TÉCNICO LISBOA

Semi-Automatic Selection and Annotation of Hate Speech from Social Media

Raquel Bento Santos

Thesis to obtain the Master of Science Degree in

Information Systems and Computer Engineering

Supervisors: Prof. Fernando Manuel Marques Batista
Prof. Paula Cristina Quaresma da Fonseca Carvalho

Examination Committee

Chairperson: Prof. José Alberto Rodrigues Pereira Sardinha
Supervisor: Prof. Fernando Manuel Marques Batista
Member of the Committee: Prof. Nuno João Neves Mamede

October 2022

Acknowledgments

Em primeiro lugar, quero agradecer à minha família e amigos por toda a paciência, o apoio e a motivação. Ao João Ricardo e a toda a gente que conheci graças a ele e que foram a minha companhia ao longo deste último ano. Quero agradecer também aos mecenas responsáveis pela Bolsa de Estudo E.A.S.S. que me apoiou nos primeiros quatro anos do meu percurso académico e que foram essenciais para chegar até aqui. Agradecer também à Fundação para a Ciência e Tecnologia que financiou o projeto Hate COVID-19.pt e a toda a equipa que contribuiu para o desenvolvimento deste trabalho. Em particular quero agradecer aos meus orientadores, Fernando Batista e Paula Carvalho, e ao Ricardo Ribeiro pelo apoio ao longo deste processo e pelo seu contributo para a elaboração desta dissertação.

A todos, o meu obrigada.

Abstract

With the proliferation of hate speech, particularly on social media, it urges to develop models able to automatically detect it. Such models typically rely on large-scale annotated data, which are still scarce in languages such as Portuguese. However, creating manually annotated corpora is a very time-consuming, expensive, and demanding task. To address this problem, we tested an ensemble of three semi-supervised models that can be used to automatically create a corpus representative of online hate speech in Portuguese. These models consist of a Convolutional Neural Networks (CNN); a model that combines Generative Adversarial Learning (GAN) and a Bidirectional Encoder Representations from Transformers (BERT) based model; and a label propagation model. Furthermore, this work explores the impact of data augmentation and domain adaptation to solve the unbalanced data and the linguistic heterogeneity, taking into consideration the geographic context, and the targets of hate speech. We have explored the annotations of three existing Portuguese corpora (CO-Hate, ToLR-BR, and HPHS) to automatically annotate FIGHT, a corpus composed of geolocated tweets produced in the Portuguese territory. Additionally, to augment our training dataset, HS English corpora were automatically translated into Portuguese. An intermediary domain between CO-Hate and FIGHT was also generated to diminish the differences in the nature of both data sources. The models obtained a performance in line with the results reported in the literature for the same domain task. Additional experiments, from FIGHT to CO-Hate, and within the same domain were also performed to analyze the potential of the proposed models.

Keywords

Hate Speech; Semi-Supervised Learning; Semi-Automatic Annotation; Self-training; Data Augmentation.

Resumo

Com a proliferação do discurso de ódio, principalmente nas redes sociais, torna-se fundamental desenvolver modelos capazes de detetá-lo automaticamente. A criação de modelos robustos normalmente requer uma grande quantidade de dados, que ainda são escassos em línguas como o português. No entanto, a criação de conjuntos de dados manualmente anotados é uma tarefa morosa e dispendiosa. Para resolver este problema, foi testado um conjunto de três modelos semi-supervisionados, que foram usados para criar automaticamente um conjunto de dados anotados, representativos do discurso de ódio online em português. Estes modelos consistem numa Rede Neural Convolucional; num modelo que combina Redes Adversariais Generativas e BERT; e num modelo de propagação de etiquetas. Além disso, este trabalho explora o impacto do aumento de dados e adaptação de domínio para resolver o desequilíbrio dos dados e a heterogeneidade linguística, considerando o contexto geográfico e os alvos do discurso de ódio. Foram exploradas as anotações de três corpora existentes para o português (CO-HATE, ToLR-BR e HPHS) para anotar automaticamente o FIGHT, uma coleção de dados composta por tweets geolocalizados no território português. Além disso, sete conjuntos de dados em inglês foram automaticamente traduzidos para o português, para aumentar os dados de treino. Um domínio intermédio entre o CO-Hate e o FIGHT também foi gerado. Os modelos obtiveram um desempenho semelhante ao reportado na literatura para a tarefa entre o mesmo domínio. Experiências adicionais partindo, neste caso, do FIGHT para o CO-Hate, e considerando o mesmo domínio, foram igualmente levadas a cabo para analisar o potencial do modelo.

Palavras Chave

Discurso de Ódio; Aprendizagem Semi-Automática; Anotação Semi-Automática; Autotreinamento; Aumento de Dados.

Contents

List of Figures

List of Tables

Acronyms

AGNN	Attention-based Graph Neural Networks
BERT	Bidirectional Encoder Representations from Transformers
BrWaC	Brazilian Web as Corpus
CBOW	Continuous Bag-of-Words
CC	Context Carrier
CO-Hate	**C**ounter, **O**ffensive and **Hate** speech
CLN	Convolutional Ladder Network
CNN	Convolutional Neural Networks
CVAE	Conditional Variational Autoencoders
ELMo	Embeddings from Language Models
FIGHT	**FI**ndin**G** **H**ate Speech in **T**witter
FlauBERT	French Language Understanding via BERT
FLUE	French Language Understanding Evaluation
GAN	Generative Adversarial Learning
GCN	Graph Convolutional Networks
GDPR	General Data Protection Regulation
GloVe	Global Vectors
GNN	Graph Neural Networks
GPT-3	Generative Pre-Trained Transformer 3
HS	Hate Speech
IAA	Inter-Annotator Agreement
LR	Logistic Regression

LSA	Latent Semantic Analysis
LSTM	Long Short-Term Memory
MLP	Multilayer Perceptron
NLP	Natural Language Processing
OTG	Offensive expression or Target Group mention
PCA	Principal Components Analysis
POS	Part-of-Speech
PMI	Pointwise Mutual Information
PPDB	Paraphrase Database
RF	Random Forest
RNN	Recurrent Neural Network
RoBERTa	Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach
SVM	Support Vector Machines
S3VM	Semi-Supervised Support Vector Machines
S4VM	Safe Semi-Supervised Support Vector Machines
SS-GAN	Semi-Supervised Generative Adversarial Networks
TF-IDF	Term Frequency–Inverse Document Frequency
TSVM	Transductive Support Vector Machines
USE	Universal Sentence Encoder
VAE	Variational Autoencoders
VAT	Virtual Adversarial Training

1

Introduction

Contents

With the rise of the use of the Internet and social media, it became easier to express opinions and participate in online debates. However, this also potentiated polarized discussions, which often introduce harsh language and hate speech to social media.

This work is developed in the scope of Hate COVID-19.pt[1], a scientific research project focused on the analysis of hate speech in Portuguese and its automated detection. The emphasis of this project is to identify the main strategies used in explicit (or overt) and implicit (or covert) Hate Speech (HS) and to develop models that automatically identify both types of speech. Moreover, the project aims at understanding the impact of COVID-19 on hate speech, particularly in the Portuguese online context.

The non-existence of a unique and consensual definition of hate speech both in literature [3] and in the policies of social media corporations, makes its detection more difficult, both for humans and algorithms. For the purpose of this work, our definition is based on the one recently proposed by the Council of Europe in its Recommendation CM/Rec/2022/16 where hate speech is defined as "all types of expression that incite, promote, spread or justify violence, hatred or discrimination against a person or group of persons, or that denigrates them, by reason of their real or attributed personal characteristics or status such as race, colour, language, religion, nationality, national or ethnic origin, age, disability, sex, gender identity and sexual orientation".[2] Considering this, hate speech is defined according to the following coexisting conditions [4]:

- 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).

A growing number of people have reported that have been exposed to hate speech on social media [5]. This phenomenon can result in a negative self-image and the marginalization of the targeted community [6]. Due to the anonymity allowed on the Internet, people feel more at ease expressing themselves and engaging in hostile behaviors [3]. This was especially aggravated with the COVID-19 lockdown [7] that forced people to use social platforms as a medium of communication and promoted discrimination against specifically targeted communities such as Chinese people [8, 9]. To solve this problem, several countries have developed legislation to hold platforms responsible for the hate speech that is published on them. The social media platforms themselves are more aware of the dangers of

[1] https://hate-covid.inesc-id.pt/
[2] https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=0900001680a67955

this type of discourse and have implemented codes of conduct and mechanisms to remove any harmful publications [10–13]. These policies are intended to remove hate speech from social media platforms and may impose yet another limitation considering that these resources are important to study the phenomenon of hate speech dissemination. However, there is a significant number of cases expressing hate speech that is still available on a variety of social media platforms. This problem is even more evident in non-English languages, like Portuguese, where there is both a lack of human moderation and models to detect hate speech [14–16].

To automatically detect hate speech, a robust supervised learning model requires a large, annotated dataset that covers not only the vocabulary but also the syntactic constructions and the rhetorical devices used to express both direct and indirect hate speech. This is especially difficult to obtain due to the diversity of languages, the regional and linguistic disparities even within the same language, and the variety of hate speech targets, who may have their own specificities [17]. The existing resources – mostly for English – cannot be easily transferable to other languages due to cultural disparities. Even within the same language, models tend to have generalization problems, dropping in performance when tested with a distinct dataset [18].

In the existing resources, several definitions are employed, mixing up especially the notions of hate and offensive speech. This results in heterogeneous resources, reducing the number of corpora that can be used in this task since they do not meet our definition of hate speech. Besides, with few exceptions, existing corpora do not usually cover implicit hate speech [19–21]. In fact, the data comprising hate speech corpora are often retrieved by using words and expressions with negative polarity, which might not be found in implicit hate speech, often based on figurative language, including irony, comparisons, rhetorical questions, and other strategies [22,23]. Besides, when extracting data from social media, hate speech corresponds to a small portion of conversations. Implicit hate speech is even more difficult to identify, considering that it usually does not contain offensive terms and requires additional context to be interpreted. This leads to unbalanced datasets, where the majority class corresponds to *Non-Relevant* examples, i.e., not offensive nor hateful speech nor counterspeech. This asymmetry deteriorates the performance of the classification models, so using a larger and more balanced dataset may prevent overfitting [24]. Furthermore, distinguishing the minority class from the remaining is impractical and manual labeling all the data requires a great human effort, being time-consuming and expensive. In addition, 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 [25]. This requires annotators that understand the language, the culture, and the context of the message.

Crowdsourcing annotation is a faster alternative, but also requires a lot of preparation and confirmation of the results to ensure the quality of the annotations. Since the annotation is done by several unqualified users, motivated in its majority by economic incentives, a high error rate can be introduced.

A common baseline considers only the majority vote, but this approach ignores important information considering that some annotators are more reliable than others [26]. As an alternative, for the annotations to be considered, the annotator should have a minimum reliability measured by the number of failed questions. A question is failed if there is no exact match of the annotation with the gold standard [27].

Moreover, the annotation process can introduce bias to the model. This is partially due to the fact that hate speech can be confused with abusive or offensive language or even counterspeech, particularly in cases where the guidelines are not clear enough. Besides this, the personal experiences, knowledge and beliefs of the annotators, as well as demographic features such as the first language, age, and education can also introduce personal bias into the dataset [28, 29].

To deal with shortage and unbalance of data, data augmentation can be used to artificially generate synthetic hateful comments [30, 31]. Moreover, semi-supervised learning requires a small set of annotated data and takes advantage of a large set of unlabeled data to improve performance [32]. This project will allow an understanding of the impact of data augmentation and the efficacy of semi-supervised methods to classify hate speech in Portuguese. Furthermore, domain adaptation will be used to cover the variability of hate speech between the different social media platforms.

This document contains sensitive content and real hate speech examples retrieved from social media in order to illustrate the problem and the behavior of the models.

1.1 Goals

Being aware that creating manually annotated corpora is a very time-consuming and expensive task, requiring a deep understanding of the language system and social practice, we propose an ensemble of three semi-supervised models to create annotated corpora representative of the hate speech present on social media platforms in Portugal. The first model corresponds to a Convolutional Neural Networks (CNN) that is used as a weaker and faster baseline and as a tiebreaker between the other two models' predictions. The second model combines Generative Adversarial Learning (GAN) and a Bidirectional Encoder Representations from Transformers (BERT) based model to add noise to the training data in order to make the model more robust. The last one is based on label propagation, assigning labels to the unlabeled data based on their similarity with the annotated corpus. The models are combined in a semi-supervised self-training approach to obtain an automatically annotated corpus.

In the end, this work proposes to fulfill the following research goals:

- Develop and implement a set of semi-supervised models to automatically annotate hate speech in social media text;
- Understand the impact of data augmentation and domain adaptation in semi-supervised learning with limited and unbalanced data;

- Determine which is the best approach to generate labels for a specific hate speech corpus;
- Apply relevant evaluation metrics that allow comparing and assessing the implemented approaches and models;
- Understand the impact of cross-domain tasks on the models' performance.

1.2 Methodology

This research makes part of Hate COVID-19.pt and aims at developing a model able to recognize hate speech. For that, a large amount of annotated data is needed. Considering that there are only two datasets publicly available for Portuguese, which cover the Brazilian scenario, we use data augmentation to generate more data. In particular, back-translation is used from Portuguese into English and back into Portuguese, in the two available Brazilian datasets. Besides, a sample of English hate speech messages is translated into Portuguese and added to the training data.

After obtaining these additional resources, the models are selected. We have initially tested several weaker models but, due to the complexity of this task, their performance was considerably low. CNN was the model with the best results, when combined with a Word2Vec SkipGram, in order to consider the context of all the sentences to capture their full meaning. Besides, GAN-BERT and the label propagation model are also used, considering that they have proven to obtain good results for hate speech classification. In the case of GAN-BERT, a re-trained BERT model for abusive language was fine-tuned with CO-HATE to be able to recognize hate speech. For the label propagation model, a Universal Sentence Encoder (USE) is used, in order to consider the context of the sentence.

Considering the difficulties of the model to identify the shorter hate speech messages from FIGHT, several hate speech tweets were added, generated from the data augmentation steps. The same was done with implicit hate speech to try to solve the difficulties of the models. However, the annotators themselves struggle to agree on the meaning of some messages, which often require access to their context (that do not exist in some cases).

1.3 Organization of the Document

The rest of this document is organized as follows. In Chapter 2 it is presented the related work on hate speech datasets, semi-supervised learning for text classification, data augmentation in text, and transfer learning. Chapter 3 presents the procedures underlying data extraction, and the transformations applied to the existent data. Chapter 4 presents the proposed architecture to solve the research problem. The

experiments conducted and the results achieved are presented in Chapter 5. Lastly, Chapter 6 presents the main conclusions and future work.

2

Related Work

Contents

Hate Speech in social media is a recent research topic that has been evolving with the increased use of these platforms. This section will start by presenting the datasets specifically created to detect Portuguese hate speech. Considering the low amount of Portuguese resources, English corpora that have adopted a similar hate speech definition will be also presented.

Considering the novelty of this topic, this chapter will provide an overview of the most relevant semi-supervised approaches focusing on text classification problems in general.

In order to solve the problem of low data resources, we will lastly present some of the most used strategies for data augmentation and transfer learning.

2.1 Hate Speech Datasets

As previously mentioned, the focus of this research is Portuguese hate speech. To the best of our knowledge, there are only four datasets covering Portuguese, all of them focused on Brazilian Portuguese.[1] From these, only two are publicly available - HPHS and ToLR-BR.

Pelle and Moreira [33] develop a corpus with comments from the most used Brazilian news website. The focus was politics and sports news considering that these topics had a higher percentage of hate speech. The corpus is composed of 10,336 comments retrieved from 115 news. Each comment is classified as "Offensive" or not. Then, the offensive comments are categorized into "Xenophobia", "Homophobia", "Sexism", "Racism", "Cursing" and "Religious Intolerance". "Cursing" is present in about 25% of the comments and "Religious Intolerance" was found in only one comment. The remaining categories correspond to 1 to 2% of the comments. Around 20% of the comments were classified as offensive. However, considering our definition, only a small percentage corresponds to hate speech. This low amount may be due to the moderation present on the news website.

Fortuna et al. [6] present a corpus of 5,670 Brazilian Portuguese tweets (HPHS) from 115 users. The tweets were retrieved by using offensive keywords and by selecting users who often post hateful comments. These messages were manually classified by three annotators in a binary scheme (hate speech or not). The hate messages were then classified according to their target, following a hierarchical scheme including 81 hate speech categories, to understand their motivation. As an example, racism is divided into "Black people", "Chinese", "Latinos", etc. Around 22% of the tweets correspond to hate speech. Most targets only have one corresponding instance, so it is hard for the models to predict these categories.

Leite et al. [34] develop 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 are classified as "Homophobia", "Obscene", "Insult", "Racism",

[1] https://hatespeechdata.com

"Misogyny", "Xenophobia" or "None" if the tweet was not offensive. Around 44% of the messages were classified as offensive by at least one annotator, 21% by two, and 7% by the three annotators. About 3% of tweets are classified as "Obscene", another 3% as "Insult", and the remaining classes have less than 0.4%.

