hate speech against inmigrants and women in English and spanish tweets, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 476–479, https://doi.org/10.18653/v1/S19-2084
[20] M. Zhang, R. David, L. Graumans, G. Timmerman, Grunn2019 at Semeval-2019 task 5: shared task on multilingual detection of hate, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 391–395, https://doi.org/10.18653/v1/S19-2069
[21] M. Benballa, S. Collet, R. Picot-Clemente, Saagie at Semeval-2019 task 5: from universal text embeddings and classical features to domain-specific text classification, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 469–475, https://doi.org/10.18653/v1/S19-2083
[22] M. Bojkovský, M. Pikuliak, STUFIIT at semeval-2019 task 5: multilingual hate speech detection on Twitter with MUSE and elmo embeddings, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 464–468, https://doi.org/10.18653/v1/S19-2082
[23] A.S. Gertner, J. Henderson, E. Merkhofer, A. Marsh, B. Wellner, G. Zarrella, Mitre at Semeval-2019 task 5: transfer learning for multilingual hate speech detection, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 453–459, https://doi.org/10.18653/v1/S19-2080
[24] M. Graff, S. Miranda-Jiménez, E. Tellez, D.A. Ochoa, INGEOTEC at semeval-2019 task 5 and task 6: a genetic programming approach for text classification, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 639–644, https://doi.org/10.18653/v1/S19-2114
[25] M. Manolescu, D. Löfflad, A.N.M. Saber, M.M. Tari, Tueval at Semeval-2019 task 5: LSTM approach to hate speech detection in English and Spanish, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 498–502, https://doi.org/10.18653/v1/S19-2089
[26] A. Montejo-Ráez, S.M. Jiménez-Zafra, M.A. Garcia-Cumbreras, M.C. Díaz-Galiano, SINAI-DL at semeval-2019 task 5: recurrent networks and data augmentation by paraphrasing, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 480–483, https://doi.org/10.18653/v1/S19-2085
[27] H. Mulki, C.B. Ali, H. Haddad, I. Babaoğlu, Tw-Star at Semeval-2019 task 5: n-gram embeddings for hate speech detection in multilingual tweets, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 503–507, https://doi.org/10.18653/v1/S19-2090
[28] V. Nina-Alcocer, Haterecognizer at Semeval-2019 task 5: using features and neural networks to face hate recognition, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 409–415, https://doi.org/10.18653/v1/S19-2072
[29] G.H. Paetzold, S. Malmasi, M. Zampieri, UTFPR at semeval-2019 task 5: hate speech identification with recurrent neural networks, in: Proceedings of the 13th

International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 519–523, https://doi.org/10.18653/v1/S19-2093
[30] A.H. Parizi, M. King, P. Cook, UNBNLP at semeval-2019 task 5 and 6: using language models to detect hate speech and offensive language, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 514–518, https://doi.org/10.18653/v1/S19-2092
[31] G.L. De la Peña, Gl at Semeval-2019 task 5: identifying hateful tweets with a deep learning approach, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 416–419, https://doi.org/10.18653/v1/S19-2073
[32] J.M. Pérez, F.M. Luque, Atalaya at Semeval 2019 task 5: robust embeddings for tweet classification, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 64–69, https://doi.org/10.18653/v1/S19-2008
[33] K. Raiyani, T. Gonçalves, P. Quaresma, V. Nogueira, Vista.ue at semeval-2019 task 5: single multilingual hate speech detection model, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 524–528, https://doi.org/10.18653/v1/S19-2094
[34] A. Ribeiro, N. Silva, INF-hateval at semeval-2019 task 5: convolutional neural networks for hate speech detection against women and immigrants on Twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 420–425, https://doi.org/10.18653/v1/S19-2074
[35] K. Winter, R. Kern, Know-center at Semeval-2019 task 5: multilingual hate speech detection on Twitter using CNNS, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 431–435, https://doi.org/10.18653/v1/S19-2076
[36] H. Zhang, M. Wojatzki, T. Horsmann, T. Zesch, Ltl.uni-due at Semeval-2019 task 5: simple but effective lexico-semantic features for detecting hate speech in Twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 441–446, https://doi.org/10.18653/v1/S19-2078
[37] H. Sohn, H. Lee, MC-bert4hate: hate speech detection using multi-channel BERT for different languages and translations, in: 2019 International Conference on Data Mining Workshops (ICDMW), IEEE, 2019, pp. 551–559, https://doi.org/10.1109/ICDMW.2019.00084
[38] F.-M. Plaza-Del-Arco, M.D. Molina-González, L.A. Ureña-López, M.T. Martín-Valdivia, Detecting misogyny and xenophobia in Spanish tweets using language technologies, ACM Trans. Internet Technol. 20 (2020) 12.
[39] T. Ranasinghe, M. Zampieri, Multilingual offensive language identification with cross-lingual embeddings, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2020, pp. 5838–5844, https://doi.org/10.18653/v1/2020.emnlp-main.470
[40] J.A. García-Díaz, M. Cánovas-García, R. Colomo-Palacios, R. Valencia-García, Detecting misogyny in spanish tweets. An approach based on linguistics features and word embeddings, Futur. Gener. Comput. Syst. 114 (2021) 506–518.
[41] F.M. Plaza-Del-Arco, M.D. Molina-González, L.A. Ureña-López, M.T. Martín-Valdivia, Comparing pre-trained language models for spanish hate speech detection, Expert Syst. Appl. 166 (2021) 114120.
[42] H.B. Zia, I. Castro, A. Zubiaga, G. Tyson, Improving zero-shot cross-lingual hate speech detection with pseudo-label fine-tuning of transformer language models, in: Proceedings of the International AAAI Conference on Web and Social Media, vol. 16, American Association for Artificial Intelligence, 2022, pp. 1435–1439, https://doi.org/10.1609/icwsm.v16i1.19402
[43] J.A. García-Díaz, S.M. Jiménez-Zafra, M.A. García-Cumbreras, R. Valencia-García, Evaluating feature combination strategies for hate-speech detection in Spanish using linguistic features and transformers, Complex Intell. Syst. 9 (2023) 2893–2914.
[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, vol. 30, NeurIPS, 2017, pp. 5998–6008.
[45] J. Cañete, S. Donoso, F. Bravo-Marquez, A. Carvallo, V. Araujo, ALBETO and distilbeto: lightweight Spanish language models, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, ACL Anthology, 2022, pp. 4291–4298.
[46] D. Rothman, Transformers for Natural Language Processing: Build Innovative Deep Neural Network Architectures for NLP with Python, Pytorch, Tensorflow, BERT, RoBERTa, and More, Packt Publishing Ltd, 2021.
[47] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained BERT model and evaluation data, in: Practical Machine Learning for Developing Countries (PML4DC) Workshop, 2020, pp. 1–9.
[48] E. Said-Hung, X. Blanco-Valencia, M.R. Pieretti, Á Martín-Gutiérrez, D.M. Calvanese, Ó.D.G. Vicente, S. Arce-García, Dataset de mensajes analizados para la detección de odio religioso alrededor de contenidos publicados por medios informativos españoles EN x [Data set]. Hatemedia project (2025) Last accessed: 15 September 2025. https://doi.org/10.5281/zenodo.15789729