Comparing Different Approaches for Detecting Hate Speech in Online Portuguese Comments

Bernardo Cunha Matos
INESC-ID Lisboa, Portugal
Instituto Superior Técnico, Lisbon, Portugal

Raquel Bento Santos
INESC-ID Lisbon, Portugal
Instituto Superior Técnico, Lisbon, Portugal

Paula Carvalho
INESC-ID Lisbon, Portugal

Ricardo Ribeiro
INESC-ID Lisbon, Portugal
Iscte – University Institute of Lisbon, Portugal

Fernando Batista
INESC-ID Lisbon, Portugal
Iscte – University Institute of Lisbon, Portugal

Abstract

Online Hate Speech (OHS) has been growing dramatically on social media, which has motivated researchers to develop a diversity of methods for its automated detection. However, the detection of OHS in Portuguese is still little studied. To fill this gap, we explored different models that proved to be successful in the literature to address this task. In particular, we have explored transfer learning approaches, based on existing BERT-like pre-trained models. The performed experiments were based on CO-HATE, a corpus of YouTube comments posted by the Portuguese online community that was manually labeled by different annotators. Among other categories, those comments were labeled regarding the presence of hate speech and the type of hate speech, specifically overt and covert hate speech. We have assessed the impact of using annotations from different annotators on the performance of such models. In addition, we have analyzed the impact of distinguishing overt and and covert hate speech. The results achieved show the importance of considering the annotator's profile in the development of hate speech detection models. Regarding the hate speech type, the results obtained do not allow to make any conclusion on what type is easier to detect. Finally, we show that pre-processing does not seem to have a significant impact on the performance of this specific task.

2012 ACM Subject Classification Computing methodologies → Transfer learning; Social and professional topics → Hate speech; Computing methodologies → Supervised learning; Computing methodologies → Machine learning approaches; Information systems → Clustering and classification

Keywords and phrases Hate Speech, Text Classification, Transfer Learning, Supervised Learning, Deep Learning

Digital Object Identifier 10.4230/OASIcs.SLATE.2022.10

Supplementary Material *Software (Code, Data and Models)*: `https://hate-covid.inesc-id.pt/?page_id=200`

Funding This research was supported by Fundação para a Ciência e a Tecnologia, through the Projects HATE Covid-19 (Proj.759274510), MIMU (FCT PTDC/CCI-CIF/32607/2017), and UIDB/50021/2020.

11th Symposium on Languages, Applications and Technologies (SLATE 2022).
Editors: João Cordeiro, Maria João Pereira, Nuno F. Rodrigues, and Sebastião Pais; Article No. 10; pp. 10:1–10:12
OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Social media environments are fertile to the dissemination of aggressive and harmful content [12, 18]. Some factors that contribute to this dissemination are the platforms' easy access, the users' potential anonymity, and the increased willingness of people to express their opinions online [8, 12]. The development of methods and tools for automatically detecting offensive and abusive language and hate speech (HS) has recently gained traction in the Natural Language Processing (NLP) and Artificial Intelligence (AI) research communities [8, 21]. The non-existence of a unique and consensual definition of HS [8] makes it difficult to clearly distinguish HS from other related phenomena, such as offensive speech, either for humans or algorithms [14]. For the purpose of this work, HS is defined according through the following coexisting conditions [4]:

- HS has a specific target that can be mentioned explicitly or implicitly, which corresponds to vulnerable or historically marginalized groups or individuals targeted for belonging to those groups;
- HS typically spreads or supports hatred, or incites violence against the targets, by disparaging, humiliating, discriminating, or even threatening them based on specific identity factors (e.g., religion, ethnicity, nationality, race, color, descent, gender, sexual orientation);
- HS can be expressed both explicitly (or overtly) and implicitly (or covertly).

The major difficulty of this task is related to the fact that most of hateful comments in social media are covert or implicit, and their interpretation requires information on the sociopolitical and pragmatic context [3]. Moreover, demographic features such as the first language, age, education, and social identity can result in subjective and biased annotations in the corpora used to both train and test OHS detection systems [1].

