Semi-Supervised Annotation of Portuguese Hate Speech Across Social Media Domains

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

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

Paula Carvalho ✉
INESC-ID Lisbon, Portugal

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

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

Abstract

With the increasing spread of hate speech (HS) on social media, it becomes urgent to develop models that can help detecting it automatically. Typically, such models require large-scale annotated corpora, which are still scarce in languages such as Portuguese. However, creating manually annotated corpora is a very expensive and time-consuming task. To address this problem, we propose an ensemble of two semi-supervised models that can be used to automatically create a corpus representative of online hate speech in Portuguese. The first model combines Generative Adversarial Networks and a BERT-based model. The second model is based on label propagation, and consists of propagating labels from existing annotated corpora to the unlabeled data, by exploring the notion of similarity. We have explored the annotations of three existing corpora (CO-HATE, ToLR-BR, and HPHS) in order to automatically annotate FIGHT, a corpus composed of geolocated tweets produced in the Portuguese territory. Through the process of selecting the best model and the corresponding setup, we have tested different pre-trained embeddings, performed experiments using different training subsets, labeled by different annotators with different perspectives, and performed several experiments with active learning. Furthermore, this work explores back translation as a mean to automatically generate additional hate speech samples. The best results were achieved by combining all the labeled datasets, obtaining 0.664 F1-score for the *Hate Speech* class in FIGHT.

2012 ACM Subject Classification Social and professional topics → Hate speech; Theory of computation → Semi-supervised learning; Computing methodologies → Transfer learning

Keywords and phrases Hate Speech, Semi-Supervised Learning, Semi-Automatic Annotation

Digital Object Identifier 10.4230/OASIcs.SLATE.2022.11

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. 11; pp. 11:1–11:14
OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

A growing number of people have reported that have already been exposed to hate speech on social media [29]. Due to the anonymity allowed on the Internet, people feel more ease at expressing themselves and engaging in hostile behaviors [12]. Therefore, it urges to develop models able to detect online hate speech automatically.

The non-existence of a unique and consensual definition of hate speech [12] makes its detection more difficult, either for humans or algorithms. For the purpose of this work, hate speech is defined according to the following coexisting conditions [6]:

- Hate speech 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;
- Hate speech 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);
- Hate speech can be expressed both explicitly (or overtly) and implicitly (or covertly).

Most hateful comments are implicit in text, making use of several rhetorical strategies, such as irony and rhetorical questions [33], turning them even harder to identify. Furthermore, hate speech is often context-dependent, meaning that specific words or expressions may have different interpretations, depending on the linguistic and pragmatic context where they are used [16]. Moreover, the personal experiences, knowledge, and beliefs of the ones studying it, as well as demographic features such as the first language, age, education, and social identity, can also introduce personal bias into the classification process [1, 32].

Robust models typically rely on large-scale annotated language resources, which have been created following different annotation guidelines. However, the existing resources – mostly for English – cannot be easily transferred to other languages due the linguistic disparities even within the same language, and the multiplicity of hate speech targets being considered in those studies [27]. Even within the same language, models tend to have generalization problems, dropping in performance when applied to a distinct dataset [34]. Besides, with few exceptions, existing corpora do not usually cover implicit hate speech [3, 15, 17]. In fact, the data comprising Hate Speech (HS) corpora is often retrieved by using negative polarity words and expressions, which are not usually found in implicit hate speech. In addition, most corpora available are imbalanced, and the majority class often correspond to neutral speech, i.e., not offensive nor hateful speech. This asymmetry may deteriorate the performance of the classification models.

Being aware that creating manually annotated corpora is a very time-consuming and expensive task, requiring linguistic and pragmatic knowledge, we propose an ensemble of two semi-supervised models to create annotated corpora representative of the hate speech present on social media platforms in Portugal. The first model combines Generative Adversarial Networks and a BERT-based model. The second one is based on label propagation, assigning labels to the unlabeled data based on their similarity with the annotated corpus. Both models are combined in a semi-supervised self-training approach to obtain an automatically annotated corpus.

The rest of this document is organized as follows. Section 2 presents the related work, focusing particularly on the hate speech datasets available for Portuguese and the most relevant semi-supervised learning models for this task. Section 3 describes the datasets used in the experiments performed, and Section 4 describes our model and the pre-processing applied to the corpora. Section 5 presents the results, and, finally, Section 6 highlights the main conclusions.