Lastly, *Vargas et al.* [35] present a corpus of 7,000 comments extracted from Instagram posts of six Brazilian political personalities. This corpus was annotated according to three criteria: "offensive" or "non-offensive"; the level of offense in three levels; and regarding the target in "Xenophobia", "Racism", "Homophobia", "Sexism", "Religious intolerance", "Partyism", "Apology to the dictatorship", "Antisemitism" and "Fatphobia". Half of the comments are offensive, being 11% highly offensive, 15% moderately offensive, and 24% slightly offensive. Of the offensive comments, 14% corresponds to "Partyism" and the remaining have less than 3% each.

For English, there are several resources. However, due to the plurality of hate speech definitions, the majority do not clearly distinguish between hate and offensive speech. Taking this into consideration, we have found seven datasets publicly available that are conceptually closer to our definition.

Kennedy et al. [36] present a dataset composed of 50,000 comments, annotated by 10,000 annotators. These comments were retrieved from Twitter, Reddit, and YouTube. The tweets were randomly selected, the Reddit comments correspond to all comments from the subreddit "/r/all" and the YouTube comments were retrieved from videos published in the top 300 most populated cities of the United States of America. The messages are classified as "Sentiment", "Respect", "Attack-Defend", "Insult", "Status", "Dehumanize", "Humiliate", "Hate Speech", "Violence" and "Genocide". Around 40% of the messages were classified as hate speech.

Samory et al. [37] present a dataset composed of 13,631 messages, recurring to crowdsourcing annotation to classify them as "Sexism" or not. These messages correspond to tweets, psychological survey items, and adversarial examples generated by machine learning models using as input the retrieved messages. Around 13% of the messages were classified as sexism.

ElSherief et al. [38] present a corpus composed of 3,222 tweets. The tweets were retrieved using hate-related keywords or expressions and annotated by crowdsourcing in a binary classification scheme (hateful or neutral). Around 73% of the tweets were classified as hateful.

Mollas et al. [39] develop a corpus of 998 messages retrieved from YouTube and Reddit. The YouTube comments were extracted from the ones automatically annotated by Hatebusters Platform. The messages from Reddit were retrieved from 4 subreddits that already have been shut down due to the dissemination of hate speech. The messages were then filtered to ensure a balance and diversity of the labels ("violence", "directed vs generalized", "gender", "race", "national origin", "disability", "sexual orientation" and "religion"). The messages were annotated by crowdsourcing in a binary classification scheme (hate speech or not) and according to the previously mentioned categories. Around 43% of the

messages were classified as hateful.

Davidson et al. [40] present a dataset of 24,802 tweets, retrieved using offensive keywords and expressions. The tweets were classified by crowdsourcing as “Hate Speech”, “Offensive Language” or “Neither”. Around 6% of the messages were classified as hate speech.

Gao and Huang [41] develop a corpus of 1,528 comments, retrieved from ten of the most read Fox News articles. The comments were annotated by two annotators in a binary classification scheme (hate speech or not). Around 28% of the messages were classified as hateful.

Röttger et al. [42] develop a corpus of 3,728 messages synthetically generated to cover 29 functional tests in 11 classes: “Derogation” (implicit and explicit), “Threatening language”, “Slur usage” (hateful and not hateful), “Profanity usage” (hateful and not hateful), “Pronoun reference”, “Negation” (hateful and not hateful), “Phrasing” (question and opinion), “Non-hate group identity”, “Counterspeech”, “Abuse against non-protected targets” and “Spelling variations” of hateful messaged. The messages were annotated by 10 annotators in a binary classification scheme (hate speech or not). Around 68% of the messages were classified as hate speech.

These existing resources will be used in combination with the extracted data in order to improve the results of our models.

2.2 Semi-Supervised Learning

Given the lack of resources, semi-supervised learning appears as a solution for hate speech classification. Supervised Learning trains classifiers using labeled data. Unsupervised Learning models attempt to learn patterns from unlabeled data in order to classify this data. Semi-Supervised Learning considers a small amount of labeled data and a large amount of unlabeled data. The categorization used in this section is represented in Figure 2.1.

Inductive methods are seen as extensions of the supervised algorithm to classify unlabeled data. These methods employ pseudo-labeled data to train a classifier that will predict the labels for the remaining unlabeled instances [43]. This can be done during pre-processing, inside the objective function, or during a pseudo-labeling step [1]. The trained classifier is then used in the testing phase to predict the label for unlabeled or unseen data points [1]. **Transductive methods** classify unlabeled data by propagating information from labeled data to unlabeled data. There is no distinction between the training and testing phases considering that the algorithms receive both labeled and unlabeled data and generate predictions for the unlabeled part [1]. These algorithms typically consist of a graph-based approach to encode data similarity so that the unlabeled points inherit the label from the most similar labeled points [1].

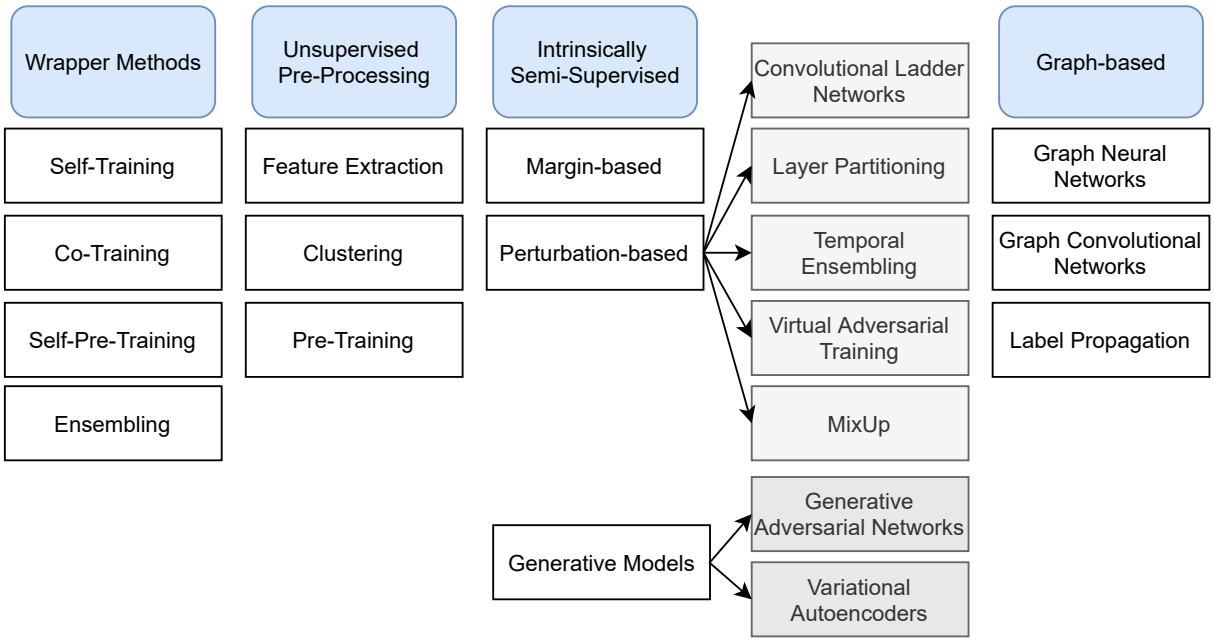

Figure 2.1: Categorization of Semi-Supervised Methods. Adapted from [1].

2.2.1 Wrapper Methods

Wrapper methods train one or more classifiers on labeled data generating pseudo-labels that will be used to predict a new set of data [1]. The final classifier will apply both the original labels and the pseudo-labels without any distinction [1, 44]. A schematic representation of the behavior of these methods is depicted in Figure 2.2.

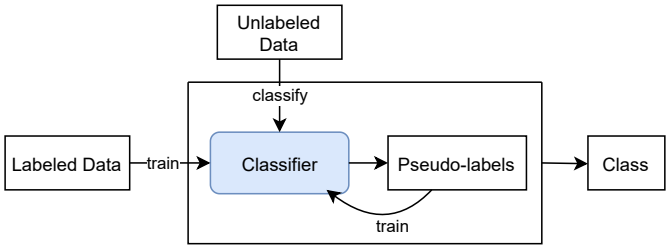

Figure 2.2: Behavior of Wrapper Methods.

2.2.1.A Self-Training

Self-training re-applies the classifier to its most confident predictions [1]. To ensure a good learning ability and good performance, it is required a sufficiently large initial training dataset [43] considering that the performance depends on the accuracy of the pseudo-labels [45]. As an advantage, this can be applied to multiple tasks considering that they do not require any assumptions [46].

Alsafari and Sadaouia [43] use semi-supervised self-training to classify Arabic tweets in "clean" or "offensive/hate" speech. 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), CNN, AraBERT, and DistilBERT. The classifiers were evaluated according to their accuracy, model size, and inference speed, being the best results for CNN. 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.

Xu et al. [46] propose a semantic space-based self-training model for multi-label text classification that combines self-training in pre-training and fine-tuning. The semantic space of the pre-trained model (SBERT) is initially fine-tuned with the labeled data. Then, a fully connected layer is added to the pre-trained model to fine-tune the classifier on the previous dataset. The most confident predictions are extracted to fine-tune the semantic space of the classifier. To avoid learning too much noise, self-training stops if the set of confident labels stops growing, if the set does not include most samples of the previous iteration or if the overlapped samples have inconsistent labels.

2.2.1.B Co-Training

Co-training applies the self-training procedure with multiple classifiers, each using different data subsets – views [1]. At each iteration, the generated pseudo-labels are added to the labeled data of the remaining classifiers, reducing the disagreement and minimizing the error rate [1]. In multi-view co-training, the classifiers are trained in distinct subsets of features that must be sufficient to obtain good results and should be conditionally independent [46] or, at least, not highly correlated [1]. Single-view co-training solves the problem of disjoint subsets of features by splitting the feature set in each iteration [1].

Rosenthal et al. [47] use a semantically oriented model to annotate a dataset of English tweets according to their offensiveness and target. The Pointwise Mutual Information (PMI) is calculated for each unigram and bigram and combined into a single score to determine the best class. FastText is used to incorporate subword representations to reduce the noise. Long Short-Term Memory (LSTM) is used to account for long-distance relations between words. BERT is used due to its high representational power and robustness. Each model is trained on a gold dataset and predicts the confidence of each unlabeled example. An aggregated score considering the average and standard deviation of each model's confidence is used to avoid over-fitting to any model, to reduce dataset biases to a specific model, and to

eliminate instances where the models disagree. After training with the pseudo-labels, the classification process is tested with BERT and FastText, with BERT producing better results.

Shayegh et al. [32] propose a co-training semi-supervised learning algorithm. Using a CNN, augmented data is added to the input by replacing random words with their synonyms. The training set is divided using Latent Dirichlet Allocation, in which topics are created considering the word distribution and each document is assigned to the most probable topic. Each view is composed of all documents with the same topic. The input goes through a CNN to extract the most relevant features and their classes. This generates a different classifier with different features for each view. Each unlabeled example is classified by the classifiers which views are in the k-nearest neighbors. The confident predictions are added to the labeled dataset until a stopping criterion is met.

Chen et al. [48] propose a co-training semi-supervised deep learning model for sentiment classification of posts on online courses' forums. The authors consider character-based Embeddings from Language Models (ELMo) and word embeddings as the two different views. The embeddings for a small subset of labeled data are used to train both models and then the pseudo-labels are generated using a CNN. The selected confident predictions have the same label in both classifiers and have high similarity with samples from the training set with that class.

2.2.1.C Self-Pre-Training

Self-pre-training is an iterative method with two classifiers where one labels a sample of unlabeled data and the second one is initialized with these labels and trained with a set of labeled data [49].

Karisani P. and Karisani N. [49] propose a model inspired in self-training but resilient to the pseudo-labels noise that increases with the number of iterations. In each iteration, self-pre-training revises the previous labels. These revised labels are then used to initialize the classifier that will be fine-tuned with the labeled data. This approach was tested on Twitter datasets and outperformed BERT-based models and a self-training approach with hundreds of labeled documents.

2.2.1.D Ensemble

Ensemble methods apply several classifiers sequentially, incorporating labeled data and the previous most confident predictions [1]. The goal is to combine multiple models with different inductive biases to reduce each one's biases [47]. The ensemble can alternate between classifying the unlabeled data and using the pseudo-labels to train the next model [50]. Alternatively, each classifier can be trained with labeled data and, in each iteration, classify a different subset of unlabeled data considering the results of the previous iteration and being weighted according to their confidence [51]. In bagging methods, each model is trained independently with a random sample, and then, the predictions result of the combination of their outputs [1]. In boosting methods, each model receives the full dataset weighted according to

the previous performance being the larger weights assigned to the wrongly classified points [1]. The predictions result from a linear combination of the predictions of all classifiers [1].

Alsafari and Sadaoui [43] show that using an ensemble-based self-training with only two iterations obtains similar results to the previous CNN approach with fifteen iterations. This was verified using a CNN and a bidirectional LSTM with Maximum Voting (select data with the highest probability) and Average Voting (select data with the average of both classifiers above threshold).

2.2.2 Unsupervised Pre-processing

Unsupervised pre-processing uses unlabeled and labeled data in two different stages [1]. Firstly, it is applied one unsupervised method, followed by a supervised classification model. A schematic representation of this behavior is depicted in Figure 2.3.

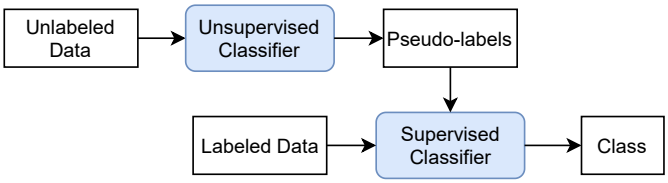

Figure 2.3: Behavior of Unsupervised Pre-processing Methods.

2.2.2.A Feature Extraction

Feature extraction methods extract or transform useful features from the unlabeled data to improve the performance of the classifier [1].

Zareapoor and Seeja [52] use Principal Components Analysis (PCA) and Latent Semantic Analysis (LSA) to extract features from emails to obtain a more compact feature space. PCA transforms the original space into a linear, uncorrelated, and smaller one by combining the original variables. These new variables are obtained by computing the mean and the covariance of the original attributes and by extracting the eigenvectors. The eigenvectors with the highest eigenvalues are the new features. LSA correlates semantically related terms to produce a smaller set of concepts to deal with synonyms and homonyms. The selected features are applied to a bagging classifier.

Lee et al. [53] use an unsupervised CNN to learn the feature's embedding and then apply the feature vectors to a supervised CNN to classify adverse drug events in Twitter. Each tweet is treated as a bag of words. The unsupervised CNN learns a vector representation for each tweet by considering its context. Then, the supervised CNN is trained with the annotated data and the previously generated embeddings.

2.2.2.B Clustering

Clustering follows the principle of applying an unsupervised or semi-supervised clustering algorithm to all data and use the results in classification [1].

As an example, *Zhang et al.* [54] propose a semi-supervised clustering approach to classify English and Chinese articles. The authors use labeled data to determine the text clusters, and the unlabeled data to adjust the centroids. The unlabeled data is classified according to the similarity with the clusters.

2.2.2.C Pre-Training

Pre-training uses unlabeled data to guide the decision boundary before fine-tuning with the labeled data.

Sun et al. [55] use BERT, Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach (RoBERTa), and Generative Pre-Trained Transformer 3 (GPT-3) pre-trained on the unlabeled dataset and fine-tuned on the labeled dataset. The goal is to perform sentiment analysis on movie and hotel reviews. The models are used to generate the pseudo-labels of all unlabeled data being the most confident ones added to the labeled dataset. Lastly, a student model is trained on the previously obtained dataset.

Le et al. [56] propose French Language Understanding via BERT (FlauBERT) to classify French text. The BERT and RoBERTa models are pre-trained on a multiple-source French corpus. Then, FlauBERT is fine-tuned on French Language Understanding Evaluation (FLUE) benchmark, a French evaluation setup for Natural Language Processing (NLP) tasks.

2.2.3 Intrinsically Semi-Supervised

Intrinsically semi-supervised methods are extensions of the supervised learning to include unlabeled data to optimize the objective function [1].

2.2.3.A Margin-Based

Margin-based methods focus on maximizing the margin between classes considering the density of the points [57].

In 1998 was proposed the concept of a Semi-Supervised Support Vector Machines (S3VM) [58] to maximize the separation margin and minimize the misclassification error of unlabeled data points by penalizing the points according to their distance to the closest margin boundary. Several methods have been proposed based on this approach, such as Transductive Support Vector Machines (TSVM) [59], MeanS3VM [60], and Safe Semi-Supervised Support Vector Machines (S4VM) [61]. However, there have been no improvements for text classification.

2.2.3.B Perturbation-Based

Perturbation-based methods are based on the smoothness assumption: if two samples are close in the input space, their labels should be the same so the model should be robust to local perturbations, i.e., the prediction for the clean and the noisy inputs should be similar [1]. This noise can be applied directly to the input considering the difference between the clean and noisy predictions in the loss function, or directly to the classifier [1].