Recently, the Commissioner for Human Rights, in a memorandum related to Portugal, has noted that, despite the information being provided by civil society organisations indicates low rates of reporting of HS, there is a rise in the number of racially motivated hate crimes and HS [19]. This highlights the need of monitoring the spread of OHS, which is only possible at a large scale by the use of computational methods.

Despite the great popularity of OHS detection, few studies have specifically dedicated to the analysis and detection of European Portuguese OHS. In fact, there is a lack of resources (particularly annotated corpora) specifically designed to support OHS detection in Portuguese [8, 9].

This work uses the recently created CO-HATE (Counter, Offensive and HS) Corpus [4], which is composed by comments on YouTube videos that tackle topics that could potentially generate hatred content. In particular, this corpus focus on the expression of afrophobia, romaphobia, and LGBTQIphobia by the Portuguese online community, since the Afro-descendant, Roma, and LGBTQI communities are among the most commonly reported targets of both offline and online HS in Portugal [11]. The corpus was labeled by five annotators, who followed detailed guidelines developed for this purpose. As expected, the Inter-Annotator Agreement (IAA) among the annotators is relatively low, achieving a Krippendorff's alpha value of 0.478, which reflects the plurality of subjective views on the HS concept. This motivated us to investigate how does the selection of different perspectives for training affects the results of HS detection.

Recent work on OHS detection relies mostly on the use of Deep Learning (DL) methods for both feature extraction and training of classifiers [2, 10, 13, 20]. As reported in literature [17], the use of models such as Convolutional Neural Networks (CNN) and Long Short-Term

Memory Networks (LSTM), among others, suffers from the lack of labelled data. Transfer learning approaches can overcome this issue since they do not require large amounts of labelled data to train models. Furthermore, they are not so time-consuming, and can outperform all the remaining approaches [17, 22, 25]. In this work, we focus on different transfer learning approaches based on an existing pre-trained model called BERT (Bidirectional Encoder Representations from Transformers) [6]. In particular, we compare the results of three different models – BERT-LinearLayer, BERT-CNN, and GAN-BERT – on the automated detection of HS in Portuguese YouTube comments, based on the CO-HATE corpus. We study the impact of using data annotated by different sets of annotators on the performance of the models and also performed some experiments related to the impact of preprocessing. We also study the differences in detecting Covert and Overt HS.

This document is organised as follows: Section 2 presents an overview of the most relevant related literature; Section 3 describes the dataset used; Section 4 describes the different models; Section 5 presents the experimental setup; Section 6 presents the results; finally, Section 7 presents the major conclusions and pinpoints future directions.

2 Related Work

Recently, HS detection has gained particular relevance in areas such as NLP and AI, and different approaches, mostly relying on deep learning methods, have been proposed in the literature.

Kamble et al. [13] explored HS detection in Hindi-English code-mixed tweets. They developed domain specific word embeddings from 255,309 Hindi-English tweets conveying HS and non-hate. They used the Word2Vec algorithm [16] to train the word embeddings model. Using those embeddings as features, they conducted classification experiments with Deep Learning algorithms such as one-dimensional CNN (CNN-1D), LSTM, and BiLSTM. According to the results reported by the authors, CNN-1D resulted in the highest precision, F1-score, and accuracy, while BiLSTM achieved the best recall. Their models were able to better capture the semantics of HS along with their context which resulted in an improvement of about 12% in F1-score over a past work that used statistical classifiers. For offensive language detection, Ong [20] experimented CNN, LSTM, BiLSTM, GRU, BiGRU, and combinations based on these models. The author used GloVe word vectors, some pre-trained using Twitter with 100 and 200 dimensions and others pre-trained with Common Crawl with 300 dimensions. The author concluded that the architecture that had the highest macro average F1-Score was BiLSTM-CNN.