2 Related Work

Hate Speech in social media is a recent research topic that has been evolving with the increased use of these platforms. To the best of our knowledge, there are only two datasets covering Portuguese hate speech publicly available.[1]

Leite et al. [18] developed ToLR-BR, a corpus composed of 21,000 tweets, retrieved by applying a list of offensive keywords and considering keywords related to influential Brazilian users that could be targets of hate speech or abuse. The messages were classified as *Homophobia*, *Obscene*, *Insult*, *Racism*, *Misogyny*, and *Xenophobia* by three annotators. Around 44% of the messages were classified as offensive by at least one annotator, 21% by two, and 7% by the three annotators.

Fortuna et al. [13] presented a Hierarchically-Labeled Portuguese Hate Speech Dataset (HPHS) of 5,670 Brazilian Portuguese tweets from 115 users. The messages were retrieved using a list of offensive keywords and by considering users that typically post hateful comments. The tweets were manually classified by three annotators in a binary scheme (hate speech or not). The hatred messages were then classified according to their target, following a hierarchical scheme including 81 hate speech categories. Around 22% of the tweets correspond to hate speech.

Given this lack of resources, semi-supervised learning surges as a solution for hate speech classification. This approach considers a small amount of labeled data and makes use of a large amount of unlabeled data.

Alsafari and Sadaouia [2] use semi-supervised self-training to classify Arabic tweets in *Clean* or *Offensive/Hate*. This approach consists of re-applying the classifier to its most confident predictions [31]. To ensure a good learning ability and good performance, it is required a sufficiently large initial training dataset [2] considering that the performance depends on the accuracy of the pseudo-labels [19]. The tweets are represented with Word2Vec SkipGram embeddings to capture their semantic and syntactic information. The model consists of one classifier based on N-Grams and two deep neural network classifiers. The authors performed multiple experiments with Support Vector Machines (SVM), Convolutional Neural Networks (CNN), AraBERT, and DistilBERT. The classifiers were evaluated according to their accuracy, model size, and inference speed, being the best results achieved by the CNN approach. This model was then used to perform fifteen iterations reusing the predictions with higher confidence. AraBERT and DistilBERT were not used due to their complexity. With the increase in the number of iterations, the model started to associate a hashtag with the tag *Offensive/Hate* so hashtags were ignored. However, the models still perform poorly when classifying implicit hate and in the presence of rare terms. Besides, tweets with counterspeech and abusive words are wrongly classified as *Offensive/Hate*. As expected, the authors also show that increasing the size of the labeled dataset led to a performance increase.

Croce et al. [8] propose GAN-BERT. In Generative Adversarial Learning (GAN), the generator is trained to produce a sample and the discriminator to distinguish between generated samples or samples belonging to the training data. With Semi-Supervised Generative Adversarial Networks (SS-GAN), the discriminator will also classify the sample. BERT is used to encode the input and as the discriminator. The generator is a multi-layer perceptron that transforms an input into a vector representation being the [CLS] token used as a sentence embedding. The discriminator is another multi-layer perceptron with a last layer with

[1] `https://hatespeechdata.com/`

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. The model was tested with a variety of datasets for multiple tasks obtaining an increase in performance for all of them when compared to BERT. Furthermore, the authors have proved that less than 200 annotated examples obtain similar results to the supervised approach. More recently, Breazzano et al. [5] extended this model to multi-task learning and applied it to hate speech classification with similar performance.

D'Sa et al. [10] represent tweets as a pre-trained sentence embedding, using the Universal Sentence Encoder (USE). The authors use a Multilayer Perceptron (MLP) to transform this generic representation into a task-specific representation using a small amount of labeled data. After training with the labeled data, the MLP classifier receives as input the pre-trained representations of a labeled sample and an unlabeled sample. The outputs of the activation function of the two hidden layers correspond to two different task-specific representations. Then, label propagation is performed to obtain the labels for the unlabeled sample. Label propagation is a graph-based semi-supervised technique where the data is represented as a graph. The vertices correspond to the data points and the edges represent the similarity between two nodes. The data points close to each other tend to have a similar label so, the labels are propagated from the labeled points to the unlabeled ones [10, 21]. Finally, the pre-trained embeddings and the labels are used to train the MLP classifier. Comparatively to the MLP classifier trained only with the labeled set and without label propagation, training using label propagation on pre-trained representations performs worse. However, the two representations from the hidden layers capture class information and have better results. In some cases, the label propagation using the representation after the first hidden layer performed better so fully fine-tuned representation may not always be the best approach.