Ladder Networks have the objective to augment a feed-forward neural network with additional branches to include unlabeled data [62]. *Cardellino et al.* [63] propose a Convolutional Ladder Network (CLN) for the classification of Wikipedia articles. The model is an autoencoder with skip connections from the encoder to the decoder so that the decoder can recover details discarded by the encoder. The CLN has two encoder paths (one corrupted that adds noise in each layer and one clean) and one decoder path. The decoder uses a denoising function to reconstruct the activations of each layer. This model obtains better results than a purely supervised CNN.

Another way to introduce a small perturbation is by using **layer partitioning**. Perturbing each word harms the understanding of the impact of each independent step considering that the perturbed embedding does not map back to any word [64]. Besides, the noisy output should still represent a correct sentence with a similar meaning. *Li and Sethy* [64] decompose a neural network into two layers. A layer *F* is frozen and serves as a feature extractor and as a perturbation function, introducing noise into the input. *F* contains the lower layers that tend to learn more general knowledge being domain-agnostic. A layer *U* is trained on the outputs of *F* and contains task-specific knowledge. This layer partitioning algorithm can be combined with any semi-supervised learning method:

- Π**-model** is a simplification of Ladder Networks where the encoder is removed [62]. All training data is classified and then these predictions are used as targets to minimize the difference between the output of two perturbed network models for the same input [62]. However, because the outputs are based on a single evaluation of the network, they are unstable [62].

- Instead, **Temporal ensembling** penalizes the difference between the outputs of the previous epochs by calculating the moving average. The goal is to reduce the error of the pseudo-labels to improve the overall accuracy [65]. The output is smoothed over multiple perturbations [1]. However, at the first training iterations, the model keeps the problems of Π-model, so the outputs are stabilized with a bias correction at each step [62]. The training is faster since there is only one forward pass through the network [62]. However, this model requires a large amount of memory to keep the previous predictions for all training data [62].

- **Mean teacher** considers the moving average over connection weights at each training iteration to reduce the time to incorporate the newly learned information [1, 62]. This model separates the

teacher and the student, being the weights of the teacher model an exponential moving average of the students' weights [62]. With a large number of training iterations, the weights of the teacher model convert to the ones of the student model and so there could be the propagation of biases and unstable predictions [62].

- **Dual students** considers two independent student models that are simultaneously trained [62]. The student with the most stable prediction will provide the targets for the other, and the weights of the unstable will be updated [62].

- The previous models introduce random noise that can leave the model vulnerable to perturbation in the adversarial direction [62]. **Virtual Adversarial Training (VAT)** takes into account the directionality of the perturbation introducing adversarial noise that approximates the perturbation to the neighbor inputs to obtain a robust model that can generalize and correctly classify original and adversarial examples [1, 66]. VAT is the extension of Adversarial Training to semi-supervised learning using regularization, to solve the problem of the inexistence of labeled data [66].

Li and Sethy [64] have chosen to use Π-model and temporal ensembling. In the first one, each input point goes through *F* twice, obtaining two perturbed outputs. Their labels will be given by *U*. For the temporal ensembling model, each input goes through *F* once, and the target is an ensemble of previous predictions. After a sufficient amount of training, *U* becomes saturated, and so *F* will be gradually unfrozen to learn task-specific features. A transformer encoder is pre-trained, and a BERT-tokenizer is used to generate a [CLS] token for each sentence having a linear layer for classification. This model has a significant performance increase when compared to the previous ones occurring the best results with Π-model using all labels and unsupervised data and with Temporal Ensembling for a smaller number of labels.

Meel and Vishwakarma [67] use self-ensembling via temporal ensembling to detect fake news articles to use the semantics of the labeled data and to understand patterns of the unlabeled data. Each word is represented by a Word2Vec embedding. A portion of the annotated corpus is used to train the CNN. The output of the previous epochs will be aggregated and compared to the current epoch to obtain a more accurate prediction for the unknown labels. The authors have proved that self-ensembling provides consensus predictions better than the most recent output for the unlabeled data.

Miyato et al. [66] proposed one of the first approaches of VAT to text classification. Considering that text input is discrete, the perturbation is done on the word embeddings. The LSTM is pre-training with the word embeddings. After that, the LSTM receives a normalized word embedding and a perturbation at each time step. The authors also used a bidirectional LSTM model by adding another LSTM in the reverse order to predict the label. The best results were obtained with VAT on bidirectional LSTM, concluding that VAT can improve the classification performance and the quality of the word embeddings.

Miao et al. [68] develop Snippext to perform sentiment analysis. The model is based on MixMatch [69] that is composed of a data augmentation step, followed by label guessing, sharpening, and a final MixUp step. Data augmentation is performed by MixDA. The sentences are divided into aspect terms, opinions, and sentiments. To keep the structure and meaning of the sentence, the authors only perform replacements, insertions, deletions, and swaps on the tokens that do not belong to any aspect or opinion term. There is also an operation over spans that replace a span with another of the same type. To select a token or span a combination of different sampling strategies is used:

- Uniform Sampling: Each token or span has an equal probability;
- Importance-based Sampling: Each token or span has a probability proportional to its importance according to Term Frequency–Inverse Document Frequency (TF-IDF) or the span's frequency;
- Semantic Similarity: Used as post-sampling where each token or span has a probability proportional to its semantic similarity with the original one according to the cosine similarity.

After this step, label guessing is performed using BERT to generate pseudo-labels for the unlabeled examples. BERT is fine-tuned with the labeled and augmented data. Then, the embeddings of the pseudo-labels and the labeled examples are interpolated to be used as training data. Sharpening is performed to adjust the probability distribution of each label to improve the confidence of the predictions. Lastly, the MixUp step interpolates both augmented and unlabeled data to produce a sample between the original and the augmented data to reduce noise. This sample will be used in back-propagation to update the model. The model was evaluated with four sentiment analysis datasets obtaining an improvement in performance for all of them. The authors also concluded that using only a small percentage of the training set is enough to obtain state-of-the-art performance.

2.2.3.C Generative Models

Generative models find the distribution of classes of the labeled data and then update it with the unlabeled data [70]. The performance of the model will depend on this distribution that may not be correct [71].

Croce et al. [2] propose GAN-BERT. In 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. In this paper, BERT is used as the discriminator. To this model are added task-specific layers and SS-GAN layers. 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 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.

Another approach to generative models consists of **Variational Autoencoders (VAE)**. *Xu et al.* [72] proposed a semi-supervised sequential VAE. This approach consists of a Seq2Seq structure where each labeled point is encoded by a Recurrent Neural Network (RNN) to extract its lexical information. The points are decoded by another RNN that receives the labels at each iteration and reconstructs the data considering the probability of the data point and the associated latent variable. Then, the points go through a LSTM classifier that generates a categorical label for the input.

Similarly, *Cheng et al.* [73] propose a LSTM-based VAE where both encoder and decoder are LSTM networks. The model is used for rumor detection considering tweets. The classifier is a RNN with bidirectional LSTM, calculating the probability of a given input belonging to a category.

In the field of hate speech detection, *Qian et al.* [74] propose a Conditional Variational Autoencoders (CVAE) for a fine-grained classification task. The input text goes through a bidirectional LSTM and the resulting output goes through a Multilayer Perceptron (MLP). The posterior networks receive the encoded variable and the true labels. The prior networks receive only the encoded variable. During testing, the prior networks substitute the posterior ones, using the input and its label. The model presents better results than the remaining and has demonstrated being more stable than the remaining with a small amount of data.

2.2.4 Graph-Based

Graph-based methods create a graph to represent the data structure where each node represents a data point (labeled and unlabeled), and each edge represents a relation of similarity between the points [1, 43, 62].

2.2.4.A Graph Neural Networks

Graph Neural Networks (GNN) are an extension of neural networks to graphs, applying a recurrent layer to each node [75]. The network consists of a propagation layer and a single layer perceptron. The propagation layer is responsible for encoding the graph structure of the adjacency matrix into the model and performs a local averaging. The perceptron is applied separately to each node and updates the shared weights.

Thekumparampil et al. [75] propose an Attention-based Graph Neural Networks (AGNN) model. This approach is a variant of the GNN, assigning bigger weights to the more relevant neighbors.

Benamira et al. [76] propose a graph-based semi-supervised model for fake news detection. Each word in each news is converted to a Global Vectors (GloVe) embedding. The article embedding is the mean vector of the individual embeddings. The graph that captures the contextual similarities is constructed considering these embeddings. The missing labels are assigned by a GNN and an AGNN. Both approaches improve the accuracy when compared to other models. However, GNN obtains considerably higher results.

Huang et al. [77] propose an AGNN to give different weights to the edges according to their importance and the importance of each type. Each node represents a document, and the edges represent a citation relation. This approach adds residual connections to increase the number of layers of the graph and allows to extract data from higher-order neighbors. The node-level attention mechanism captures the importance of the neighbors of the target node to give greater weights to the most significant ones to reduce the useless information for the classification result. The class-level attention mechanism learns the importance of neighbor nodes of different categories and merges them to represent the target node. The resulting adjacency matrix goes through two stacked layers with the activation function and SoftMax to generate the outputs. This approach was tested with three citation network datasets obtaining an increase in performance.

To detect abusive language, *Mishra et al.* [78] propose a Graph Convolutional Networks (GCN) approach. GCN are composed of two stacked layers with the activation function and SoftMax to generate the outputs [79]. The nodes of the graph correspond to the authors and their tweets. Two authors are connected if one follows the other, and each tweet is connected to its author. Each node is represented as a Node2Vec embedding or by a zero embedding for the authors without connections. GCN is applied to the graph to propagate information about whether the authors of the tweets tend to produce abusive tweets. This approach obtains good results, especially in identifying sexism. In the case of racism, all tweets were written by five unique authors, so the model has limitations when classifying abusive tweets from other authors.

2.2.4.B Label Propagation

Label propagation is a graph-based semi-supervised technique analogous to the k-nearest-neighbors algorithm [80]. It assumes that data points close to each other tend to have a similar label and propagates labels from the labeled points to the unlabeled ones [62, 80].

D'Sa et al. [80] represent tweets as a pre-trained sentence embedding from the USE. The authors use a 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. 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. The intra-class and inter-class distance are similar and so the representations belonging to the same class are not close and those from different classes are not far from each other. 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.

2.3 Data Augmentation

Considering that most of the interactions present in social media do not correspond to hate speech, and given the difficulty to extract them, all these datasets are generally unbalanced. **Data augmentation** is the process of expanding an existing training dataset by implementing transformations to the already labeled data or by creating synthetic examples from this data [45, 81]. This can reduce the data scarcity by generating new comments for the minority classes [28], balancing the dataset labels, and reducing the overfit [82]. It can also help the model to better generalize to unseen data, increasing its overall performance [45]. However, data augmentation in NLP tasks is limited considering that most operations can distort the meaning of the sentence and the synonyms of a word are reduced. The most common data augmentation strategies are the following:

- Token Insertion: Insert a random token (character or word) in the sentence [83];
- Token Deletion: Delete a random token of the sentence [31, 83];
- Token Replacement: Replace a token with another. At the word level, the replacement can be a synonym [30], a hypernym, a random word, or an inflection to express a different tense, voice, person, etc. [84]. The token selection can be random or can exclude words that do not have synonyms or that would change the meaning of the sentence such as pronouns, conjunctions, prepositions, and articles [85]. These synonyms can be given by databases such as WordNet or Paraphrase Database (PPDB); pre-trained word embeddings such as FastText, Word2Vec, or GloVe; or language models such as BERT or RoBERTa. The disambiguation can be done with Part-of-Speech (POS) tagging;
- Token Swap: Swap two tokens of the sentence [83];

- Back-Translation: Translate the sentence to a different language and back to the original one [31, 86]. The paraphrases generated by this approach tend to preserve the semantics of the message [87] and are grammatically correct but the diversity is limited by the translation models [88];
- Example generation with language models: Models such as GPT-3 or DistilBERT can be trained on the minority class to generate new samples [84, 89]. It considers the context semantics being able to solve the ambiguity of the words but it requires large amounts of training data [88].

Easy Data Augmentation correspond to the steps of token insertion, deletion, replacement and swap. It is one of the simplest strategies of data augmentation and the addition of noise can improve the model robustness [88]. However, the number of additional sentences generated is limited considering that there is a limited number of synonyms and the semantics may be altered when performing too many replacements [88]. Besides, the words can have several meanings according to the context that is not considered for the alteration and the syntax and semantics may be distorted [88].

In order to balance the dataset without the need to generate new sentences, it is also frequent to use the following methods:

- Oversampling: Copy minority class points to improve the relevance of the class [84];
- Majority class sentence addition: Add majority class sentences to the minority class to make relevant features stand out and reduce the sensibility to irrelevant ones [84].

Considering that it is common to find spelling mistakes in comments retrieved from social media, it is important to add natural noise. Furthermore, there is a growth of the use of masked words in social media with an intent to avoid being detected [90], so the addition of character-level noise can be relevant. *Belinkov and Bisk* [91] add synthetic noise at character level. The authors followed four approaches: randomly swapping two letters; randomizing the order of the middle letters; randomizing the order of all letters; and finally, randomly replacing a letter for another adjacent in the keyboard. The first two methods are applied to words with a length bigger than four and the remaining to all words. Natural noise was also introduced at the word level by substituting a random word of the sentence with a common typo.

2.4 Transfer Learning

Transfer Learning consists of extracting pre-trained vector representations from large amounts of data. Then, this knowledge is transferred to a target domain [92]. In inductive transfer learning, the data in the source domain can be labeled or unlabeled being the target data labeled. In transductive transfer learning, the source domain is labeled and the target domain is unlabeled being this target the only

data to be classified. In unsupervised transfer learning, the data in both domains is unlabeled. In semi-supervised transfer learning, there is a small amount of labeled data in both source and target domains.

Domain adaptation is a type of transfer learning when the source domain has limited labeled data, and the target domain has only unlabeled data [93]. The source and target domains have the same feature space, i.e., their variables have the same n-dimensions, but different distributions.

With supervised learning, several approaches use transfer learning with models such as Naive Bayes, Logistic Regression (LR), CNN, LSTM [94], bidirectional LSTM [95, 96], and BERT. *Mozafari et al.* [97] propose a transfer learning approach based on BERT. BERT is pre-trained on general corpora being the model initialized with the pre-trained parameters and then fine-tuned on task-specific annotated datasets. The [CLS] token is used to represent the tweet and will be the input to a fully connected network that performs classification. Another approach to fine-tuning is using a CNN where all outputs of the transformer encoders are concatenated into a matrix that is used as the network input. To evaluate the model, the authors used a dataset of tweets classified in racism, sexism, neither, or both.

To solve the problem of limited labeled data, *Yuan and Wen* proposed Co-Transfer, a semi-supervised inductive transfer learning approach [92]. The model is based on TrAdaBoost, re-weighting the source domain data to give more importance to the good examples. There are three TrAdaBoost classifiers for transfer learning from the source to the target domain and another three from the target to the source domain. There are two ensemble classifiers trained on the labeled data of the source and target, respectively. At each iteration, the unlabeled samples are labeled if two other classifiers agree with the label and then, the sample is added to the labeled set to refine the group of TrAdaBoost classifiers. The final labels are generated by the group of TrAdaBoost classifiers learned to transfer from the source to the target domain. Co-Transfer uses only the source and target labeled data. Compared with TrAdaBoost trained only with the labeled data, Co-Transfer performs better. Considering TrAdaBoost using more than 40% of all available labeled data, Co-Transfer has better results. However, with less data, TrAdaBoost has a lower error rate.

Depending on the target of the hate speech, the linguistic variants, or the culture, hate speech can vary a lot. Considering this variance from domain to domain, *Sarwar and Murdock* [17] propose an unsupervised domain adaptation to detect hate speech. The source datasets are annotated tweets and the target corresponds to unlabeled tweets and posts from forums and Facebook. The goal is to generate a corpus that serves as a bridge between the source and the target datasets. For this, a sequential tagger is trained on the labeled data to divide the sentences into Context Carrier (CC) and Offensive expression or Target Group mention (OTG) present in the source domain and in the hate speech lexicon. The authors use the cosine similarity to extract the context terms that are present in source and target and contain at least two OTG spots. These examples are completed with the random OTG, labeled as hate speech, and added to the source. The same process is done with sentences that

contain zero or one OTG spot and are classified as non-hate speech to distinguish between offensive and hate speech. The authors used both character and word representations that are concatenated and fed into a bidirectional LSTM layer. This layer is used to obtain a contextual word representation and a SoftMax layer is applied to obtain the probability distribution. The tagger will then be applied to the unlabeled data to derive a lexicon of hate terms and CC for the target domain. The most similar templates to the target sentences are used to augment the labeled set. This approach is evaluated with word bidirectional LSTM, character CNN and sub-word BERT. This approach is limited to the specific templates and performs poorly with implicit hate or implicit mentions to the target group. However, it obtains better results than using only the source domain.