New architectures, namely the BERT-based ones, seem to outperform all the remaining approaches. Ranasinghe et al. [22] presented a multilingual Deep Learning model to identify HS and offensive language in social media, in their submission to the sub-task A of the HASOC 2019 shared task. To make the system portable to all the languages in the dataset, they used only minimal preprocessing methods, such as removing usernames, removing urls, and, depending on the architecture, converting all tokens into lowercase. They experimented multiple DNNs architectures: pooled GRU, LSTM+GRU+attention, two-dimensional CNN (CNN-2D)+Pooling, GRU+capsule, and LSTM+capsule+attention, using FastText as word embeddings. Furthermore, they also experimented with fine-tuned BERT, which outperformed every above mentioned DNN for German, English, and Hindi. Mozafari et al. [17] used two Twitter datasets that were annotated for racism, sexism, hate, and offensive content. They experimented different combinations of BERT with other models, such as CNN and LSTM. The evaluation results indicated that BERT-CNN outperformed previous works by profiting from the syntactical and contextual information embedded in different transformer encoder layers of the BERT using a CNN-based fine-tuning strategy. In OffensEval 2019, a shared task that focus on the detection of offensive language, Zampieri et al. [25] mentioned that among the top-10 teams, seven used BERT with variations in the parameters and in the preprocessing steps. The top-performing team [15] achieved a macro average F1-Score of 0.829, using pre-trained BERT with fine-tuning on the OLID dataset, and hashtag segmentation and emoji substitution as preprocessing.

An improvement to these BERT-based models was proposed by Croce et al. [5]. The authors implemented GAN-BERT using Generative Adversarial Learning. In this model, the generator is trained to produce a sample, and the discriminator to distinguish between generated samples or samples belonging to the training data. BERT is used to encode the input and as the discriminator. The model was tested with a variety of datasets for multiple tasks (topic classification, question classification, and sentiment analysis) obtaining an increase in performance for all of them when compared to BERT.

With these successful approaches in mind, we will use similar models for the classification of HS in social media text, in particular we will fine-tune BERT, and test combinations of BERT with CNN and GAN.

3 Data

The dataset used in our experiments, the CO-HATE Corpus [4], is composed by the comments retrieved from 39 YouTube videos: 20,590 written comments (795,111 tokens), posted by 8,485 different online users. The corpus was annotated by five recruited annotators, who are currently enrolled in a bachelor's or a master's degree in Communication or in Political and Social Sciences. The average age of the annotators is 23 (ranging from 21 to 27 years old) and the annotation team is composed by both individuals belonging to the communities monitored in this study, and by annotators that do not belong to any potentially marginalized group. More specifically, the annotation team includes Portuguese individuals as follows: a female of African descent, a White male who identifies himself as part of the LGBTQ+ community, a female of Roma descent, a White cisgender hetero male, and a White cisgender hetero female. The corpus is subdivided into five parts, each containing approximately 4,000 messages, on average. Each part was randomly assigned to a different annotator. Additionally, all the annotators were assigned to a common part comprehending 534 messages, which was used to measure the agreement between the annotators, and assess the reliability of the annotations assigned to the entire corpus.

These 534 messages are also used as the test set. Given the task subjectivity, and assuming that the profile of human annotators may influence the data annotation, all the messages labeled as conveying HS by at least two annotators will be considered as hatred content. With this voting type the test set is composed of 50% HS messages. We did not consider the messages containing at least one vote in order to discard unintentional errors introduced by the annotator; the possibility of two annotators making a mistake would be a more unlikely scenario. Table 1 presents the distribution of hatred messages used in training, considering both the annotations performed by each annotator and the ones performed by different groups of annotators that were selected following the criteria defined in Section 5.

4 Model Description

In this section, we describe the models we use for detecting OHS, based on the data previously described. We decided to test such models because they have already shown good performance in similar tasks [5, 17, 23].

Table 1 Class distribution for the training data.