Considering that most of the interactions present in social media do not correspond to hate speech, and given the difficulty to extract them, the percentage of hate speech present in hate speech corpora is low (around 8%). Data augmentation allows to expand an existing training dataset by implementing transformations to the already labeled data or by creating synthetic examples from this data [19, 22]. This can reduce the data scarcity by generating new instances for the minority classes [1], balancing the dataset labels, and reducing the overfit [28]. It can also help the model to better generalize to unseen data, increasing its overall performance [19]. However, data augmentation in NLP tasks is limited since most operations can distort the meaning of the sentence and the number of synonyms of a word is not very high. Considering these limitations, we opted to use back translation since the paraphrases generated by this approach tend to preserve the semantics of the message [4].

This work will follow a self-training approach with an ensemble of two models to reduce the bias of each one. Considering the good results of the previous two models, our proposal will consist of an adaptation of both.

3 Data

We use CO-HATE [6] and FIGHT [7], two corpora recently created from Portuguese online data containing potential hate speech. CO-HATE is composed of comments retrieved from YouTube, and has been manually annotated. FIGHT is composed of tweets, lacks from annotations, and our goal is to provide such annotations. Two additional datasets, focusing on Brazilian Portuguese, were also considered as additional labeled sources, as described in Section 3.3.

3.1 CO-HATE Corpus

The CO-HATE (**C**ounter, **O**ffensive and **Hate** speech) corpus [6] is composed of 20,590 written messages, posted by 8,485 different online users on 39 YouTube videos covering topics and events targeting, directly or indirectly, three specific focus groups: African descent, Roma, and the LGBTQ+ communities. The first two communities correspond to the most representative minorities in Portugal. The LGBTQI community was reported as the most targeted group in terms of online hate speech [11, 23, 25]. The CO-HATE corpus was manually annotated by five annotators, each being responsible for annotating approximately 4,000 messages. Additionally, all annotators were assigned to a common part consisting of 534 messages to assess the inter-annotator agreement (IAA) and the reliability of the annotations.

The annotators 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 three annotators are female. The annotators A, B and C belong to the communities monitored in this study. 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 [6].

The final labels for the messages are obtained considering the majority of the annotations. The IAA (using Krippendorff's alpha) between all the annotators was considerably low (0.478), despite providing the annotators with detailed guidelines. This demonstrates the subjectivity (and difficulty) of this task, even for humans [6]. Table 1 shows the percentage of messages classified as conveying hate speech by each annotator individually, and the group of annotators (ABCDE).

Table 1 Proportion of messages containing hate speech in CO-HATE corpus, by annotator.

Annotators	Number of messages	HS (%)
A	4,008	25
B	4,011	36
C	4,017	29
D	4,014	39
E	4,006	48
Total	20,590	35

3.2 FIGHT Corpus

The FIGHT (**FI**ndin**G** **H**ate Speech in **T**witter) corpus [7] is composed of 56,546 geolocated tweets in the Portuguese territory. This corpus was obtained with two retrieval methods: selecting tweets that include non-ambiguous words that may be used to mention one of the aforementioned target groups (54,352 tweets); and selecting tweets containing a potential mention to the target group, and at least one offensive or insulting word or expression (9,796 tweets). The second approach prevents from retrieving a multiplicity of hate speech forms, including implicit or covert hate speech, but allows retrieving potential offensive or hatred content [7]. In order to evaluate the performance of our models, we have manually annotated a sample of 300 tweets, which is used as our test set.

3.3 Additional Datasets

Since the previously mentioned corpora are focused only on three specific hate speech targets, we decided to consider two additional hate speech Brazilian Portuguese datasets, ToLR-BR and HPHS, covering other HS targets. Taking into account the subjectivity of this task and the personal bias that can be introduced in the annotation process, only the messages labeled as hate speech by the majority of the annotators will be considered as such in order to select only clear cases of hate speech and considering that it is the standard approach in the literature. Regarding ToLR-BR corpus [18], we assumed as hate speech all tweets with one of the labels *Homophobia*, *Racism*, *Misogyny*, and *Xenophobia* given by the majority of the annotators. Of the 21,000 tweets, 403 were classified as hate speech. From these, 192 correspond to *Homophobia*, 96 correspond to *Racism*, 158 to *Misogyny* and 60 to *Xenophobia*. For the HPHS dataset [13], we considered as hate speech the tweets classified as *Hate Speech* by at least two out of the three annotators. From the 5,670 tweets, 1,788 correspond to hate speech.