Gupta et al. [98] propose a semi-supervised and transfer learning approach to sentiment analysis. Each sentence is represented as a Doc2Vec embedding. The experiments were done with a single corpus partially annotated and a cross-corpora setting with two corpora. For the single corpus setting, after extracting the feature representations, the authors performed semi-supervised training to classify the unlabeled data. This was compared with a supervised approach obtaining an increase of accuracy, especially with a low proportion of training data. For the second approach, the previous corpus was used to pre-train the neural network. The model training uses manifold regularization that introduces a penalty term to the supervised loss. This is used to train a statistical model to use both labeled and unlabeled data being the unlabeled data from the previous dataset and in-domain data. The goal is to minimize the distance between the labeled outputs and the near unlabeled data. For evaluation purposes, the authors compared a pre-trained model with supervised training, a model with manifold regularization without pre-training, and a pre-trained model with manifold regularization. In one of the datasets, both pre-trained models perform better being the manifold regularization better with more labeled data. On the other, semi-supervised learning without pre-trained has drastically better results than the remaining. This is possibly because Euclidean distance fails to capture the geometry of the manifold in real-world scenarios and fails to represent the similarities of the domain.

Kang et al. [93] propose a semi-supervised transfer learning model to cross-language text classification. The authors had a large amount of labeled data from the source language and a small amount for the target language. Each sentence is represented as a TF-IDF bag of words. The authors developed a semi-supervised discriminative transfer learning method that transfers new data from the target into the source and then reconstructs the source subspace representation to the original space. Then, the classifier (Linear Ridge Regression and SVM) is trained on the labeled data in the source domain and used to classify the reconstructed data. The unlabeled data is used to improve the reconstruction. Both models provide an accuracy improvement for the majority of the languages.

Bashar et al. [99] propose a progressive domain adaptation for hate speech detection. The goal is to obtain a deep feature representation that captures the domain invariance and the differences between

the source and target domains. Considering that word embeddings ignore the order of the features and their context, the approach uses a language model based on LSTM with multiple datasets (from general to specific) to cover multiple domains. The source domain has a limited amount of labeled data while the target domain corresponds to an unlabeled dataset. Each input is mapped to a deep feature vector by a language model that will then be mapped to a label by a classification model. The language model is fine-tuned in the source domain and evaluated in the target domain to obtain a smooth probability distribution. The classifier is responsible for transforming the features to reflect the differences between each dataset. Both language model and classifier are LSTM in which each layer (one for each dataset) learns the feature vectors for the previous datasets freezing the initial layer to keep the domain invariance during training. The classifier has a linear and a SoftMax layer to classify each sentence. This approach obtains better performance than the sixteen models to which it was compared.

3

Data Collection

Contents

This chapter describes the extraction process of the data that will be used in training and testing. Section 3.1 starts by comparing the most used social media platforms in order to select the ones that are prone to be used as data sources. Section 3.2 presents the linguistic, spatial, and temporal dimensions used underlying data extraction. Sections 3.3 and 3.4 describe the properties of CO-Hate and FIGHT corpora that were specifically developed under the scope of Hate COVID-19.pt project to support hate speech detection in Portuguese. Furthermore, these sections also describe the annotation guidelines and the results of this annotation process. Lastly, Section 3.5 presents the adaptations of the existing resources to be used in the scope of our project to fit our hate speech definition.

3.1 Analysis of Data Sources

Being the detection of Portuguese hate speech the focus of this research, the first step was selecting the most adequate data sources for extracting data that may convey this phenomenon. We started by inspecting Twitter, Reddit, Facebook, YouTube, and Instagram, considering that these are the most used social media platforms in Portugal.

As mentioned by *Poletto et al.* [22], the most frequently used data source for the collection of hate speech is **Twitter**. The Twitter API[1] allows to retrieve tweets by keyword, hashtag, user, or publication date. The social media platform is based on short posts with a maximum of 280 characters. Usually, each tweet is independent of any type of context such as a video, an image, or a post, besides the tweet that it is replying to if it is the case. However, Twitter tends to have flatter conversation structures considering that it is more frequent to publish a tweet than to initiate a discussion [100]. Another advantage of Twitter is the possibility to filter tweets by geolocalization. Despite the smaller amount of geolocalized tweets, this information is crucial since we are interested in analyzing the tweets posted by the Portuguese community, and we want to exclude Portuguese tweets published by users from other countries.

In **Reddit**, it is possible to extract comments using a list of keywords or to extract comments and posts from a subreddit. Reddit allows retrieving posts as a tree, keeping the context of the conversation. The API also allows recovering the “top” or “hottest” posts, to consider the number of “upvotes” and “downvotes” or the controversy of a comment (if it has a similar number of “upvotes” and “downvotes”). In Portugal, the most used subreddits have active moderation that forbids offensive or hate speech. The remaining have less activity and so it is hard to retrieve a significant amount of relevant data. For this lack of resources, Reddit will not be used as a data source.

Facebook is also a commonly used data source in the literature. However, due to the changes in the privacy policies in compliance with the General Data Protection Regulation (GDPR) made in May

[1] https://developer.twitter.com/en/docs/twitter-api

of 2018, it is only possible to access data from a public page and it is not possible to retrieve any information about the user that has made a public comment. Using applications such as Facepager[2], it is possible to retrieve posts and comments from the public pages. Any post or comment extraction based on keywords is not possible due to the API limitations. As a direct comparison between Facebook and Twitter, "people are more vocal and overtly aggressive on Facebook" [101], making the former a potentially more interesting source in terms of hate speech. However, the limitation of only retrieving data from selected pages can limit the amount of data to perform a statistical analysis [100]. Hence, Facebook will not be used as a data source.

YouTube allows extracting comments on a video or channel. The videos can be searched by keywords to select the content that covers the chosen topic. Considering that the comments are often a reaction to the video, their interpretation may require to access the video in order to understand their context. However, being an environment that encourages discussions, there is a great potential to find hate speech.

There are very few works that use **Instagram** as a data source. This is possibly due to the access restrictions of the platform. It is only possible to extract posts and comments through hashtags or user profiles. Using tools such as Instagram Scraper[3], it is possible to obtain pictures or videos along with their captions and other metadata from selected users. However, these accounts have to be manually selected. For these reasons, Instagram will not be used as a data source in this project.

3.2 Data Selection Criteria

Taking into consideration the pros and cons associated with each social media platform, the data was retrieved from YouTube and Twitter. The composition of both the datasets used in our experiments is described in Sections 3.3 and 3.4.

The data selection criteria were the following for both corpora:

- Linguistic dimension: Portuguese language.
- Spatial dimension: Focus on the Portuguese community, considering the YouTube videos covering events that occurred in Portugal and tweets geolocated in the Portuguese territory.
- Temporal dimension: To understand the impact of the COVID-19 pandemic on the evolution of hate speech, the data was retrieved in order to include the period before and after the first lockdown in Portugal (March 19, 2021).

[2] https://github.com/strohne/Facepager
[3] https://github.com/arc298/instagram-scraper

3.3 CO-HATE Corpus

The **C**ounter, **O**ffensive and **Hate** speech (CO-Hate) corpus [4] was compiled by the project's team. This corpus is composed of 20,590 written messages, including comments and replies, posted by 8,485 different online users on 39 YouTube videos.

3.3.1 Data Collection

The collection of videos was manually selected by searching for keywords present in their title or description. These keywords can be references or be associated with topics and events targeting, directly or indirectly, three specific focus groups: African descendants, Roma, and the LGBTQI+ communities. The first two correspond to the most represented racialized minorities in Portugal. The LGBTQI+ community was reported as the most targeted group analyzed in terms of hate speech on social media [102–104]. These keywords include *negro* ['black'], *racista* ['racist'], *colonização* ['colonization'], *cigano* ['Roma'], *subsídio-dependência* ['subsidy dependence'], *RSI* ['Social Integration Income'], *LGBT*, *género* ['gender'] or *homofóbico* ['homophobic']. Some of the extracted videos are illustrated in Examples (1), (2) and (3).

(1) *A manifestação antirracista em Portugal*
The anti-racist manifestation in Portugal

(2) *Afinal há ou não há um problema com ciganos em Portugal?*
After all, is there a problem with Roma in Portugal or not?

(3) *Jovem homofóbico bate em casal gay - Experiência Social*
Homophobic hits gay couple - Social Experience

To meet the linguistic and spatial dimensions, the videos were only posted by Portuguese authors and spoken in European Portuguese. All the comments on the videos were retrieved without any further selection to understand the real distribution of hate speech and other phenomena, such as counter-speech and offensive speech. To obtain these phenomena, it is fundamental to have dialogues in the comment section, so the selected videos have more than 100 comments.

The distribution of the comments according to our potential target groups is represented in Table 3.1. The most represented class is African descendants corresponding to 40% of the retrieved comments, followed by LGBTQI+ (31%), and lastly, Roma with 28% of the comments.

Table 3.1: Distribution of the retrieved data according to the mentioned target.

Target	CO-Hate	FIGHT-Target	FIGHT-Offensive
African descendants	8,278 (40%)	22,896 (42%)	6,678 (69%)
Roma	5,862 (28%)	3,036 (06%)	346 (04%)
LGBTQI+	6,450 (31%)	27,867 (52%)	2,647 (27%)
Total	**20,590**	**54,352**	**9,671**

3.3.2 Annotation Guidelines

The guidelines were created and discussed by the senior members of the project's team. In order to consider the context of each comment, the annotations were performed after watching the video and taking into account the sequence of the conversations.

The annotation process considers four dimensions of analysis represented in Table 3.2. The same comment can contain several attributes of the same dimension.

Table 3.2: Dimensions and attributes of CO-Hate and FIGHT corpus annotation.

Discourse Type	Target	Rhetorical Strategy	Sentiment
Explicit Hate Speech	African descendants	Fear appeal	Very Negative
Implicit Hate Speech	Roma	Call to action	Negative
Offensive Speech	LGBTQI+	Personal attack	Neutral
Counterspeech	Racism	Stereotype	Positive
	Xenophobia	Irony/Sarcasm/Humor	Very Positive
	Other	Rhetorical Question	
		Other	

A comment is annotated as *Explicit Hate Speech* if it meets the previously given definition of hate speech resorting typically to explicit offensive lexicon, as seen in Example (4). *Implicit Hate Speech* corresponds to hate speech expressed using rhetorical figures, such as irony and sarcasm, as represented in Example (5). Considering our definition, any comment supporting hate speech will also be considered as such, as in Example (6). *Offensive Speech* is distinguished from hate speech considering that it does not attack a person or a group based on their social identity characteristics, as in Example (7). *Counterspeech* corresponds to any comment that tries to correct or denounce any hate speech [105], as represented in Example (8). In case of not meeting any of the described discourses types, the comment is considered as *Non Relevant*.

(4) *Racismo o c@ralho! se não fossem esses parasitas da sociedade que não querem fazer nada, Portugal era um paraíso.*

F@ck the racism! If it were not those social parasites that don't want to do anything, Portugal was

a paradise.

(5) *Pura verdade. E se for ver os exemplos de países mais evoluídos como Holanda e França, já nem ciganos lá existem. Foram corridos de lá para fora.*
Pure truth. And if you look at the examples of more developed countries like the Netherlands and France, there aren't even Roma there anymore. They were kicked out.

(6) *@User* 👍👍👍 in response to *É um facto que imigrantes provenientes do norte de Africa e da Africa subsariana, estão mais predispostos a cometer crimes*
It is a fact that immigrants from North Africa and Sub-Saharan Africa are more predisposed to commit crimes

(7) *É tudo a mesma bosta, todos esses vermes são racistas e xenofóbicos.*
It's all the same crap, all these worms are racist and xenophobic.

(8) *Nao, nao é racismo. Só é racismo se o indivíduo branco for agredido ou maltratado por causa da cor da sua pele. No caso aqui relatado, é sobre a constante actuacao abusiva da polícia contra cidadaos negros.*
No, it is not racism. It is only racism if the White individual is attacked or mistreated because of the color of his skin. The case reported here is about the constant abusive action of the police against black citizens.
in response to *e quando um branco é atacado por um cidadão de outra cor... pra v6 não é racismo?!... bando de hipócritas inúteis...*
And when a white person is attacked by a citizen of another color... Is it not racism for you? Bunch of useless hypocrites...

Besides the three previously mentioned targets, the annotation also includes racism and xenophobia to cover a more generic target. The rhetorical strategies used in the comments were also annotated. Lastly, the sentiment of the message was considered to understand the patterns of the most polarized comments and the evolution of these patterns according to the intensity of the discussions.

3.3.3 Annotation

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 annotation takes into consideration the type of discourse, the target of hate speech, the rhetorical strategies used, and the sentiment polarity and intensity, as presented in Section 3.3.2.

The annotators are Portuguese students enrolled in a bachelor's or a master's degree in Communication or in Political and Social Sciences, having between 21 and 27 years old. The annotators A, B and C belong to the mentioned target groups. More specifically, the annotation team includes: a female of African descent, a White male who identifies himself as part of the LGBTQI+ community, a female of Roma descent, a White cisgender hetero male, and a White cisgender hetero female. This selection was made in order to consider the multiplicity of perspectives, including the ones of the target groups involved.

The final labels for the golden set messages are obtained considering the majority of the annotations. Table 3.3 shows the percentage of messages classified as conveying hate speech by each annotator individually, and by the group of annotators (ABCDE). About 35% of the comments were classified as *Hate Speech*, being *Implicit Hate Speech* more frequent than *Explicit Hate Speech*. Around 23% of the comments were annotated as *Offensive Speech* and 17% as *Counterspeech*.

Table 3.3: Proportion of messages classified as 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**

The IAA was measured using Krippendorff's alpha. In order to assess the impact of social identity on the perception of hate speech, we calculated the IAA for all the annotators and the group composed of the annotators belonging to the target groups (A, B and C) and the remaining. As represented in Table 3.4, the IAA between all the annotators for the classification of hate speech was considerably low (0.478), despite providing the annotators with detailed guidelines. This is also verified for the remaining attributes, demonstrating the subjectivity and difficulty of this task, and the fragility of the existing models that usually do not consider multiple perspectives. As expected, implicit hate speech is harder to classify than explicit hate speech. Besides, the offensive speech seems to be even harder to identify, especially between annotators A, B and C, due to its similarity with hate speech. Directly comparing the two groups, the one composed by the annotators not belonging to the communities targeted reached a good agreement in the majority of the attributes and a higher agreement than the group composed by the annotators A, B and C. This corroborates the idea that the annotators' social identity may influence the perception of hate speech.

Although African descendants are the most predominant target, corresponding to 44% of the hate speech messages, only 40% of the retrieved messages for this target were classified as hate speech, as

Table 3.4: IAA by discourse type for CO-Hate corpus.

Attribute	All	ABC	DE
Hate Speech	0.478	0.360	0.735
Explicit Hate Speech	0.416	0.383	0.548
Implict Hate Speech	0.237	0.145	0.421
Offensive Speech	0.143	0.005	0.472
Counterspeech	0.419	0.358	0.762

shown in Table 3.5. Surprisingly, even though the Roma community was the theme of only 20% of the comments, 56% of these comments were classified as hate speech, corresponding to 44% of the total hate speech messages. For the LGBTQI+ community, 58% of the messages retrieved for this target were classified as hate speech, corresponding to 33% of the total hate speech messages. Except for the LGBTQI+ community, implicit hate speech is more common than explicit hate speech. Concerning counterspeech, the Roma community has less only 15% counterspeech while African descendants and LGBTQI+ have 34 and 38%, respectively.

Table 3.5: Distribution of hate speech in annotated corpora by target.

Target	CO-Hate	FIGHT
African descendants	3,288 (40%)	1,938 (18%)
Roma	3,278 (56%)	484 (45%)
LGBTQI+	2,413 (58%)	3,006 (37%)
HS Total	**7,394 (35%)**	**5,207 (29%)**

3.4 FIGHT Corpus

The **FI**ndin**G** **H**ate Speech in **T**witter (FIGHT) corpus [106] is composed of 63,450 geolocated tweets in the Portuguese territory that fulfill the previously mentioned criteria, posted by 6,728 different users. The FIGHT corpus is divided into FIGHT-Target and FIGHT-Offensive corpus according to the lexicon used to retrieve the tweets.

3.4.1 Data Collection

The majority of the tweets were retrieved from an existing database composed of tweets that have been daily collected since 2015. Tweets are frequently deleted by their owners or by Twitter, in case of violating Twitter's hateful conduct policy.[4] Besides, an account can be made private or deleted and

[4] https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

all the published tweets become unavailable. In this way, the database allowed us to retrieve a higher amount of tweets. The highest percentage of deleted tweets are related to the Roma community (18%), followed by the LGBTQI+ community (17%) and African descendants (15%).

Twitter API was used to retrieve the updated metrics from the existent tweets and to fill potential gaps in data collection. In this way, combining both data sources, we obtained an updated and more robust dataset. The tweets were filtered to corresponds to the period between August 1, 2018 and October 31, 2021, to give a similar margin before and after the first lockdown in Portugal.

The distribution of messages retrieved from each source is represented in Table 3.6.

Table 3.6: Distribution of tweets in FIGHT, according to their data source.