Set of annotators	Number of messages	HS (%)
A+B+C+D+E	20,056	35
A+B+D+E	16,039	37
A+B+C	12,036	30
D+E	8,020	43
A	4,008	25
B	4,011	36
C	4,017	29
D	4,014	39
E	4,006	48

4.1 BERT-LinearLayer

BERT [6] is a multi-layer bidirectional transformer encoder. In our experiments, we used BERT base, which contains an encoder with 12 layers (transformer blocks), 12 self-attention heads, and 110 million parameters. As the BERT model is pre-trained on general corpora, we had to fine-tune BERT using our annotated dataset. BERT-LinearLayer is inspired in [23] and [17]. In this architecture, the [CLS] token output of the 12th transformer encoder, a vector of size 768, is given as input to a fully connected network W/o hidden layer. Then, the Sigmoid activation function is applied to the hidden layer in order to make the prediction.

4.2 BERT-CNN

This model is inspired in [23] and consists of two main components. The first one is the BERT model, in which the text is passed through 12 layers of self-attention to obtain contextualized vector representations. The other one is a CNN, which is used as a classifier.

First, the text is given as input to BERT, then the output of the last four hidden layers of the pre-trained BERT are concatenated to get vector representations. Next, these embeddings are passed in parallel into 160 convolutional filters of five different sizes (768x1, 768x2, 768x3, 768x4, and 768x5), 32 filters for each size. Each kernel takes the output of the last four hidden layers of BERT as 4 different channels and applies the convolution operation on it. After that, the output is passed through the ReLU Activation function and a Global Max-Pooling operation. Finally, the output of the pooling operation is concatenated and flattened to be later on passed through a dense layer and a Sigmoid function to get the final binary label.

4.3 GAN-BERT

The GAN-BERT approach is based on GAN-BERT [5]. The input is encoded by a BERT model. The GAN is composed by a generator, trained to produce a sample, and a discriminator to distinguish between generated samples or samples belonging to the training data. The generator is a multi-layer perceptron (MLP) that transforms an input into a vector representation being the [CLS] token used as a sentence embedding. The discriminator is another MLP with a last layer with SoftMax as an activation function to classify the received embedding. The training process consists of optimizing both generator and discriminator losses. The generator loss considers the error induced by the generated examples correctly identified by the discriminator. The discriminator loss considers the error induced by wrongly classifying the labeled data and by not being able to recognize generated samples. The BERT weights will be updated when updating the discriminator. After training, the generator is discarded.

5 Experimental setup

The maximum sequence length of each text sample was set to 350 tokens to avoid overloading the GPU. A substantial amount of messages does not exceed that length and it does not degrade performance. All the models were trained for 15 epochs and the model with the best positive class F1-Score on the development set was saved. The training data was split into 80% and 20% for training and development sets, respectively, preserving the same proportions of examples in each class.

The first experiment was done using the entire training data, i.e., the 20,056 messages annotated by annotators A, B, C, D, and E. Considering the subjectivity of this topic, we have also experimented using the corpus annotated by each user independently (A, B, C, D, and E). Since annotator C was the one having the worse IAA, when compared to the remaining annotators (0.23 on average, while the others have at least 0.531), we tested the combination A+B+D+E. Also, the annotators D and E were the ones that achieved better results independently and the highest agreement rate between them, so we also have tested the combination D+E. Since annotators D and E do not belong to any potential historically marginalized group, we decided to test also the combination A+B+C, composed by annotators that belong to the target communities. This may help us to understand how the annotators' social identity may affect the performance of OHS detection.

For all models, we used two different pre-trained BERT models, namely **Multilingual BERT** (mBERT),[1] and **BERTimbau** (brBERT) [24]. When using the entire corpus, BERTimbau had an overall performance better than mBERT, so for all the other experiments we only tested with BERTimbau and we only report the results obtained using it.

We report both macro and positive class F1-scores, but when assessing the models performance we give particular importance to the positive class F1-score, since it evaluates the models performance on the class that we want to detect.