4 Modeling Approaches

The goal of this work is to present a model capable of automatically classifying hate speech, aiming at contributing to solve the scarcity of annotated hate speech corpora in Portuguese. The model should be able to transfer knowledge from the CO-Hate corpus in order to annotate the FIGHT corpus. This is a particularly complex task considering the different nature of the two corpora. While CO-Hate is composed by YouTube comments contextualized by the videos, FIGHT is composed by individual tweets that are published without a context. Besides, YouTube comments can have an arbitrary size while tweets are limited to 280 characters.

The proposed model corresponds to an ensemble of two semi-supervised models, to reduce the bias of each model. The first model combines Generative Adversarial Networks and a BERT-based model, based on GAN-BERT [8]. The goal is to find the distribution of classes for the labeled data and update it with the unlabeled data. The second model, label propagation, uses the similarities between the instances (points) of the datasets to propagate the existing labels to the unlabeled data. The label of a given point is determined by the labels of the closest points (the implementation used the scikit-learn library [24]). Both models have been recently tested for the hate speech domain, obtaining performance improvements when compared to other previously developed models [5, 10].

Both classifiers are trained with a sample of labeled data. Then, at each iteration, both classifiers classify a subset of unlabeled data. The most confident predictions are added to the labeled set and the models are fine-tuned with them. The maximum sequence length of each message was defined as 350 tokens to ensure the efficiency of the model without loosing too much information. The GAN-BERT model was trained for 15 epochs with 5 patience, considering the model with the best F1-score for the positive class. The training data was randomly split into 80% for the train set and 20% for the development set. The label propagation model used the k-nearest neighbors algorithm with a maximum of 100 iterations and neighbors between 3 and 50.

Considering the GAN-BERT model, we fine-tuned two different pre-trained BERT-based models: **Multilingual BERT**[2] and **BERTimbau** [30] to find which performed better. Similarly, for the label propagation model, we tested representing the sentences with **Doc2Vec** and **Universal Sentence Encoder** (USE).[3]

5 Results

This section describes three different types of experiments. Section 5.1 starts by exploring different embeddings for each model, and assessing the impact of different types of pre-processing. Section 5.2 presents experiments performed by each model individually, considering several subsets of training data. Lastly, Section 5.3 presents the results of the ensemble model, and reveals the impact of using additional labeled datasets and back translation.

5.1 Different Embeddings and Pre-processing Experiments

The experiments here described use CO-HATE as training data, and a sample of 300 tweets from FIGHT corpus that were manually annotated for this purpose, as testing data.

We have started by combining our modeling approaches with different embeddings. The results achieved are summarized in Table 2, where the baseline consists of a dummy classifier that classifies all examples as *Hate Speech*. Results show that BERTimbau achieves an overall better performance when combined with GAN-BERT. This was expected considering that BERTimbau was trained using web corpora that are more likely to include toxicity than the Google Books corpus used for Multilingual BERT. For the label propagation model, we have adopted USE, considering that it has a higher recall and F1-score for the positive class, two relevant metrics when detecting hate speech.

Table 2 Performance of different embeddings for each model.

		Acc	HS Class			Macro Average		
			Prec	Rec	F1	Prec	Rec	F1
Dummy Classifier (all HS)		0.177	0.177	1.000	0.300	0.088	0.500	0.150
GAN-BERT	Multilingual	0.633	0.315	0.133	0.187	0.450	0.430	0.458
	BERTimbau	0.647	0.188	0.302	0.232	0.508	0.511	0.501
Label Propagation	Doc2Vec	0.707	0.260	0.358	0.302	0.555	0.569	0.557
	USE	0.653	0.248	0.472	0.325	0.553	0.582	0.546

In order to understand the impact of pre-processing, for each model, we have also performed experiments without any pre-processing, and with two levels of pre-processing. The partial pre-processing is composed of the following steps:

- Noise removal: remove processing errors in the data retrieval;
- Removal of repetitions of three or more punctuation signals and emojis. This step may remove some noise and shorten the message to fit the maximum sequence length. However, it may lose some of the meaning of the sentence;
- Anonymization of users' mentions: replace a user tag with "@UserID" to represent a username with a single word.

[2] `https://github.com/google-research/bert/blob/master/multilingual.md`
[3] `https://tfhub.dev/google/universal-sentence-encoder-multilingual/3`

The full pre-processing approach is composed of the previous steps plus:

- Removal of user's mentions;
- Removal of links.