Data Source	FIGHT-Target	FIGHT-Offensive	Total
DB (existing)	35,832	5,576	41,408
Twitter API	17,947	4,095	22,042
Total	**53,779**	**9,671**	**63,450**

FIGHT-Target corpus was retrieved using a list of 174 non-ambiguous words that may be used to mention one of the target groups. Some of these keywords include *africano* ['African'], *racista* ['racist'], *cigano* ['Roma'], *feirante* ['street marketer'], *LGBT* and *homofóbico* ['homophobic']. Examples (9), (10) and (11) correspond to tweets extracted with this method. This approach can possibly exclude hate speech targets not included in the mentioned focus groups. Additionally, if a group has a predominantly higher percentage of hate speech, the models can start to associate words related to this topic as HS [22]. Furthermore, the number of keywords covering each target should be similar in order to ensure a more natural distribution [22].

FIGHT-Offensive corpus is composed of potentially hateful tweets. In addition to a potential mention to the target groups, the messages also contain at least one word of a lexicon of 800 offensive or insulting words or expressions related to the mentioned targets, such as *assassino* ['killer'], *cobarde* ['coward'] and *feio* ['ugly']. In this approach, the mentions to the target groups also include ambiguous words like *preto* ['black'], *não binário* ['non-binary'], *nómada* ['nomad'], in a total of 259 words. Examples (12), (13) and (14) were extracted with this method.

(9) *Ele é africano, mas é muito bom atleta.*
He is African but is a very good athlete.

(10) *Dentro não tem marroquino, tem #cigano*
Inside there is no Moroccan, there is #Roma.

(11) *ela protege os direitos dos cidadãos, das mulheres, sem abrigos, dos animais, dos lgbt, tem uma empresa que doa dinheiro e ajuda as pessoas!*

she protects the rights of citizens, women, homeless, animals, lgbts, she has a company that donates money and helps people!

(12) *O cão dos meus vizinhos chama-se preto, é racismo tbm?*
My neighbors dog is called black, is that racism too?

(13) *Sou lésbica e tenho namorado mas ele é não binário, por isso ele não tem género então eu continuo a ser lésbica*
I'm a lesbian and I have a boyfriend but he's non-binary so he doesn't have a gender so I'm still a lesbian

(14) *Miúda és mesmo ignorante. Os árabes nem existiam quando Israel foi fundada originalmente. Os Judeus estão na região há mais de 4000anos. Os árabes são nómadas e vieram em ondas invasoras. Eles próprios admitem que Israel pertence aos Judeus.*
Girl you really are ignorant. Arabs didn't even exist when Israel was originally founded. Jews have been in the region for over 4000 years. The Arabs are nomads and they came in invading waves. They themselves admit that Israel belongs to the Jews.

The lexical approach used for FIGHT-Offensive may condition the data, making it hard to retrieve implicit hate speech. However, this limitation can be overcome by exploring the FIGHT-Target corpus, which may contain both explicit and implicit hate speech. However, while CO-Hate is composed of conversations, which tend to contain all types of discourse, FIGHT was retrieved considering a list of keywords. This may condition the results considering that this retrieval method typically falls short when extracting implicit hate speech.

The distribution of the retrieved comments according to our target groups is represented in Table 3.1. The most represented class for both FIGHT corpora is African descendants, followed by LGBTQI+, and lastly, with a much smaller percentage, the Roma community.

3.4.2 Annotation

To understand the potential of this collection, a sample of 300 tweets from FIGHT-Offensive was annotated by two elements of the project's team. This sample is composed of 100 randomly selected tweets from each one of the target groups. The obtained results suggest that 40% of these are effectively offensive or hateful. For the *Hate Speech* class, the IAA was 0.753 which is surprisingly high considering the subjectivity of this task. Additionally, each tweet was also classified as *Unclear* or not, whenever the context did not allow us to understand the intention of the message. Around 11% of tweets required additional context in order to understand the intention of the user. This is another important limitation when using tweets considering that the short messages can be insufficient to correctly identify the presence of hate speech.

Later, a larger sample of 19,148 tweets was annotated by five annotators, each being responsible for annotating approximately 4,000 messages, including a sample of 1000 common to all. We opted to annotate FIGHT-Offensive and the tweets from FIGHT-Target that have been deleted considering that these have more potential to contain hate speech. Of these messages, 10,971 target African descendants, 1,068 the Roma community, and 8,116 the LGBTQI+ community. The annotation followed the previously mentioned guidelines.

The annotators follow the same criteria from CO-Hate. In this case, the annotation team includes: a female and a male who identify themselves as part of the LGBTQI+ community, a female of Roma descent, a White cisgender hetero male, and a White cisgender hetero female.

Table 3.7 shows the percentage of messages classified as conveying hate speech by each annotator individually and the group of annotators. About 29% of the tweets were classified as *Hate Speech*. Contrarily to what was seen for CO-Hate, for FIGHT *Explicit Hate Speech* is more frequent (19%) than *Implicit Hate Speech* (9%), due to the retrieval method. Around 11% of the comments were annotated as *Offensive Speech* and 20% as *Counterspeech*. In comparison to CO-Hate, we have a smaller percentage of hate speech, especially implicit, and offensive speech but a slightly higher percentage of counterspeech.

Table 3.7: Proportion of messages classified as hate speech in annotated sample of FIGHT corpus, by annotator.

Annotators	**Number of messages**	**HS** (%)
A	4,630	24
B	4,629	24
C	4,630	32
D	4,629	34
E	4,630	23
Total	**19,148**	**29**

As represented in Table 3.8, the IAA between all the annotators for the classification of hate speech is lower than the one for CO-Hate. Especially for implicit hate speech, the agreement is almost nonexistent. This may be due to the lack of context that makes it even harder to identify what could be implicit hate speech.

As seen in Table 3.5 and similarly to what happened for CO-Hate, the African descendants is the most predominant target, corresponding to 57% of the annotated messages, has only 18% of these messages classified as hate speech. Roma community was the theme of only 6% of the comments, but 45% of these comments were classified as hate speech. LGBTQI+ community was mentioned in 42% of the messages and from these, 37% were classified as hate speech. For all the targets, explicit hate speech is more common than implicit hate speech. Concerning counterspeech, as also seen in CO-Hate, the Roma community has a lower percentage of counterspeech (4%) while the remaining have

Table 3.8: IAA by discourse type for FIGHT sample.

Attribute	IAA
Hate Speech	0.362
Explicit Hate Speech	0.324
Implict Hate Speech	0.080
Offensive Speech	0.214
Counterspeech	0.268

20% each.

3.5 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 [34], 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 [6], 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

Contents

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, already manually labeled, 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 of YouTube comments contextualized by the videos, FIGHT is composed of individual tweets that are published without context. Besides, YouTube comments can have an arbitrary size while tweets are limited to 280 characters. As we have seen in Section 3.4.2, these characteristics can be a limitation in the annotation process.

This section describes the approach used to solve the research problem. Section 4.1 presents the experiments to obtain an initial baseline model. Section 4.2 describes the ensemble model. Lastly, Sections 4.3 and 4.4 present the steps to increase the amount of training data, using data augmentation and domain adaptation, respectively.

4.1 Initial Experiments

In order to establish a baseline, we have started by considering a dummy classifier that classifies all examples as *Hate Speech*.

As seen by *Alsafari and Sadaouia* [43], CNN obtained the best results for hate speech classification and SVM showed to be the best model in terms of complexity. For these reasons, we have tested these two approaches. SVM is a binary classifier that finds a hyper-plane that separates two classes and maximizes the margin between the points and the hyper-plane [43]. CNN acts as an N-gram feature extractor with an embedding layer to convert the input sequence into a 2-D matrix and a dropout and max pooling layers that transform the embedding matrix into a one-dimensional vector [43]. Random Forest (RF) and LR algorithms were also applied to explore different approaches considering that are often applied for text classification [107]. RF are an ensemble of decision trees using different samples of the training data with replacement being the classification done by majority voting. LR estimates the probability of each independent variable [107].

The initial experiments use pre-trained Word2Vec embeddings[1] available for Portuguese.[2] Word2Vec generates a vector representation for each word in order to capture its semantic. Words closer in the vector space appear in similar contexts so the neighbors will give the context of the word. These dimensional vectors were generated from 17 different corpora from Brazilian and European Portuguese.

As seen in Table 4.1, from all these models, only CNN obtained better performance than the weak baseline, so it was the only one used going forward. These results were obtained with the 300 annotated tweets from FIGHT as the test set and using Word2Vec SkipGram 50 as embeddings.

[1] https://www.tensorflow.org/tutorials/text/word2vec
[2] http://nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc

Table 4.1: Performance of baseline models.

Model	Acc	HS Class			Macro Average		
		Prec	Rec	F1	Prec	Rec	F1
Dummy Classifier (all HS)	0.245	0.245	1.000	0.393	0.122	0.500	0.196
SVM	0.675	0.209	0.036	0.062	0.454	0.489	0.433
CNN	0.600	0.411	0.483	0.444	0.566	0.570	0.566
RF	0.668	0.230	0.052	0.085	0.465	0.489	0.441
LR	0.675	0.236	0.044	0.075	0.469	0.492	0.439

4.2 Ensemble Model

This work will follow a self-training approach with an ensemble of three models to reduce the bias of each one.

The first model corresponds to the CNN model, which obtained the best results from the previously tested models. The remaining are GAN-BERT and a label propagation model. Both models have been tested in hate speech detection obtaining a performance improvement when compared to the previous models [80, 108].

The classifiers are trained with a sample of labeled data. Then, at each iteration, they classify the unlabeled data. The most confident predictions (above 0.99) are added to the labeled set and the models are fine-tuned with them.

4.2.1 CNN

CNN will serve as a weaker and faster baseline and as a tiebreaker between the other two models' predictions. Besides the previous experiment, we have tested Word2Vec, FastText[3] and GloVe[4] embeddings with 50, 100 and 300 dimensions.[5]

GloVe considers the local context of the word and its co-occurrences in the corpus. The embeddings relate to the probabilities that two words appear in the same context in a large corpus [109]. FastText is an extension of Word2Vec. It is faster and considers character N-grams. The embedding of a word is given by the sum of the N-grams embedding. This allows the generation of representations for rare words or words not present in the training data and to deal with misspelling [109], which are particularly frequent in social media comments. For Word2Vec and FastText, we have experimented using Continuous Bag-of-Words (CBOW) model, where the order of the words in the sentence is not considered, and SkipGram, which considers the context by giving a higher weight to the closer words [109].

[3] https://fasttext.cc/
[4] https://nlp.stanford.edu/projects/glove/
[5] http://nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc

The training data was randomly split into 80% for the training and 20% for validation. The model was trained for 25 epochs with patience 5, considering the model with the best validation AUC score and using the Adam optimizer.

4.2.2 GAN-BERT

The second model combines a GAN and a BERT-based model, based on GAN-BERT [2], and is an improvement of BERT-based models. The goal is to find the distribution of classes for the labeled data and update it with the unlabeled data.

The input will be encoded by a BERT-based model. The generator will produce messages similar to the original data in order to introduce noise and improve the classifier. The discriminator will try to distinguish between these artificially generated messages and the ones belonging to the original data. In the end, the discriminator will also perform the classification. The proposed architecture is described in Figure 4.1.

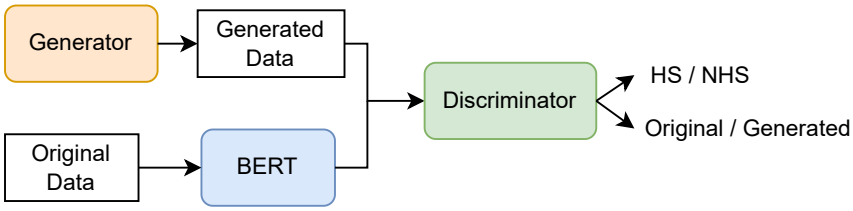

Figure 4.1: GAN Architecture. Adapted from [2].

We have tested three different pre-trained BERT-based models to find which performed better:

- Multilingual BERT pre-trained on Wikipedia articles on 104 languages;[6]
- BERTimbau [110] pre-trained on Brazilian Web as Corpus (BrWaC), a web corpus for Brazilian Portuguese;
- Fine-tuned HateBERT, a re-trained BERT model for abusive language detection in English [111] fine-tuned with CO-Hate.

The maximum sequence length of each message was defined as 350 tokens to ensure the efficiency of the model without losing too much information. The model was trained for 10 epochs with 5 patience, considering the model with the best F1-score for the positive class.

4.2.3 Label Propagation

The label propagation model represents the data as a graph where each data point corresponds to a comment and the edges represent the similarity between two comments. Each edge has a weight

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

where a larger weight w_{ij} represents a higher similarity between the nodes *i* and *j*. The model uses the similarities between the points 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 is based on the one from *D'Sa et al.* and uses the scikit-learn library [112].

We tested representing the sentences with Doc2Vec[7] and USE.[8] While Word2Vec generates a vector representation for each word, Doc2Vec computes an additional vector for every document to encapsulate the entire meaning of the message [109]. USE uses a transformer architecture to consider the context of the sentence. These two embeddings were chosen considering that have obtained better results than the remaining, such as Word2Vec, GloVe and FastText [113].

The model used the k-nearest neighbors algorithm with a maximum of 1000 iterations and neighbors between 3 and 50.

4.3 Data Augmentation

Hate speech represents a small amount of the retrieved data, creating an unbalanced dataset that can induce bias in the models. The goal is to obtain a balanced corpus so that all labels have similar relevance. For this purpose, the manually annotated corpus will be augmented to obtain a more balanced amount of hate speech.

As mentioned in Section 2.3, there are several approaches to data augmentation in NLP. To keep the meaning and the sentiment of each comment, this processing will be done with back translation. Google translate API – Googletrans[9] – will be used to generate paraphrases of the minority class examples. Each one of the 2,191 tweets from ToLR-BR and HPHS that have been classified as hate speech will be translated into English and back into Portuguese in order to generate more examples while preserving the semantics of the message [87]. Character level operations would be useful to incorporate spelling mistakes and masked words however, being both corpora retrieved from social media, these phenomena should already be naturally present.

To incorporate more data, due to the low resources in Portuguese, a sample of the previously mentioned English datasets will also be translated into Portuguese. To consider more domain variety, a maximum of 4,000 comments were selected from each corpus, obtaining a total of 21,476 messages.

The proposed approach will be tested both with and without the augmented data to understand its impact.

[7] https://radimrehurek.com/gensim/models/doc2vec.html
[8] https://tfhub.dev/google/universal-sentence-encoder-multilingual/3
[9] https://py-googletrans.readthedocs.io/en/latest/

4.4 Domain Adaptation

To improve the initial models, domain adaptation will be used in combination with semi-supervised classification. Considering CO-Hate as the source domain and FIGHT corpus as the target domain, the goal is to generate an intermediary corpus following an approach similar to the one mentioned by *Sarwar and Murdock* [17].

Following the same nomenclature, each message will be divided into an OTG and a CC. The OTG should be a noun phrase identifying the target of the hate speech or an insult. The noun phrase is identified using spaCy Part-of-Speech Tagging[10] and must contain one of the keywords of the lexicon used to extract the tweets. The CC corresponds to the remaining portion of the sentence. Considering that the sentences from FIGHT tend to be simpler, we will use the 2,191 tweets annotated as hate speech and find the most similar ones from the source domain using cosine similarity. The sentences of the source domain are filled with the target OTG, labeled with the same label, and added to the labeled data. This new labeled dataset will be then used with the previously described model. This approach performs better with explicit hate speech but has proven to be an overall improvement when compared to using only the source domain. The proposed architecture is described in Figure 4.2.

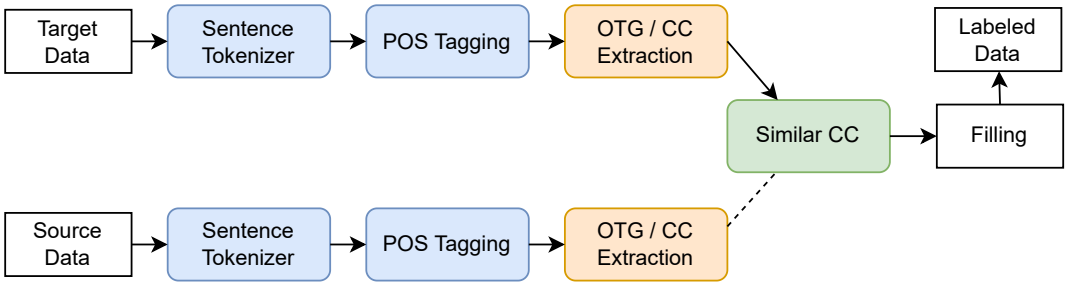

Figure 4.2: Domain Adaptation Implementation Scheme.

To illustrate this procedure, we will consider Example A from Figure 4.3 extracted from the FIGHT corpus. In this case, the OTG corresponds to the noun that is present in our lexicon of hate speech targets. The most similar sentence in the CO-Hate corpus is Example B so the noun phrase in B will be replaced by the OTG identified in A. The sentence, as illustrated in Example (15), will be then added to the labeled data.

A.	ele	é	tão	paneleiro	não	aguento						
	PRON	VERB	ADV	NOUN	ADV	VERB						
				OTG								
B.	Ele	é	tão	boa	pessoa	que	nem	o	próprio	país	o	quer
	PRON	VERB	ADV	ADJ	NOUN	PRON	CCONJ	DET	ADJ	NOUN	PRON	VERB

Figure 4.3: Domain Adaptation Example with OTG Identification.