6 Results and Analysis

The results of all models for the different sets of training data are represented in Table 2. In these experiments, the text was not pre-processed and in the test set we considered a minimum of two votes to decide if a comment contains or not HS. We present the results of our baseline, a dummy classifier that classifies all instances as HS. Considering the positive class F1-score as the benchmark metric, we can see that using the data of A, B, C, and A+B+C always obtains worse results than the baseline model. Also, when using the data of all annotators, A+B+C+D+E, the best result (Positive Class F1-Score of 0.667) was not able to outperform the baseline model. On the other hand, when using the data of D, E, D+E, and A+B+D+E, the results outperformed the baseline.

Comparing the different model architectures, the results suggest that BERT-LinearLayer and BERT-CNN can attain better results than GAN-BERT. In particular, the best result was obtained using the data of D+E and the BERT-CNN model with an F1-score of 0.721.

[1] `https://github.com/google-research/bert/blob/master/multilingual.md`

Table 2 Performance of all models with different data used for train.

Training Data	Model	Positive Class			Macro Avg		
		Prec	Rec	F1	Prec	Rec	F1
[redacted]	Baseline	0.500	1	0.667	0.250	0.500	0.333
A	BERT-LinearLayer	0.604	0.629	0.617	0.609	0.609	0.608
	BERT-CNN	0.601	0.648	0.623	0.609	0.609	0.608
	GAN-BERT	0.578	0.390	0.465	0.559	0.552	0.540
B	BERT-LinearLayer	0.721	0.532	0.612	0.675	0.663	0.657
	BERT-CNN	0.695	0.581	0.633	0.667	0.663	0.661
	GAN-BERT	0.625	0.468	0.535	0.600	0.594	0.587
C	BERT-LinearLayer	0.611	0.629	0.620	0.614	0.614	0.614
	BERT-CNN	0.649	0.547	0.593	0.629	0.625	0.623
	GAN-BERT	0.615	0.562	0.587	0.606	0.605	0.604
D	BERT-LinearLayer	0.603	0.809	0.691	0.657	0.639	0.628
	BERT-CNN	0.589	0.790	0.675	0.636	0.620	0.608
	GAN-BERT	0.623	0.682	0.651	0.636	0.635	0.634
E	BERT-LinearLayer	0.665	0.663	0.664	0.665	0.665	0.665
	BERT-CNN	0.702	0.697	0.699	0.700	0.700	0.700
	GAN-BERT	0.631	0.633	0.632	0.631	0.631	0.631
D+E	BERT-LinearLayer	0.645	0.768	0.701	0.679	0.672	0.669
	BERT-CNN	0.625	0.850	**0.721**	0.696	0.670	0.659
	GAN-BERT	0.630	0.682	0.655	0.641	0.640	0.640
A+B+C	BERT-LinearLayer	0.649	0.610	0.629	0.641	0.640	0.640
	BERT-CNN	0.683	0.614	0.647	0.666	0.665	0.664
	GAN-BERT	0.618	0.442	0.515	0.592	0.584	0.576
A+B+D+E	BERT-LinearLayer	0.712	0.622	0.664	0.688	0.685	0.684
	BERT-CNN	0.679	0.697	0.688	0.684	0.684	0.683
	GAN-BERT	0.656	0.622	0.638	0.648	0.648	0.648
A+B+C+D+E	BERT-LinearLayer	0.636	0.700	0.667	0.651	0.650	0.649
	BERT-CNN	0.688	0.618	0.651	0.670	0.669	0.668
	GAN-BERT	0.648	0.509	0.570	0.622	0.616	0.612

Intuitively, it was expected that BERT-CNN would yield better results, since it uses information that contains both syntactical and contextual features from the last four layers of BERT, which encode more information than the output of the top layer [6], while BERT-LinearLayer and GAN-BERT models only use the [CLS] token of the last transformer encoder. We also believe that the convolutions of the CNN architecture highlighted features related to different writing patterns, which may occur in HS samples, and therefore provide better detection for the aforementioned category.