As presented in Table 3, the best results for GAN-BERT were obtained with the full pre-processing, contrarily to what we expected. This pre-processing puts the emphasis on the message. However, some of the meaning of the messages can be lost by removing the repetitions of punctuation signals and emojis, and some context can be removed by deleting the user's mentions and links. For the label propagation model, the best results were obtained without any pre-processing, potentially because the pre-processing removes too much context from the messages. The following section use GAN-BERT with the pre-processed training set, and the label propagation model with the original data, which are the most promising combinations. Our goal is to obtain the most promising model, so, for each experiment, we will select the options that result in the best performance.

Table 3 Impact of pre-processing.

	Pre-processing	Acc	HS Class			Macro Average		
			Prec	Rec	F1	Prec	Rec	F1
GAN-BERT	Without	0.647	0.188	0.302	0.232	0.508	0.511	0.501
	Partial	0.680	0.228	0.340	0.273	0.535	0.546	0.534
	Full	0.707	0.294	0.472	0.362	0.580	0.546	0.586
Label Propagation	Without	0.653	0.248	0.472	0.325	0.553	0.582	0.546
	Partial	0.633	0.234	0.472	0.313	0.544	0.570	0.531
	Full	0.623	0.222	0.453	0.298	0.536	0.556	0.520

5.2 Considering Different subsets, from Different Annotators

As previously mentioned, the CO-HATE corpus annotation process involved five annotators, with the achieved low values of IAA evincing the difficulty and subjectiveness of the task. Therefore, in order to assess the perspective of each annotator in the hate speech classification, we have tested several combinations of data subsets. We have used the corpus annotated by each user independently, the corpus composed of messages labeled by all the annotators, and multiple combinations taking into consideration the annotators that have shown best inter-annotator agreement results. Table 4 presents the results for GAN-BERT, using subsets annotated by each annotator and by combined datasets of the most relevant associations. The following experiments were carried out with the two samples that achieved better results, namely the combination of data annotated by annotators B, C, and D, and the data annotated by all the annotators. The results for the label propagation models are shown in Table 5. For this second model, we opted to use the data annotated by annotators A, B, and C, and by all the annotators.

As mentioned by Carvalho et al. [6], annotators A, B, and C belong to the target groups considered in the corpus. Comparing the IAA between this group and the one composed by annotators D and E, who do not belong to any potential marginalized group, we observe that hate speech is perceived differently by individuals from both groups. In fact, the agreement rate was lower among the individuals of the target groups for almost all dimensions considered in the guidelines. Considering the classification of hate speech, the annotators A, B, and C had an IAA of 0.360, while the annotators D and E had an IAA of 0.735. This corroborates the idea that hate speech identification is a very subjective task, and that the annotators'

Table 4 Impact of the perspective of annotators in the performance of GAN-BERT model.

	Acc	HS Class			Macro Average		
		Prec	Rec	F1	Prec	Rec	F1
A	0.667	0.169	0.226	0.194	0.495	0.494	0.492
B	0.530	0.225	0.679	0.338	0.552	0.589	0.487
C	0.477	0.201	0.660	0.308	0.529	0.529	0.444
D	0.687	0.275	0.472	0.347	0.570	0.602	0.571
E	0.463	0.190	0.623	0.291	0.515	0.526	0.430
BD	0.597	0.254	0.660	0.366	0.571	0.622	0.535
DE	0.677	0.266	0.472	0.340	0.565	0.596	0.563
ABC	0.667	0.169	0.226	0.194	0.495	0.494	0.492
BCD	0.700	0.287	0.472	0.357	0.578	0.610	0.581
ABDE	0.620	0.161	0.280	0.282	0.492	0.493	0.484
ABCDE	0.707	0.294	0.472	0.362	0.580	0.610	0.586

Table 5 Performance of the label propagation model based on the perspective of annotators.

	Acc	HS Class			Macro Average		
		Prec	Rec	F1	Prec	Rec	F1
A	0.730	0.259	0.283	0.270	0.551	0.554	0.552
B	0.637	0.225	0.434	0.297	0.537	0.557	0.526
C	0.707	0.267	0.358	0.302	0.555	0.570	0.558
D	0.537	0.179	0.453	0.257	0.502	0.504	0.460
E	0.463	0.186	0.604	0.284	0.511	0.518	0.428
AC	0.697	0.250	0.358	0.295	0.549	0.564	0.551
DE	0.583	0.223	0.547	0.317	0.541	0.569	0.509
ABC	0.693	0.259	0.396	0.313	0.557	0.577	0.558
BCE	0.563	0.213	0.547	0.307	0.533	0.557	0.494
ABCD	0.677	0.250	0.415	0.312	0.552	0.574	0.550
ABCDE	0.643	0.240	0.472	0.318	0.549	0.576	0.538

social identity may influence the perception of HS. Taking this into account, we tried to investigate the impact of each group on the performance of the models and assess whether higher IAA lead to better performance. For GAN-BERT, from Table 4, it is clear that the sample composed by annotators D and E obtained globally better results. For the label propagation model, from Table 5, although using the data from annotators A, B and C led to higher accuracy and precision, the F1-score for the positive class is slightly higher for D and E.