[10] https://spacy.io/models/pt

Additionally, more variations of hate speech for more targets will be artificially generated. For each one of the 2,191 tweets, each OTG identified will be replaced by a random word in the lexicon. For this purpose, the lexicon was subdivided into "target group mention" (*lésbica* ['lesbian']), "source of hate speech" (*homofóbico* ['homophobic']) and "insult" (*feio* ['ugly']). Example (16) corresponds to one of the generated variations.

(15) *ele é tão **boa pessoa** que nem o próprio país o quer* → *ele é tão **paneleiro** que nem o próprio país o quer*
he is such a good person that not even his own country wants him → *he is such a faggot that not even his own country wants him*

(16) *O Daniel é tão **gay*** → *O Daniel é tão **preto***
Daniel is so gay → Daniel is so black

5

Experimental Results

Contents

This chapter describes the conducted experiments. Section 5.1 starts by exploring different embeddings for each model. Section 5.2 assesses the impact of different types of pre-processing. Section 5.3 presents experiments performed by each model individually, considering several subsets of training data. Section 5.4 presents results of including the additional resources generated from back-translation and by translating the English datasets. Section 5.5 includes the generated examples to create an intermediary corpus and the additional artificially generated tweets.

Section 5.6 presents the results of the ensemble model using CO-Hate as training data, and the 1300 annotated tweets from the FIGHT corpus as test data. Section 5.7 presents the error analysis for these experiments.

In order to understand the potential of the models, in Section 5.8 and Section 5.9 the models are applied to the inverse task, i.e., from FIGHT to CO-Hate, and in the same domain. Lastly, Section 5.10 tries to compare this approach with the previously presented in the literature, and Section 5.11 summarizes these results.

5.1 Different Embeddings Experiments

We have started by combining our modeling approaches with different embeddings. The results achieved are summarized in Table 5.1. With respect to CNN model, Skip-Gram demonstrated to be better than CBOW since the order of the words in a sentence is important to understand its meaning. Comparing the three embedding types, GloVe was considerably lower performance than the remaining. For all embeddings, using 300-dimensional vectors led to worse performance, possibly due to overfitting. The best results were obtained with Word2Vec Skip-Gram with 50-dimensional vectors.

In what concerns GAN-BERT, contrarily to the expected, the best results were obtained with Multilingual BERT. We expected better performance for HateBERT considering that it was fine-tuned with hate speech data. Then, it would be expected that BERTimbau would perform better than Multilingual BERT, considering that BERTimbau was trained using web corpora that are more likely to include toxicity than the Google Books corpus used for Multilingual BERT. We decided to include some pre-processing steps since we believe that the results obtained may be influenced by the noise present in the messages. For the label propagation model, Doc2Vec obtained better results than USE. This result was not expected considering that USE has a higher synonym rank [114] and has proven to perform better in similar contexts [115].

Table 5.1: Performance of different embeddings for each model.

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
CNN	Word2Vec CBOW 50	0.333	0.200	0.634	0.304	0.446	0.439	0.333
	Word2Vec CBOW 100	0.384	0.209	0.607	0.311	0.470	0.462	0.377
	Word2Vec CBOW 300	0.293	0.188	0.628	0.289	0.412	0.411	0.293
	Word2Vec Skip-Gram 50	0.325	0.211	0.708	0.325	0.459	0.459	0.325
	Word2Vec Skip-Gram 100	0.340	0.201	0.631	0.305	0.449	0.442	0.338
	Word2Vec Skip-Gram 300	0.378	0.189	0.520	0.277	0.446	0.428	0.366
	FastText CBOW 50	0.304	0.194	0.648	0.299	0.426	0.425	0.304
	FastText CBOW 100	0.312	0.192	0.591	0.290	0.437	0.426	0.334
	FastText CBOW 300	0.335	0.174	0.507	0.259	0.416	0.395	0.328
	FastText Skip-Gram 50	0.325	0.196	0.628	0.299	0.438	0.431	0.324
	FastText Skip-Gram 100	0.359	0.198	0.658	0.305	0.476	0.472	0.358
	FastText Skip-Gram 300	0.365	0.184	0.513	0.271	0.437	0.417	0.354
	GloVe 50	0.318	0.202	0.671	0.311	0.444	0.442	0.318
	GloVe 100	0.338	0.193	0.594	0.292	0.439	0.428	0.336
	GloVe 300	0.338	0.186	0.557	0.279	0.430	0.415	0.334
GAN-BERT	Multilingual	0.705	0.333	0.289	0.309	0.565	0.558	0.561
	BERTimbau	0.660	0.260	0.262	0.261	0.520	0.520	0.520
	HateBERT	0.608	0.265	0.403	0.320	0.528	0.536	0.522
Label Propagation	Doc2Vec	0.609	0.273	0.359	0.310	0.521	0.525	0.519
	USE	0.619	0.266	0.314	0.288	0.514	0.516	0.514

5.2 Pre-Processing Experiments

In order to understand the impact of pre-processing, for each model, we have applied the following steps [47]:

- Noise removal: remove processing errors in the data retrieval;
- Removal of repetitions of two 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;
- Removal of user's mentions;
- Removal of links.

As presented in Table 5.2, for the CNN model, the performance decreased by applying the pre-

processing steps. This is potentially because 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. The best results for GAN-BERT were obtained with the full pre-processing considering that pre-processing puts the emphasis on the message. Now, HateBERT obtains the best results, as expected. For the label propagation model, the best results were obtained with pre-processing and USE, as predicted.

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 5.2: Pre-processing impact for each model.

Model		Acc	HS Class			Macro Average		
			Prec	Rec	F1	Prec	Rec	F1
CNN	Word2Vec CBOW 50	0.342	0.189	0.570	0.284	0.435	0.422	0.337
	Word2Vec CBOW 100	0.366	0.185	0.517	0.272	0.438	0.419	0.355
	Word2Vec CBOW 300	0.370	0.185	0.513	0.272	0.439	0.420	0.358
	Word2Vec Skip-Gram 50	0.384	0.197	0.550	0.291	0.456	0.442	0.373
	Word2Vec Skip-Gram 100	0.310	0.198	0.661	0.305	0.435	0.433	0.310
	Word2Vec Skip-Gram 300	0.322	0.188	0.591	0.285	0.426	0.416	0.320
	FastText CBOW 50	0.408	0.184	0.463	0.264	0.447	0.427	0.384
	FastText CBOW 100	0.321	0.193	0.617	0.294	0.432	0.425	0.320
	FastText CBOW 300	0.349	0.177	0.503	0.262	0.425	0.403	0.340
	FastText Skip-Gram 50	0.338	0.207	0.668	0.316	0.458	0.454	0.337
	FastText Skip-Gram 100	0.393	0.201	0.554	0.295	0.462	0.450	0.381
	FastText Skip-Gram 300	0.309	0.184	0.584	0.279	0.416	0.406	0.308
	GloVe 50	0.333	0.209	0.688	0.321	0.460	0.458	0.333
	GloVe 100	0.359	0.186	0.530	0.275	0.437	0.419	0.350
	GloVe 300	0.354	0.181	0.517	0.268	0.431	0.411	0.345
GAN-BERT	Multilingual	0.634	0.377	0.367	0.372	0.557	0.556	0.557
	BERTimbau	0.608	0.361	0.424	0.390	0.550	0.554	0.550
	HateBERT	0.614	0.369	0.435	0.400	0.557	0.562	0.557
Label Propagation	Doc2Vec	0.576	0.300	0.326	0.312	0.503	0.503	0.503
	USE	0.573	0.308	0.357	0.331	0.509	0.510	0.509

5.3 Considering Different Subsets from Different Annotators

As previously mentioned, the annotation process involved five annotators that led to a low IAA, demonstrating the difficulty and subjectivity of the task. To assess the perspective of each annotator in the

Table 5.3: Performance of the CNN model based on the perspective of annotators.

Annotators	Acc	HS Class			Macro Average		
		Prec	Rec	F1	Prec	Rec	F1
A	0.373	0.203	0.591	0.302	0.460	0.449	0.366
B	0.535	0.267	0.591	0.368	0.539	0.555	0.500
C	0.372	0.191	0.537	0.281	0.446	0.430	0.362
D	0.425	0.220	0.594	0.322	0.488	0.485	0.412
E	0.582	0.218	0.319	0.259	0.492	0.490	0.484
BD	0.647	0.288	0.366	0.322	0.541	0.548	0.542
DE	0.412	0.211	0.570	0.308	0.476	0.468	0.399
ABC	0.335	0.207	0.671	0.316	0.456	0.453	0.334
ABD	0.405	0.236	0.711	0.354	0.511	0.513	0.402
ABDE	0.311	0.203	0.688	0.314	0.442	0.443	0.311
ABCDE	0.325	0.211	0.708	0.325	0.459	0.459	0.325

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 the best inter-annotator agreement results.

Table 5.3 presents the results for the CNN model. As we can see, the selection of the training data severely impacts the performance of the model. The following experiments were carried out with the sample that achieved better results, namely the data annotated by annotators B and D, given the macro F1 score and the overall performance. Considering that this subset has a low amount of data, we will also perform the experiments using the data annotated by all the annotators.

Table 5.4 presents the results for GAN-BERT. For this model, we opted to use the data annotated by annotators D and E, and by all the annotators given the lower dimension of the subset.

The results for the label propagation model are shown in Table 5.5. For this model, we opted to use the data annotated by all the annotators.

As mentioned by Carvalho et al. [4], annotators A, B, and C that belong to the target groups considered have a lower agreement rate than the one composed by annotators D and E, who do not belong to any potential marginalized group. 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 CNN, from Table 5.3, although recall and F1 are higher for ABC, the sample composed of annotators D and E obtained globally better results. For GAN-BERT, from Table 5.4, it is clear that the sample composed of annotators D and E obtained globally better results. For the label propagation model, from Table 5.5, we can see the opposite behavior of the CNN. Considering this variability according to the

Table 5.4: Performance of the GAN-BERT model based on the perspective of annotators.

Annotators	Acc	HS Class			Macro Average		
		Prec	Rec	F1	Prec	Rec	F1
A	0.658	0.262	0.272	0.267	0.522	0.522	0.522
B	0.564	0.250	0.450	0.321	0.517	0.524	0.500
C	0.580	0.246	0.403	0.305	0.513	0.518	0.502
D	0.605	0.240	0.336	0.280	0.508	0.510	0.504
E	0.567	0.249	0.440	0.318	0.516	0.522	0.500
BE	0.548	0.246	0.470	0.323	0.515	0.520	0.492
DE	0.557	0.295	0.671	0.410	0.569	0.597	0.528
ABC	0.598	0.259	0.406	0.316	0.523	0.530	0.516
BCE	0.565	0.275	0.547	0.366	0.542	0.559	0.518
BCDE	0.599	0.259	0.403	0.315	0.523	0.530	0.516
ABCDE	0.614	0.369	0.435	0.400	0.557	0.562	0.557

models, we cannot extract a conclusion but, by the majority, it seems that a higher IAA leads to higher performance.

5.4 Experiments with Additional Labeled Resources

The low results obtained in the previous sections may be caused by the different nature of the train and test datasets. For this reason, we have also added 26,670 tweets from ToLR-BR and HPHS (represented as BR) to the training data in order to include messages with a more similar structure. These datasets cover Brazilian topics and are written in Brazilian Portuguese so they can introduce noise to the training data. As seen from Table 5.6, only the CNN model benefited from this addition.

Additionally, the 18,148 tweets manually annotated by only one of the five annotators were added as training data. Considering that each tweet was only classified by one annotator, there can be annotation errors and the annotations can be biased to the personal opinions of each annotator, especially considering the additional ambiguity of the Twitter messages. As seen from Table 5.6, this noise only affected the performance of the CNN model, improving the results for the remaining two models.

Adding both the Brazilian tweets and the ones annotated from FIGHT, the CNN model decreased its performance, as expected. The GAN-BERT model obtained better results than the baseline but is more sensitive to the noise from the Brazilian tweets so the best results were obtained using only the FIGHT ones. For the label propagation model, the best results are obtained using both additional datasets.

Back translation was also used to generate more examples from the additional Brazilian datasets. For this, the sentences were translated from Portuguese into English and then, back to Portuguese,

Table 5.5: Performance of the label propagation model based on the perspective of annotators.

Annotators	Acc	HS Class			Macro Average		
		Prec	Rec	F1	Prec	Rec	F1
A	0.638	0.203	0.198	0.201	0.483	0.484	0.484
B	0.575	0.230	0.366	0.283	0.501	0.501	0.490
C	0.571	0.254	0.450	0.324	0.521	0.528	0.505
D	0.483	0.174	0.336	0.229	0.451	0.431	0.420
E	0.489	0.225	0.503	0.311	0.496	0.494	0.453
CE	0.565	0.250	0.450	0.322	0.518	0.525	0.501
DE	0.449	0.202	0.477	0.284	0.471	0.459	0.418
ABC	0.632	0.252	0.309	0.278	0.516	0.518	0.515
BCE	0.569	0.248	0.433	0.315	0.516	0.521	0.501
BCDE	0.552	0.233	0.416	0.299	0.503	0.504	0.485
ABCDE	0.573	0.308	0.357	0.331	0.509	0.510	0.509

using Googletrans. Except for the CNN model, the results for the remaining reveal a significantly lower performance, possibly due to loss of context during the translation process.

Considering the large amount of hate speech resources for English, a sample of these messages were translated into Portuguese, also using the Google translate API. In order to consider the most variety of domains, a maximum of 4,000 comments from each corpus were considered, obtaining a total of 21,476 messages. The CNN model benefited from this addition although the back translation step was more effective. For GAN-BERT, although this step lead to better results than back translation, the baseline results were still higher. Concerning the label propagation model, this step led to a better performance than back translation and was a significant improvement in comparison to the baseline.

The combinations of these resources were tested. For the CNN model, the best results were obtained using the Brazilian tweets and the results of the back translation; and the results of the back translation and the translation of the English datasets. For GAN-BERT, the best results were obtained by adding the additional FIGHT messages. Lastly, for the label propagation model, the best results were obtained with the FIGHT tweets; and the additional Brazilian messages and the FIGHT tweets.

5.5 Domain Adaptation

To surpass the differences in the nature of the train and test data, an intermediary corpus was generated. Besides, artificially generated tweets were also introduced, as described in 4.4.

The results are described in Table 5.7. For the CNN model, the additional examples contributed to better performance when considering the Brazilian datasets and back translation. Indeed, the best

Table 5.6: Performance of the models with additional labeled resources.

Model	Additional Data	Acc	HS Class			Macro Average		
			Prec	Rec	F1	Prec	Rec	F1
CNN	Baseline	0.325	0.211	0.708	0.325	0.459	0.459	0.325
	BR	0.335	0.218	0.738	0.337	0.476	0.476	0.335
	FIGHT	0.205	0.136	0.460	0.210	0.291	0.295	0.205
	BR, FIGHT	0.225	0.150	0.510	0.232	0.320	0.325	0.225
	BR, BckTrns	0.327	0.225	0.795	0.351	0.490	0.491	0.326
	BR, Trns	0.341	0.216	0.711	0.331	0.472	0.471	0.341
	FIGHT, BckTrns	0.230	0.154	0.527	0.239	0.328	0.334	0.230
	FIGHT, Trns	0.198	0.139	0.480	0.215	0.281	0.297	0.197
	BckTrns, Trns	0.752	0.409	0.188	0.257	0.600	0.554	0.554
	BR, BckTrns, Trns	0.388	0.206	0.587	0.305	0.467	0.458	0.379
	BR, FIGHT, BckTrns	0.230	0.139	0.453	0.212	0.320	0.308	0.230
	BR, FIGHT, Trns	0.205	0.148	0.517	0.230	0.293	0.314	0.204
	BR, FIGHT, BckTrns, Trns	0.203	0.135	0.456	0.208	0.288	0.292	0.203
GAN-BERT	Baseline	0.614	0.369	0.435	0.400	0.557	0.562	0.557
	BR	0.636	0.320	0.520	0.396	0.572	0.595	0.568
	FIGHT	0.796	0.555	0.560	0.558	0.712	0.713	0.713
	BR, FIGHT	0.712	0.405	0.544	0.464	0.627	0.653	0.634
	BR, BckTrns	0.707	0.267	0.358	0.302	0.555	0.570	0.558
	BR, Trns	0.639	0.390	0.308	0.344	0.555	0.547	0.548
	FIGHT, BckTrns	0.660	0.260	0.262	0.261	0.520	0.520	0.520
	FIGHT, Trns	0.705	0.333	0.289	0.309	0.565	0.558	0.561
	BckTrns, Trns	0.608	0.265	0.403	0.320	0.528	0.536	0.522
	BR, BckTrns, Trns	0.511	0.411	0.454	0.431	0.502	0.502	0.501
	BR, FIGHT, BckTrns	0.575	0.293	0.309	0.301	0.498	0.498	0.498
	BR, FIGHT, Trns	0.573	0.307	0.356	0.330	0.509	0.510	0.508
	BR, FIGHT, BckTrns, Trns	0.777	0.546	0.530	0.538	0.697	0.693	0.695
Label Propagation	Baseline	0.573	0.308	0.357	0.331	0.509	0.510	0.509
	BR	0.619	0.245	0.319	0.277	0.512	0.514	0.509
	FIGHT	0.779	0.520	0.470	0.494	0.684	0.671	0.676
	BR, FIGHT	0.802	0.559	0.648	0.600	0.559	0.648	0.600
	BR, BckTrns	0.610	0.224	0.285	0.251	0.497	0.496	0.494
	BR, Trns	0.719	0.416	0.554	0.475	0.634	0.661	0.642
	FIGHT, BckTrns	0.782	0.531	0.430	0.475	0.685	0.658	0.669
	FIGHT, Trns	0.792	0.558	0.433	0.488	0.700	0.666	0.678
	BckTrns, Trns	0.634	0.229	0.252	0.240	0.500	0.500	0.499
	BR, BckTrns, Trns	0.605	0.224	0.292	0.253	0.496	0.495	0.493
	BR, FIGHT, BckTrns	0.712	0.405	0.540	0.463	0.626	0.652	0.633
	BR, FIGHT, Trns	0.719	0.416	0.554	0.475	0.634	0.661	0.642
	BR, FIGHT, BckTrns, Trns	0.719	0.415	0.550	0.473	0.634	0.660	0.641

results for this model were obtained with these data. From the scenario with back translation and translation, the performance decreased possibly considering that these datasets are artificially generated so a higher amount of noise is introduced. As expected, the GAN-BERT model has presented better performance with these additional examples. For the label propagation model, there were improvements in the recall, F1 score, and the macro average metrics but the overall better performance was still given by the combination of the Brazilian and the FIGHT corpora, considering that our priority is to recognize hate speech.