The low IAA obtained between all the annotators, a Krippendorff's alpha of 0.478, suggested the difficulty of this task even for humans. Nevertheless, reasonable results were obtained when considering the multiplicity of perspectives given by all the annotators: a Positive Class F1-Score of 0.667 (and a macro averaged F1-score of 0.649, considerably better than the one achieved by the baseline). The annotators belonging to the communities targeted in this study (A, B, and C) tended to disagree more with each other than the annotators not belonging to these communities (D and E) [4]. This aspect is also reflected in the models trained with the information provided by such annotators. Every single Positive Class F1-Score obtained by D+E is greater than the Positive Class F1-Scores obtained by A+B+C.

actual	Non-HS	131	136
values	HS	40	227
		Non-HS	HS
		predicted values	

Figure 1 Confusion matrix of the best performing model: BERT-CNN with D+E training data.

We also find important to put into perspective the results we had with the results reported by other researchers on the same task. We did not find any works using Portuguese pre-trained BERT models, so we had to compare our results with works focusing on other languages. Safaya et. al [23] achieved, on average, a macro average F1-score of 0.851 with their BERT-CNN model, using language-specific pre-trained models for Arabic, Greek, and Turkish languages. Although we did not test all the architectures they have tested, just like them, in our work, BERT-CNN yielded better results than using BERT-LinearLayer. Dowlagar et. al [7] achieved a macro average F1-score of 0.883 by fine-tuning an English-specific pre-trained BERT model, the biggest macro F1-score when compared to other machine learning approaches. To the best of our knowledge, no experiences were performed in the supervised case using GAN-BERT. The best macro average F1-score we had was 0.7, which is not in line with the values that we have now reported. Several factors may help understanding this difference. For example, most corpora used by researchers are often created through the use of generic lexical-based approaches, used to retrieve content containing words or expressions with negative polarity. This selection method leaves out an immense set of potentially relevant hatred content, including covert (indirect or implicit) HS, often resorting to rhetorical figures, and apparently neutral words and constructions used to attack or humiliate the HS targets. CO-HATE data was selected using a different approach, allowing the inclusion of messages conveying either overt or covert HS. In fact, covert HS surpasses the frequency of overt HS in this corpus. Given covert HS is much harder to detect than overt HS, we believe that this may rend the task more difficult to the classifiers.

The use of BERTimbau, a Brazilian Portuguese model trained on the BrWaC (Brazilian Web as Corpus), with European Portuguese text extracted from the YouTube platform may also influence the results, since there are lexical differences between these variants of Portuguese that are not covered by BERTimbau Tokenizer. More general aspects such as hyper-parameter optimization can also impact the performance. For instance, varying values of batch size, learning rate, dropout rate (when applicable) and different max length values for text sequences are all variables that may affect the results.

6.1 Error Analysis

This section discusses the major classification errors derived from our best model, BERT-CNN, trained with the data from annotators D and E. In Figure 1, we present the confusion matrix of our best model in order to better visualize the classification errors. As we had already shown in the Table 2, we have interesting results in terms of recall, but the precision is not at the same level. We have inspected the 136 messages that were incorrectly classified as HS and found that a high proportion contain counter-speech. Hence, this suggest that future experiments should include other related HS categories, such as 'Counter-speech', instead of considering only a dichotomous classification. Also, some of these messages include words and expressions that are also highly frequently used in messages classified as conveying hate speech (e.g., *Rendimento Social de Inserção (RSI)*, *abonos*, and *subsidiodependência*; "Social Integration Income", "subsidies", and "subsidy dependence"), which may influence the classifier in the training phase. We have also manually inspected the 40 messages that were incorrectly classified as Non-HS. This analysis suggest that most errors are associated with messages that, out-of-context, are difficult to interpret, as illustrated in Examples 1-4.