In order to understand the potential of GAN-BERT using the FIGHT corpus, similarly to what was done by Croce et al. [8], each message was labeled as *Unknown* and added to the CO-HATE corpus. The final label given to each point was the one with the highest confidence between *Hate Speech* and *Non Hate Speech*. This increased the accuracy of the model, but reduced the remaining metrics so the idea was discarded.

5.3 Ensemble Model with Additional Labeled Resources

After assessing the potential of both models individually, they were combined in order to produce the labels for the FIGHT corpus. Each individual model used the best training set, i.e., the pre-processed CO-HATE corpus for GAN-BERT and the original one for the label propagation model. Table 6 shows the corresponding results for six different experiments, all of them considering an iterative training over five epochs.

Experiment 1 consisted in adding the most confident predictions (above 0.9) given simultaneously by both models to the training set of the following epoch. Considering that the majority of the labels were *Non Hate Speech* (non-HS), we observed that the training set was getting too unbalanced (only around 20% *Hate Speech*) and the performance of the models was decreasing. To overcome this issue, Experiment 2 consisted of adding only the *Hate Speech* labels, thus obtaining around 42% of hate speech. Experiment 3 consisted of adding all the most confident predictions, including the ones given by only one model. Although the recall increased due to the higher number of hate speech instances, the accuracy and precision decreased, which is possibly explained by the lower confidence associated with these labels.

In order to increase the amount of hate speech present in the training set, we have also added ToLR-BR and HPHS. This solution significantly increased the performance of the model, as reported in Experiments 4 and 5. However, similarly to the previous Experiments 1 and 2, adding only the positive examples turned out to be a better approach (Experiment 5). Additionally, Experiment 6 used back translation to generate more annotated examples from the additional datasets. For this, the hate speech sentences were translated from Portuguese into English and then, back to Portuguese, using the Google translate API – Googletrans.[4] The results reveal a significantly lower performance, possibly due to loss of context during the translation process.

Table 6 Impact of data additions to the training set in label propagation model.

Setup	Acc	HS Class			Macro Average		
		Prec	Rec	F1	Prec	Rec	F1
Baseline	0.650	0.245	0.472	0.323	0.552	0.580	0.543
1) labels in common, HS+non-HS	0.672	0.602	0.583	0.592	0.643	0.667	0.659
2) labels in common, HS	0.693	0.618	0.647	0.632	0.655	0.692	0.684
3) all labels, HS	0.669	0.568	0.789	0.660	0.665	0.669	0.668
4) additional datasets, HS+non-HS	0.713	0.676	0.573	0.620	0.693	0.602	0.695
5) additional datasets, HS	0.708	0.629	0.693	0.659	0.687	0.723	0.702
6) additional datasets, HS, back translation	0.644	0.579	0.472	0.520	0.628	0.618	0.619

Considering all the experiments performed, the final results were obtained using as initial training set the CO-HATE corpus with the data classified by the corresponding annotators and the instances corresponding to hate speech of the two additional datasets. Table 7 reports the results of these experiments after five iterations, revealing that GAN-BERT requires more data in order to obtain better results. Using the data from the annotators B, C, and D, the majority of the metrics decreased when adding the additional dataset, especially precision. This corroborates the theory that GAN-BERT is more susceptible to noise, as seen when assessing the impact of pre-processing. However, with the entire corpus, we can obtain better results than the previous baseline. For the label propagation model,

[4] `https://py-googletrans.readthedocs.io/en/latest/`

using the corpus correspondent to the annotators A, B, and C led to an increase in precision but a decrease in the remaining metrics. Considering the entire dataset, there is a clear increase in terms of recall and F1-score.

Table 7 Models' performance after 5 iterations.