Table 5.7: Performance of the models with domain adaptation examples.

Model	Additional Data	Acc	HS Class			Macro Average		
			Prec	Rec	F1	Prec	Rec	F1
CNN	BR, BckTrns	0.327	0.225	0.795	0.351	0.490	0.491	0.326
	BR, BckTrns, DAdp	0.357	0.258	0.963	0.407	0.600	0.570	0.352
	BckTrns, Trns	0.752	0.409	0.188	0.257	0.600	0.554	0.554
	BckTrns, Trns, DAdp	0.530	0.104	0.138	0.118	0.410	0.392	0.399
GAN-BERT	FIGHT	0.796	0.555	0.560	0.558	0.712	0.713	0.713
	FIGHT, DAdp	0.763	0.622	0.588	0.604	0.721	0.714	0.718
Label Propagation	FIGHT	0.779	0.520	0.470	0.494	0.684	0.671	0.676
	FIGHT, DAdp	0.660	0.376	0.732	0.497	0.632	0.685	0.620
	BR, FIGHT	0.802	0.559	0.648	0.600	0.559	0.648	0.600
	BR, FIGHT, DAdp	0.655	0.361	0.661	0.467	0.614	0.657	0.606

5.6 Ensemble Model

After assessing the potential of the models individually, they were combined in order to produce the labels for the FIGHT corpus. Each individual model used the best training set. For the CNN model, the best results were obtained using the results of the back translation and the translation of the English datasets with domain adaptation. For GAN-BERT, the best results were obtained by adding the additional FIGHT messages and domain adaptation. Lastly, for the label propagation model, the best results were obtained with the additional Brazilian messages and the FIGHT tweets.

Each model will classify the unlabeled FIGHT corpus at each iteration. The predictions with confidence above 0.99 will be saved. Then, the predictions given by at least two models are added to the training data. This was done in order to only keep the most reliable labels. We opted to consider two models instead of three considering the low number of predictions given simultaneously by the three models in the first iterations (between 5 and 200).

Table 5.8 represents the results of the models after 5 iterations, revealing that CNN requires more

data in order to obtain better results. For GAN-BERT the majority of the metrics decreased, corroborating that GAN-BERT is more susceptible to noise, as seen when assessing the impact of pre-processing. For the label propagation model, the metrics for the hate speech class decreased. However, the macro average metrics improved.

Table 5.8: Performance of the models after five iterations.

Model	Additional Data	Acc	HS Class			Macro Average		
			Prec	Rec	F1	Prec	Rec	F1
CNN	BR, BckTrns, DAdp	0.357	0.258	0.963	0.407	0.600	0.570	0.352
	5 iter	0.534	0.310	0.846	0.454	0.608	0.643	0.524
GAN-BERT	FIGHT, DAdp	0.763	0.622	0.588	0.604	0.721	0.714	0.718
	5 iter	0.707	0.517	0.713	0.599	0.682	0.708	0.684
Label Propagation	BR, FIGHT	0.802	0.559	0.648	0.600	0.559	0.648	0.600
	5 iter	0.764	0.486	0.540	0.512	0.673	0.685	0.678

Considering the lower IAA obtained for the 1,000 FIGHT tweets annotated by the five annotators, we intended to verify if the lower performance of the models was directly related to the IAA. While the 300 tweets obtained an IAA of 0.753, the 1,000 achieved only 0.362. As seen from Table 5.9, the models can recognize better the hate speech from the 300 tweets, especially the CNN model that obtains the best results.

Table 5.9: Performance of the models after five iterations considering the three different test sets.

Model	Acc	HS Class			Macro Average		
		Prec	Rec	F1	Prec	Rec	F1
Dummy Classifier (300 tweets)	0.215	0.215	1.000	0.354	0.108	0.500	0.177
CNN	0.747	0.811	0.903	0.854	0.425	0.461	0.440
GAN-BERT	0.693	0.600	0.743	0.664	0.691	0.683	0.673
Label Propagation	0.692	0.601	0.743	0.664	0.667	0.748	0.690
Dummy Classifier (1,000 tweets)	0.245	0.245	1.000	0.394	0.123	0.500	0.197
CNN	0.512	0.804	0.334	0.472	0.604	0.591	0.509
GAN-BERT	0.775	0.729	0.559	0.633	0.760	0.724	0.735
Label Propagation	0.661	0.510	0.585	0.545	0.635	0.643	0.637
Dummy Classifier (1,300 tweets)	0.177	0.177	1.000	0.300	0.088	0.500	0.150
CNN	0.534	0.310	0.846	0.454	0.608	0.643	0.524
GAN-BERT	0.707	0.517	0.713	0.599	0.682	0.708	0.684
Label Propagation	0.764	0.486	0.540	0.512	0.673	0.685	0.678

5.7 Error Analysis

Considering that our focus is to correctly classify hate speech, there is a small number of false negatives, i.e., cases where the models were not able to recognize hate speech.

In particular, for the CNN and GAN-BERT models, the majority of false negatives correspond to **implicit hate speech**, where is used mainly irony, which is harder to identify, even for humans. Examples (17) and (18) were wrongly classified by CNN and Examples (19) and (20) by GAN-BERT. There are also several **shorter messages** that were wrongly classified, possibly considering that CO-Hate is composed of longer messages and so, there is a smaller amount of shorter hate speech messages. Examples (21) and (22) were wrongly classified by CNN and Examples (23) and (24) by GAN-BERT.

(17) *Só cego não vê. A carne mais barata do mercado é a carne negra.*
Only the blind do not see it. The cheapest meat on the market is black meat.

(18) *Mas nos EUA existe realmente racismo policial...ou os negros estão fortemente ligados ao mundo do crime?...é que são apenas 12% da população nacional e mais de 80% na população prisional!?*
But in the USA is there really police racism...or are blacks strongly linked to the world of crime?... they are only 12% of the national population and more than 80% of the prison population!?

(19) *Ai coitadinho de mim eu sou uma vítima da sociedade homofóbica*
Oh poor me I am a victim of the homophobic society

(20) *Alguém me pode explicar porque raio os animais são todos heteros e os humanos têm diferentes sexualidades*
Can someone explain to me why the hell animals are all straight and humans have different sexualities

(21) *Foi o q eu perguntei gay*
That is what I asked gay

(22) *Que puta de gay*
What a fucking gay

(23) *Filha da puta de ciganos*
Gypsy motherfucker

(24) *Preto nojento*
Disgusting black

In what concerns the label propagation model, the majority of the cases correspond to messages where the annotators had **less concordance**. Example (25) was classified as explicit hate speech

by one annotator, implicit hate speech by two, and offensive speech by the other two. Example (26) was classified as explicit hate speech by two annotators, counterspeech by another two, and even non relevant by the last one.

(25) *Juro voces sao uns conas do crl. Se fosse um preto a pintar-se de branco para fazer de branco nunca alguem iria ficar triggered. Quem cria o preconceito são vocês com o negro, o de cor, o africano e afins. Quem não sente preconceito nao tem medo de usar a palavra preto.*
I swear you guys are fucking cunts. If it was a black painting himself white to make him white, no one would ever get triggered. The ones that create the pre-concept are you with the black, the person of color, the African, and so on. Those who don't feel pre-concept are not afraid to use the word black.

(26) *Que racismo? Essa família de cigano fez merda. Em vez de admitirem ou sei lá pedirem desculpa, não , preferem defenderem se dizendo "não ao racismo" assim tão a falar oq? Vocês foram Bater numa funcionaria e numa professora pq se sentiram vítimas de racismo? Muito menos vá next*
What racism? That gypsy family fucked up. Instead of admitting or apologizing, no, they prefer to defend themselves by saying "no to racism" what are they talking about? Did you hit an employee and a teacher because you felt like a victim of racism? much less, next

Concerning the false positives, i.e., the cases where the models classified incorrectly a sentence as hate speech, the behavior was similar for the three models. The majority of the situations correspond to the presence of **lexically ambiguous words** and expressions that the models started to associate as negative because they are **often present in hate speech** (Examples (27)- (30)). There are also some cases of **counterspeech** (Examples (31) and (32)) or **offensive speech** (Examples (33) and (34)). Counter, offensive, and hate speech tend to share a lot of the vocabulary, so it is harder for the models to distinguish them.

(27) *A cena éq sou bue esquisita a conjugar cores e tb n sei qual cor poderia comprar. Os pretos são aquela cena*
The thing is, I'm really picky at conjugating colors and I don't know which color I could buy. Black is that thing

(28) *wtf há pretos giros e pretos feios, tal como brancos. Não é só por ser preto q é sexy, tal como não é só por ser branco q é sexy*
wtf there are cute blacks and ugly blacks, just like whites. It is not just because they are black that they are sexy, just as it is not just because they are white that they are sexy.

(29) *A gente já sermes cães agora com o Covid sermes pitbulls. JJ ou Cigana no Beatriz Ângelo?*
We already are dogs, now with Covid we are Pitbulls. JJ or Roma at Beatriz Ângelo?

(30) *Eu dou block em homofobicos e racistas*
I block homophobes and racists

(31) *Generalização injusta. Um cigano criminoso, não faz dos ciganos todos criminosos. Um GNR simpatizante do Chega não faz DA GNR uma força de segurança fascista. Não usemos as mesmas armas que eles, Maçã.*
Unfair generalization. A criminal Roma does not make all Roma criminals. A GNR sympathetic to Chega does not make the GNR a fascist security force. Let's not use the same weapons as them

(32) *fake news, a comunidade LGBT não compactua com a pedofilia. pedofilia é crime.*
fake news, the LGBT community does not condone pedophilia. pedophilia is a crime.

(33) *es tao feio que ando olhaste pó céu, Deus te atirou com pedra kakaka*
You are so ugly that when you looked at the sky, God threw you with a stone kakaka

(34) *é assim, eu não gosto de ver uma rapariga toda macaca. No meu caso, gosto de estar depilada em todo o lado. Mas se a miúda gosta não é preciso meterem este tipo de treta, deixem lá a miúda e respeitem a escolha dela smh. Fodasse idiotas da merda vocês. lmao*
that's how it is, I don't like to see a girl all monkey. In my case, I like to be shaved everywhere. But if the girl likes it, there's no need to get into this kind of bullshit, leave the girl alone and respect her choice smh. Fuck you fucking idiots. lmao

5.8 From FIGHT to CO-Hate

Considering that FIGHT's messages tend to be more complex to classify, we want to analyze the results for the opposite task, i.e., using the FIGHT corpus to annotate CO-Hate messages. In this case, the training data may have more noise considering the ambiguity of the tweets from FIGHT. However, the agreement for the CO-Hate corpus is higher than the one for FIGHT. Besides, the test set is more balanced, being hate speech the majority class.

The models use the previously selected embeddings and the combinations of training data were used in the performed experiments. Two test sets were considered: the one composed of the 534 comments annotated simultaneously by the 5 annotators; and the one composed of the totality of the annotated comments, which can contain annotation errors and is more susceptible to personal bias.

As we can see from Table 5.10, CNN is the worst performing model, especially for the second test set where it is not able to surpass the baseline. Besides, while the remaining models show better results with higher amounts of data, CNN decreased its performance. The GAN-BERT shows a lower recall, having difficulties identifying hate speech instances. This may be due to the fact that only 29% of the

training data corresponds to hate speech and only 9% corresponds to implicit hate speech. Surprisingly, the label propagation is able to obtain results close to 1 for every metric.

Considering the two test sets, we can see that GAN-BERT and label propagation models were able to surpass the baseline in both situations. However, the CNN model performed poorly to classify the 20,590 YouTube comments possibly due to the higher noise.

Table 5.10: Performance of the models using FIGHT to annotate CO-Hate corpus.

Model	Data	Acc	HS Class			Macro Average		
			Prec	Rec	F1	Prec	Rec	F1
Dummy Classifier (534 comments)		0.408	0.408	1.000	0.580	0.204	0.500	0.290
CNN	FIGHT	0.517	0.644	0.411	0.502	0.542	0.541	0.516
	FIGHT, 20,056 CO-Hate	0.655	0.692	0.753	0.721	0.641	0.633	0.635
	FIGHT, BR, BckTrns, DAdp	0.448	0.538	0.472	0.503	0.444	0.442	0.441
GAN-BERT	FIGHT	0.612	0.545	0.303	0.389	0.589	0.564	0.553
	FIGHT, 20,056 CO-Hate	0.710	0.638	0.670	0.653	0.701	0.704	0.702
	FIGHT, DAdp	0.837	0.985	0.610	0.754	0.886	0.802	0.816
Label Propagation	FIGHT	0.575	0.474	0.372	0.416	0.548	0.543	0.541
	FIGHT, 20,056 CO-Hate	0.663	0.604	0.505	0.550	0.649	0.638	0.640
	FIGHT, BR, FIGHT	0.974	0.977	0.959	0.968	0.974	0.971	0.973
Dummy Classifier (20,590 comments)		0.354	0.354	1.000	0.523	0.177	0.500	0.262
CNN	FIGHT	0.449	0.654	0.313	0.423	0.506	0.506	0.448
	FIGHT, BR, BckTrns, DAdp	0.176	0.209	0.099	0.134	0.185	0.207	0.174
GAN-BERT	FIGHT	0.663	0.593	0.153	0.244	0.632	0.548	0.513
	FIGHT, DAdp	0.856	0.967	0.615	0.752	0.895	0.802	0.825
Label Propagation	FIGHT	0.543	0.352	0.346	0.349	0.498	0.498	0.498
	FIGHT, BR, FIGHT	0.995	0.991	0.995	0.993	0.994	0.995	0.995

Analyzing the classification errors, the same situations occur with the three models and are similar to the ones observed for the FIGHT corpus.

In what concerns the false negatives, the majority corresponds to implicit hate speech (Examples (35) and (36)). One difference from the FIGHT corpus is the fact that the CO-Hate messages were annotated considering the context of the conversation and, according to our definition, any message that supports hatred will also be classified as hate speech. Considering this, some messages are difficult to classify when out of the original context. Examples (37) to (39) were wrongly classified as non hate speech due to this factor.

(35) *Mas ... Mas... Ela era gaga e negra, coitada, tinha de votar nela.*
But... But... She was stuttering and black, poor thing, I had to vote for her.

(36) *@User Não falei de generalizações, mas sim de factos. Se outros povos não querem portugueses nas suas terras estão no seu direito. A história nos diz que multirracialismo + proximidade = conflitos étnico-culturais. Queira pesquisar verá que em matéria de dados criminais no ocidente comunidades africanas tem uma maior prevalência a cometer crimes. Miscigenação cria problemas de identidade, basta ver o paradigma que é o Brazil, EUA, AS. Migração proveniente de Africa pertence a uma classe que nem é muito rica (pois esses ficam em Africa) nem é muito pobre(não tem posses para migrar sequer) os que migram são aqueles que justamente são necessários para que as nações africas se desenvolvam(são aqueles com algumas posses e predispostos a melhorar as suas vidas). Não foi minha intenção lhe ofender. Estime o seu marido, pois pelo que disse ser é um cidadão de bem. Portugal é o solo dos indígenas, autóctones, nativos Portugueses ao qual me insiro. Se todo o mundo pode ser português, o que é ser português?*
@User I didn't speak of generalizations, but of facts. If other countries don't want Portuguese on their land, that's their right. History tells us that multiracialism + proximity = ethnic-cultural conflicts. If you want to research you will see that in terms of criminality, the West African communities have a higher prevalence of committing crimes. Miscegenation creates identity problems, just look at Brazil, USA, AS. Migration from Africa belongs to a class that is neither very rich (as they stay in Africa) nor is it very poor (they do not even have the means to migrate) those who migrate are those who are necessary for African nations to develop (they are those with some possessions and predisposed to improve their lives). It was not my intention to offend you. Cherish your husband, from what you said, he is a good citizen. Portugal is the soil of the indigenous, autochthonous, native Portuguese to which I belong. If everyone can be Portuguese, what does it mean to be Portuguese?