(1) *Adoro Portugual*
I love Portugal
(2) *Sim :)*
Yes :)
(3) *Correto !*
Correct !
(4) @João Francisco 👍👍👍👍

6.2 Detecting Overt and Covert HS

We conducted additional experiments on the identification of Overt HS and Covert HS. This is particularly relevant because Covert HS is quite frequent in our data, even more frequent than Overt HS. In fact, our training data contains about 16% of Overt HS and 19% of Covert HS, while the test set contains about 22% of Overt HS and 34% of Covert HS. All the experiments were conducted using BERT-CNN, the model that achieved the best performance in the binary classification of HS, and all the training data were used for trained. The results are presented in Tables 3 and 4, where the baseline consists of classifying all instances as the positive class. When considering the positive class F1-Score, the detection of Overt HS (0.495) was more successful than the detection of Covert HS (0.488). This is an expected result, since covert HS usually tends to use linguistic phenomena, such as sarcasm or irony, much more difficult to detect with an automatic approach. In fact, the lower frequency of Overt HS in the data makes the detection of Overt HS more challenging due to the unbalanced data, but our model was able to still attain a good performance, specially considering the baseline. The detection of Covert HS proved to be difficult to perform using the proposed model.

Table 3 Classification of Overt Hate Speech, using the BERT-CNN model trained with all data.

Model	Overt HS			Macro Avg		
	Prec	Rec	F1	Prec	Rec	F1
Baseline	0.221	1	0.362	0.110	0.500	0.181
BERT-CNN	0.580	0.432	**0.495**	0.715	0.672	0.687

Table 4 Classification of Covert Hate Speech, using the BERT-CNN model trained with all data.

Model	Covert HS			Macro Avg		
	Prec	Rec	F1	Prec	Rec	F1
Baseline	0.335	1	0.502	0.168	0.500	0.251
BERT-CNN	0.648	0.391	0.488	0.696	0.642	0.650

6.3 Pre-processing Impact

We tested the impact of applying pre-processing, which consisted of removing processing errors generated in the data retrieval; anonymizing users' mentions by replacing a user tag with "@UserID" to represent a username with a single word and keep the context and removing repetitions of three or more punctuation signals and emojis. We compiled in Table 5 the best result of each model for each case. We were expecting that linguistic clues just like the repetitions of punctuation signals and emojis could aid the models in capturing HS but their removal accompanied by some other pre-processing techniques led to better results for BERT-LinearLayer and BERT-CNN models, even though they are minor improvements (+0.003 for BERT-LinearLayer and +0.005 for BERT-CNN). On the other hand, the results of GAN-BERT met our expectations and got worsen (-0.008), but again the differences were not significant.

Table 5 Positive Class F1-Scores of the best models on the test data with and without pre-processing.

Model	W/o preproc	W/ preproc
BERT-LinearLayer	0.701	**0.704**
BERT-CNN	0.721	**0.726**
GAN-BERT	**0.655**	0.647

7 Conclusion

In this paper, we compared three different models based on BERT on the task of identifying HS in Portuguese YouTube comments: BERT-LinearLayer, BERT-CNN, and GAN-BERT. BERT-CNN achieved the highest Positive Class F1-Score, taking advantage of both the syntactical and contextual information embedded in the last four transformer encoder layers of the BERT model and the convolutions of the CNN architecture. We have shown how the data chosen to train the models might affect the results. Our best model, BERT-CNN, achieves a Positive Class F1-Score of 0.72, surpassing the baseline in almost 6 p.p. The best result was obtained by combining the data of the annotations from the annotators with the best IAA, which suggests that if we intent to have a good performance using data annotated by different annotators, we should consider the aggregation of annotations from annotators sharing similar perceptions.Although our best macro average F1-score was reasonable, it was not at the same level of other HS detection works that involve other languages. Factors such as the high frequency of Covert HS in our dataset, the use of BERTimbau with European Portuguese, and hyper-parameter optimization may explain these differences in the performance. We performed an Error Analysis of the results provided by our best model and found that some of the messages misclassified as Non-HS do not provide the necessary context for a good classification. Also, a high proportion of messages misclassified as HS correspond, in fact, to counter-speech, and contain words that are frequently used in hatred content. We conducted some experiments for the binary classification of Overt HS and the binary classification of Covert HS in order to investigate if the HS type could affect the performance of our approaches. As expected, the detection of Covert HS proved to be slightly harder than the detection of Overt HS. We also tested the impact of pre-processing the text, and the results show that pre-processing does not seem to have a significant impact on the performance of this specific task.