		Acc	HS Class			Macro Average		
			Prec	Rec	F1	Prec	Rec	F1
GAN-BERT	Baseline	0.707	0.294	0.472	0.362	0.580	0.546	0.586
	BCD	0.317	0.194	0.906	0.319	0.549	0.548	0.317
	ABCDE	0.693	0.600	0.743	0.664	0.691	0.683	0.673
Label Propagation	Baseline	0.708	0.629	0.693	0.659	0.687	0.723	0.702
	ABC	0.672	0.657	0.413	0.507	0.667	0.632	0.631
	ABCDE	0.692	0.601	0.743	0.664	0.667	0.748	0.690

5.4 Final Considerations

In an attempt to compare our results with other work reported literature, we considered the work of Breazzano et al. [5] and D'Sa et al. [10] involving Italian and English, respectively, and a similar task, since we did not find any other previous similar work for Portuguese. However, it is important to stress that results can not be directly compared, not only because of the different languages and cultural aspects, but mostly because the testing datasets are different. Breazzano et al. [5] applied GAN-BERT to several Italian hate speech. Both the HaSpeeDe[5] and the DANKMEMES [20] datasets were used in a binary classification task, with the best model achieving a macro average F1-score of 0.633 and 0.584 and an accuracy of 0.693 and 0.562, respectively. D'Sa et al. [10] applied a label propagation model to two English datasets from Founta et al. [14] and Davidson et al. [9] to distinguish hate speech from offensive and normal speech, obtaining a macro average F1-score around 0.670 and 0.710, respectively. Considering that our task corresponds to a cross-domain scenario, we excepted this would negatively impact the results. Additionally, since BERTimbau was trained with Brazilian Portuguese, the GAN-BERT model can have been impacted by vocabulary used only in European Portuguese. Besides that, for the label propagation model, the comparison is done with English datasets, so we expected lower results due to the existence of more morphological variations in Portuguese [26]. However, with GAN-BERT we obtained a macro average F1-score of 0.673, and 0.702 for the label propagation model, which are in line with the above mentioned results, reinforcing the potential of this approach.

6 Conclusions and Future Directions

In the literature, several semi-supervised learning methods have been applied in the field of text classification and adapted to hate speech detection. However, this task is extremely complex and subjective, and its success often depends on the creation of robust and large-coverage language resources, which are still scarce for Portuguese. To address this gap, we have implemented an ensemble of two semi-supervised models. The first one employs a GAN in combination with a BERT-based model. The second model is based on label propagation,

[5] https://github.com/msang/haspeede/

which propagates labels based on similarities. The two models were combined to extract the most confident predictions, which were added to the training data of the next iteration, in an active-learning fashion.

We have explored the annotations of three existing corpora (CO-HATE, ToLR-BR, and HPHS) in order to automatically annotate FIGHT, a corpus composed of geolocated tweets produced in the Portuguese territory. Several pre-processing strategies were tested, demonstrating good results, particularly for the GAN-BERT model. Back translation was also tested in an attempt to generate more hate speech examples, but no performance increase was obtained. The best results were obtained using all the corpora. Specifically, we obtained an F1-score of 0.664 for the *Hate Speech* class for both models. The label propagation approach proved to be more stable and less susceptible to noise, with similar performance to the existing models, besides being a less complex model, and hence faster to train with larger amounts of data. However, both models obtained good performance, especially considering the different nature of the corpora.

In terms of future directions, we plan to manually annotate the entire FIGHT corpus, in an semi-automatic way, and to perform further extensive cross-domain experiments involving CO-HATE and FIGHT, using the proposed models.

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. Association for Computational Linguistics, November 2020. doi:10.18653/v1/2020.alw-1.21.

2 Safa Alsafari and Samira Sadaoui. Semi-Supervised Self-Training of Hate and Offensive Speech from Social Media. *Applied Artificial Intelligence*, pages 1–25, October 2021. doi:10.1080/08839514.2021.1988443.

3 Fabienne Baider and Maria Constantinou. Covert hate speech: A contrastive study of greek and greek cypriot online discussions with an emphasis on irony. *Journal of Language Aggression and Conflict*, 8(2):262–287, 2020.

4 Djamila Romaissa Beddiar, Md Saroar Jahan, and Mourad Oussalah. Data expansion using back translation and paraphrasing for hate speech detection. *Online Social Networks and Media*, 24:100153, 2021. doi:10.1016/j.osnem.2021.100153.

5 Claudia Breazzano, Danilo Croce, and Roberto Basili. MT-GAN-BERT : Multi-Task and Generative Adversarial Learning for sustainable Language Processing. In *Proceedings of the Fifth Workshop on Natural Language for Artificial Intelligence (NL4AI 2021)*. CEUR Workshop Proceedings, November 2021. URL: http://ceur-ws.org/Vol-3015/.

6 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. *Discourse and Society*, submitted.