References

1 Hala Al Kuwatly, Maximilian Wich, and Georg Groh. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online, November 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.alw-1.21.

2 Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760, 2017.

3 Fabienne Baider. Covert hate speech, conspiracy theory and anti-semitism: Linguistic analysis versus legal judgement. *International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique*, pages 1–25, 2022.

4 Paula Carvalho, Danielle Caled, Cláudia Silva, Fernando Batista, and Ricardo Ribeiro. The expression of Hate Speech against Afro-descendant, Roma and LGBTQ+ communities in YouTube comments. *submitted*, 2022.

5 Danilo Croce, Giuseppe Castellucci, and Roberto Basili. GAN-BERT: Generative adversarial learning for robust text classification with a bunch of labeled examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2114–2119, Online, July 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.191.

6 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi:10.18653/v1/N19-1423.

7 Suman Dowlagar and Radhika Mamidi. Hasocone@fire-hasoc2020: Using BERT and multilingual BERT models for hate speech detection. *CoRR*, abs/2101.09007, 2021. arXiv:2101.09007.

8 Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30, 2018.

9 Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104, Florence, Italy, August 2019. Association for Computational Linguistics. doi:10.18653/v1/W19-3510.

10 Björn Gambäck and Utpal Kumar Sikdar. Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online*, pages 85–90, 2017.

11 European University Institute. *Monitoring media pluralism in the digital era: application of the media pluralism monitor in the European Union, Albania and Turkey in the years 2018 2019: country report Portugal.* Publications Office, 2020. doi:10.2870/292300.

12 Md Saroar Jahan and Mourad Oussalah. A systematic review of hate speech automatic detection using natural language processing. *arXiv preprint*, 2021. arXiv:2106.00742.

13 Satyajit Kamble and Aditya Joshi. Hate speech detection from code-mixed hindi-english tweets using deep learning models. *arXiv preprint*, 2018. arXiv:1811.05145.

14 György Kovács, Pedro Alonso, and Rajkumar Saini. Challenges of hate speech detection in social media. *SN Computer Science*, 2(2), February 2021. doi:10.1007/s42979-021-00457-3.

15 Ping Liu, Wen Li, and Liang Zou. Nuli at semeval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 87–91, 2019.

16 Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013. doi:10.48550/ARXIV.1301.3781.

17 Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. A bert-based transfer learning approach for hate speech detection in online social media. In *International Conference on Complex Networks and Their Applications*, pages 928–940. Springer, 2019.

18 Karsten Müller and Carlo Schwarz. Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, 19(4):2131–2167, 2021.

19 Council of Europe. Portugal should act more resolutely to tackle racism and continue efforts to combat violence against women. https://www.coe.int/en/web/commissioner/-/portugal-should-act-more-resolutely-to-tackle-racism-and-continue-efforts-to-combat-violence-against-women, June 2021.

20 Ryan Ong. Offensive language analysis using deep learning architecture. *arXiv preprint*, 2019. `arXiv:1903.05280`.

21 Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523, 2021.

22 Tharindu Ranasinghe, Marcos Zampieri, and Hansi Hettiarachchi. Brums at hasoc 2019: Deep learning models for multilingual hate speech and offensive language identification. In *FIRE (Working Notes)*, pages 199–207, 2019.

23 Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online), December 2020. International Committee for Computational Linguistics. URL: `https://www.aclweb.org/anthology/2020.semeval-1.271`.

24 Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. Bertimbau: pretrained bert models for brazilian portuguese. In *Brazilian Conference on Intelligent Systems*, pages 403–417. Springer, 2020.

25 Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint*, 2019. `arXiv:1903.08983`.