7 Paula Carvalho, Bernardo Matos, Raquel Santos, Fernando Batista, and Ricardo Ribeiro. Hate Speech Dynamics Against African descent, Roma and LGBTQI Communities in Portugal. *LREC*, 2022.

8 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.

9 Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*, pages 512–515, 2017. arXiv:1703.04009.

10 Ashwin Geet D'Sa, Irina Illina, Dominique Fohr, Dietrich Klakow, and Dana Ruiter. Label Propagation-Based Semi-Supervised Learning for Hate Speech Classification. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 54–59. Association for Computational Linguistics, November 2020. doi:10.18653/v1/2020.insights-1.8.

11 Tim Fitzsimons. Nearly 1 in 5 hate crimes motivated by anti-LGBTQ bias, FBI finds. *NBC News*, November 2019. URL: https://www.nbcnews.com/feature/nbc-out/nearly-1-5-hate-crimes-motivated-anti-lgbtq-bias-fbi-n1080891.

12 Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4):1–30, July 2019. doi:10.1145/3232676.

13 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.

14 Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

15 Akshita Jha and Radhika Mamidi. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the second workshop on NLP and computational social science*, pages 7–16, 2017.

16 György Kovács, Pedro Alonso, and Rajkumar Saini. Challenges of hate speech detection in social media. *SN Computer Science*, 2(2):1–15, 2021.

17 Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. Benchmarking aggression identification in social media. In *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*, pages 1–11, 2018.

18 Joao A Leite, Diego F Silva, Kalina Bontcheva, and Carolina Scarton. Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis. *arXiv preprint*, October 2020. arXiv:2010.04543.

19 Changchun Li, Ximing Li, and Jihong Ouyang. Semi-Supervised Text Classification with Balanced Deep Representation Distributions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5044–5053. Association for Computational Linguistics, August 2021. doi:10.18653/v1/2021.acl-long.391.

20 Martina Miliani, Giulia Giorgi, Ilir Rama, Guido Anselmi, and Gianluca E Lebani. DANKMEMES@ EVALITA 2020: The Memeing of Life: Memes, Multimodality and Politics. In *EVALITA*, 2020.

21 Yassine Ouali, Céline Hudelot, and Myriam Tami. An Overview of Deep Semi-Supervised Learning. *arXiv:2006.05278*, pages 1–43, June 2020. arXiv:2006.05278.

22 Maria Papadaki. Data Augmentation Techniques for Legal Text Analytics. Master's thesis, Athens University of Economics and Business, October 2017. URL: http://nlp.cs.aueb.gr/theses.html.

23 Haeyoun Park and Iaryna Lyshyn. L.G.B.T. people are more likely to be targets of hate crimes than any other minority group. *The New York Times*, June 2016. URL: https://www.nytimes.com/interactive/2016/06/16/us/hate-crimes-against-lgbt.html.

24 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. URL: https://scikit-learn.org/.

25 Wyatt Ronan. New FBI Hate Crimes Report Shows Increases in Anti-LGBTQ Attacks. *Human Rights Campaign*, November 2020. URL: https://www.hrc.org/press-releases/new-fbi-hate-crimes-report-shows-increases-in-anti-lgbtq-attacks.

26 Diana Santos and Alberto Simões. Portuguese-English word alignment: some experiments. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA). URL: `http://www.lrec-conf.org/proceedings/lrec2008/pdf/760_paper.pdf`.

27 Sheikh Muhammad Sarwar and Vanessa Murdock. Unsupervised Domain Adaptation for Hate Speech Detection Using a Data Augmentation Approach. *arXiv:2107.12866*, July 2021. `arXiv:2107.12866`.

28 Connor Shorten and Taghi M. Khoshgoftaar. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1), July 2019. `doi:10.1186/s40537-019-0197-0`.

29 Alexandra A. Siegel. Online hate speech. In Joshua A. Tucker Nathaniel Persily, editor, *Social Media and Democracy*, chapter 4, page 67. Cambridge University Press, August 2021. `doi:10.1017/9781108890960`.

30 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.

31 Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine Learning*, 109:373–440, November 2020. `doi:10.1007/s10994-019-05855-6`.

32 Zeerak Waseem. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142. Association for Computational Linguistics, November 2016. `doi:10.18653/v1/w16-5618`.

33 Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. Implicitly Abusive Language – What does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587. Association for Computational Linguistics, June 2021. `doi:10.18653/v1/2021.naacl-main.48`.

34 Wenjie Yin and Arkaitz Zubiaga. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598, 2021.