(37) *Adoro Portugal*
I love Portugal

(38) *Correto !*
Correct !

(39) 👍👍👍

The messages that were incorrectly classified as hate speech correspond to the same cases as seen for FIGHT. The majority correspond to messages using words that are frequently present in hate speech. Due to the nature of the retrieved videos, the majority of these words are *Rendimento Social de Inserção (RSI)*, *abonos*, and *subsidiodependência*, as seen in Examples (40) and (41). The remaining correspond mainly to counterspeech (Examples (42) and (43)) and offensive speech (Examples (44) and (45)).

(40) *Algo não bate certo... O valor da prestação mensal do RSI Final equivale à diferença entre os rendimentos da família e o valor do RSI. Calcula-se o valor do RSI somando: 188,68 euros por titular; 130,68 euros pelos restantes adultos; e 93,34 euros por cada criança ou jovem menor de 18 anos. Assim, neste exemplo, este senhor teria direito a receber 188,68 + 130.68 + (93,34 * 6 [visto que o 7 ainda não nasceu]) = 879,4 - o rendimento do agregado. Logo este video é FALSO.*
Something doesn't make sense... The monthly installment of the Final RSI is equivalent to the difference between the family's income and the summed RSI value. The RSI value is calculated by adding: 188.68 euros per holder; 130.68 euros for the remaining adults; and 93.34 euros for each child or young person under 18 years of age. So, in this example, this man would be entitled to receive 188.68 + 130.68 + (93.34 * 6 [since 7 is not yet born]) = 879.4 - the household income. So this video is FAKE.

(41) *Quem dúvida que é verdade, é fazer contas. Entre RSI e abonos, este senhor deve receber qualquer coisa entre os 2,5 e os 3 mil €! Contas por alto!*
Anyone who doubts that it's true, do the math. Between RSI and allowances, this man should receive anything between 2.5 and 3 thousand €! broadly!

(42) *Inclusive esse tipo de comentário generalista já oiço desde sempre (a depender do RSI e a viver à custa do estado, assim como ignorantes e delinquentes existem de todas a cores). Nem todos são iguais.*
This type of generalist comment I have always heard (depending on the RSI and living at the expense of the state, ignorant and criminals exist in all colors). Not all are the same.

(43) *@User pois, como o meu marido é preto, nascido em África, e já cá vive desde muito novo, e os meu filhos seus descendentes, logo pretos, lamento imenso, mas não preciso de pesquisar pois vivo com eles. Com muito orgulho, posso dizer que o meu marido sempre trabalhou e descontou, nunca dependendo de subsídios nem a chular ninguém, muito pelo contrário. Nem todos são iguais e existem bons e maus exemplos, infelizmente, mas isso de não querer uma Europa africanizada, temos pena, pois aposto que há muitos povos por aí fora que não querem emigrantes portugueses lá na terras deles, e não é por isso que deixamos de emigrar. Não ponham todos no mesmo saco!*
@User because, as my husband is black, born in Africa, and has lived here since a very young age, and my children, their descendants, therefore black, I am very sorry, but I don't need to research because I live with them. I can proudly say that my husband has always worked and deducted, never depending on subsidies or hustling anyone, quite the opposite. Not all are the same and there are good and bad examples, unfortunately, but that of not wanting an Africanized Europe, we are sorry, because I bet there are many peoples out there who do not want Portuguese emigrants

in their lands, and that is not why we stop emigrating. Don't put them all in the same bag!

(44) *Mas se votarmos nos outros eles retiram a este e a nós....é assim que eles fazema justiça deles...por isso em Portugal todos os partidos são lixo....só tem gente sem escrúpulos que nem a casa deles sabem governar...vendilhoes.*
But if we vote for the others, they remove from this one and from us....that's how they do their justice...that's why in Portugal all parties are rubbish....there are only unscrupulous people who don't even know how to govern their house...sellers.

(45) *Policia so bandidos esses porcos.*
Police are thugs these pigs.

5.9 In-Domain Experiments

Considering the additional complexity of testing a model in a different nature corpus, we want to understand the potential of the model applied to the same training domain.

In what concerns the FIGHT corpus, the training data corresponds to the 18,148 tweets annotated by each one of the five annotators. To have a better perception, we considered the three previously mentioned test sets. To understand the impact of the nature of the training domain, the comparison was done using only the 20,590 messages from CO-Hate and the 18,148 tweets from FIGHT.

As seen from Table 5.11, the majority of the models perform significantly better when using FIGHT as training data, i.e., using the same domain. The exception is the CNN model that appears to require higher amounts of data in order to obtain better results. Besides, CNN is highly sensitive to noise and cannot reach the baseline for the last test case.

Concerning the CO-Hate corpus, the training data corresponds to the 20,056 comments annotated by each one of the five annotators and the test data to the 534 annotated simultaneously by them. These results were compared with the use of the 19,448 tweets from FIGHT.

In this case, as represented in Table 5.12 and as expected, all models perform significantly better within the same domain.

5.10 Comparing with the Related Literature

In an attempt to compare our results with other work reported literature, we considered the work of Breazzano et al. [108] and D'Sa et al. [80] 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

Table 5.11: Performance of the models considering FIGHT corpus as training and test data.

Model	Data	Acc	HS Class			Macro Average		
			Prec	Rec	F1	Prec	Rec	F1
Dummy Classifier (300 tweets)		0.215	0.215	1.000	0.354	0.108	0.500	0.177
CNN	CO-Hate	0.723	0.860	0.794	0.825	0.576	0.595	0.581
	FIGHT	0.730	0.850	0.780	0.814	0.655	0.678	0.662
GAN-BERT	CO-Hate	0.647	0.188	0.302	0.232	0.508	0.511	0.501
	FIGHT	0.740	0.376	0.717	0.494	0.650	0.731	0.659
Label Propagation	CO-Hate	0.653	0.248	0.472	0.325	0.553	0.582	0.546
	FIGHT	0.753	0.402	0.811	0.538	0.675	0.776	0.685
Dummy Classifier (1,000 tweets)		0.245	0.245	1.000	0.394	0.123	0.500	0.197
CNN	CO-Hate	0.624	0.687	0.779	0.730	0.565	0.555	0.555
	FIGHT	0.300	0.408	0.159	0.229	0.335	0.362	0.294
GAN-BERT	CO-Hate	0.624	0.446	0.349	0.392	0.568	0.560	0.560
	FIGHT	0.777	0.546	0.531	0.538	0.698	0.694	0.696
Label Propagation	CO-Hate	0.567	0.367	0.343	0.355	0.515	0.515	0.514
	FIGHT	0.788	0.563	0.604	0.583	0.716	0.726	0.720
Dummy Classifier (1,300 tweets)		0.177	0.177	1.000	0.300	0.088	0.500	0.150
CNN	CO-Hate	0.325	0.211	0.708	0.325	0.459	0.459	0.325
	FIGHT	0.292	0.475	0.211	0.292	0.343	0.343	0.292
GAN-BERT	CO-Hate	0.614	0.369	0.435	0.400	0.557	0.562	0.557
	FIGHT	0.745	0.579	0.508	0.541	0.583	0.566	0.570
Label Propagation	CO-Hate	0.573	0.308	0.357	0.331	0.509	0.510	0.509
	FIGHT	0.771	0.616	0.594	0.605	0.724	0.719	0.722

social practices, but mostly because the testing datasets are different. Breazzano et al. [108] applied GAN-BERT to several Italian hate speech. Both the HaSpeeDe[1] and the DANKMEMES [116] 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. [80] applied a label propagation model to two English datasets from Founta et al. [117] and Davidson et al. [40] 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 to negatively impact the results. 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 Por-

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

Table 5.12: Performance of the models considering CO-Hate corpus as training and test data.

Model	Data	Acc	HS Class			Macro Average		
			Prec	Rec	F1	Prec	Rec	F1
Dummy Classifier (534 comments)		0.500	0.500	1.000	0.667	0.250	0.500	0.333
CNN	FIGHT	0.517	0.644	0.411	0.502	0.542	0.541	0.516
	CO-Hate	0.614	0.690	0.633	0.660	0.607	0.610	0.607
GAN-BERT	FIGHT	0.612	0.545	0.303	0.389	0.589	0.564	0.553
	CO-Hate	0.703	0.656	0.622	0.638	0.648	0.684	0.683
Label Propagation	FIGHT	0.575	0.474	0.372	0.416	0.548	0.543	0.541
	CO-Hate	0.678	0.614	0.569	0.590	0.665	0.661	0.663

tuguese [118]. The CNN model, our weakest model, obtained a macro average F1-score of 0.524, which is not far from the expected. However, GAN-BERT obtained a macro average F1-score of 0.678, and 0.684 for the label propagation model, which are in line with the above-mentioned results, reinforcing the potential of this approach.

5.11 Final Conclusions

Our main goal is to obtain a combination of the best possible models, i.e., the ones that are better at identifying hate speech.

The first experiments considered different embeddings and pre-processing. For CNN, Skip-Gram leads to better results as it considers the order of the words. Besides, generally, the best results are obtained with 50-dimensional vectors since the model starts to overfit with higher dimensions. The best results are obtained with Word2Vec. For the remaining models, the pre-processing steps were essential to remove the noise of the messages. In the case of GAN-BERT, HateBERT obtains the best results considering that it was fine-tuned with hate speech data, recognizing it better. For the label propagation model, the best results were obtained with USE which takes into consideration the context of the sentence. We have experimented with several subsets of data, using various combinations of annotators and additional datasets. From this, we concluded that the agreement between the annotators has a direct impact on the performance of the model. This is expected considering that a higher agreement is obtained with less controversial messages so the model will also have fewer difficulties in the classification process. Besides, the use of the additional data resulting from the annotated tweets, back translation, and translation resulted, generally, in an improvement of the performance of the models. However, the data translated from English into Portuguese introduced too much noise, deteriorating the performance of GAN-BERT and, back-translation was harmful to GAN-BERT and the label propagation

model.

Comparing both tasks, the same errors occur. For false negatives, the models struggle to identify implicit hate speech, especially when irony is used, and with messages where the annotators themselves had less concordance in the hate speech classification. When testing in FIGHT, the models have difficulties with several shorter messages that were wrongly classified, possibly considering that CO-Hate is composed of longer messages and so, there is a smaller amount of shorter hate speech messages. When testing in CO-Hate, there were some messages difficult to classify when out of the original context, corresponding to messages that are supporting hate speech that occurred previously in the conversation. Concerning false positives, the majority of the cases are due to lexical ambiguity where the models started to associate words as negative because they are often present in hate speech. Besides, there are also some cases of counter and offensive speech that are confused as hate speech.

Comparing the three models, CNN requires more data in order to obtain better results and is the worse performing model in almost all experiments. GAN-BERT appears to be more susceptible to noise while the label propagation is more stable and obtains especially good results when classifying CO-Hate messages with a higher amount of training data. Directly comparing the different tasks, as expected, using the same training and test domains leads to better results.

6

Conclusion

Contents

This chapter describes the main conclusions of this work, as well as its limitations and the directions for future work.

6.1 Conclusions

In the literature, several 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 three semi-supervised models. The first model is a CNN using Word2Vec Skip-Gram as embedding. The second one employs a GAN in combination with HateBERT, a pre-trained BERT model for English abusive language that was fine-tuned with the CO-Hate corpus. The last model is based on label propagation, using USE as embedding. The three 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 CO-Hate to automatically annotate FIGHT, a corpus composed of geolocated tweets produced in the Portuguese territory.

In order to remove some of the noise that we may encounter in social media comments, several pre-processing steps were applied. This pre-processing led to an improvement in performance for GAN-BERT and the label propagation model. The CNN model, being a simpler model, suffers the most with this partial removal of meaning and context.

To understand the impact of the social identity of the annotators on the annotations and the models' performance, several samples of CO-Hate have been used as training data. We have found that annotators not belonging to any of our target groups tend to agree more. This higher agreement reflects directly on the model performance, leading to higher results.

The additional Brazilian datasets and the annotated sample of FIGHT were also added to the training data, improving the performance of GAN-BERT and the label propagation model. CNN only benefited from the addition of the Brazilian datasets possibly due to the noise of the tweets.

Back translation was also tested in an attempt to generate more hate speech examples. Additionally, the English resources were translated into Portuguese due to the low amount of Portuguese annotated data that follows our definition of hate speech. Only the CNN model benefited from the back translation and the translated examples were also beneficial for the model's performance. Both steps deteriorated the GAN-BERT performance. Concerning the label propagation model, the translated examples led to a better performance than back translation and was a significant improvement in comparison to the baseline. This may be due to the fact that the extra translation step introduces additional noise.

The reverse task, from FIGHT to CO-Hate, was also tested. In this case, the training data - FIGHT

corpus - may have more noise considering the lack of context of the messages. However, the IAA for the CO-Hate corpus is higher than the one for FIGHT. CNN is the worst performing model and performs worse with more amounts of training data. The GAN-BERT shows a lower recall, having difficulties identifying hate speech instances. This may be due to the fact that only 29% of the training data corresponds to hate speech and only 9% corresponds to implicit hate speech. The label propagation model obtains surprising results, close to 1 for every metric.

Directly comparing these tasks with the ones using the same training and test domain, using the same domain obtained better results, as expected. However, our cross-domain results are still in line with the ones seen in the literature for the same domain. Our three models obtained between 0.454 and 0.854 F1-score for the Hate Speech class and a macro average F1-score between 0.524 and 0.684. As expected, CNN proved to be a weaker model and more susceptible to noise. The label propagation approach proved to be more stable, with similar performance to the GAN-BERT model, besides being a less complex model, and hence faster to train with larger amounts of data. However, all models obtained good performance, especially considering the different nature of the corpora.

6.2 Limitations and Future Work

As limitations, the models presented difficulties in recognizing implicit hate speech considering that it requires a context to understand some of the rhetorical figures used. Besides, the models struggle to distinguish between counter, offensive, and hate speech considering that they share a lot of the vocabulary.

When classifying FIGHT messages, several shorter messages were wrongly classified, possibly considering that CO-Hate is composed of longer messages and so, there is a smaller amount of shorter hate speech messages. Concerning the CO-Hate corpus, these messages were annotated considering the context of the conversation and, according to our definition, any message that supports hatred will also be classified as hate speech. With this, several messages are difficult to classify when out of the original context and are wrongly classified by the models.

Another important limitation to consider is the disagreement between the annotators. Indeed, a considerable amount of classification errors were messages where the annotators had less concordance.

Besides, some messages are incorrectly classified as hate speech if they contain lexical ambiguity, so the models start to associate these ambiguous words that are often present in hate speech as negative. This limitation is also described in the literature.

In terms of future directions, considering the problems in distinguishing between counter, offensive, and hate speech, it would be important to consider all these classes in the classification. Although, this would require a larger amount of annotated data for each one of these classes. Especially when

classifying CO-Hate messages, the context of the entire conversation should be used. In the case of FIGHT messages, to solve the ambiguity problem, the replies to the tweet could be used. However, there are several cases where there are not replies or where the replies still are not enough to understand the real meaning of the message. Lastly, the automatic classification should also cover the remaining dimensions used in the annotation process.

Bibliography

[1] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, pp. 373–440, Nov. 2020. [Online]. Available: https://doi.org/10.1007/s10994-019-05855-6

[2] D. Croce, G. Castellucci, and R. Basili, "GAN-BERT: Generative Adversarial Learning for Robust Text Classification with a Bunch of Labeled Examples," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2114–2119, Jul. 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.acl-main.191

[3] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM Computing Surveys*, vol. 51, no. 4, pp. 1–30, Jul. 2019. [Online]. Available: https://doi.org/10.1145/3232676

[4] P. Carvalho, D. Caled, C. Silva, F. Batista, and R. Ribeiro, "The expression of Hate Speech against Afro-descendant, Roma and LGBTQ+ communities in YouTube comments," *submitted*, 2022.

[5] A. A. Siegel, "Online hate speech," in *Social Media and Democracy*, J. A. T. Nathaniel Persily, Ed. Cambridge University Press, Aug. 2021, ch. 4, p. 67. [Online]. Available: https://doi.org/10.1017/9781108890960

[6] P. Fortuna, J. Rocha da Silva, J. Soler-Company, L. Wanner, and S. Nunes, "A Hierarchically-Labeled Portuguese Hate Speech Dataset," *Proceedings of the Third Workshop on Abusive Language Online*, pp. 94–104, Aug. 2019. [Online]. Available: https://doi.org/10.18653/v1/w19-3510

[7] D. Whitehead, "Covid-19: Online hate speech has increased by 20% in the uk since start of the pandemic, research finds," *Sky News*, Nov 2021. [Online]. Available: https://news.sky.com/story/covid-19-online-hate-speech-has-increased-by-20-in-the-uk-since-start-of-the-pandemic-research-finds-12469

[8] "Covid-19 Fueling Anti-Asian Racism and Xenophobia Worldwide," *Human Rights Watch*, May 2020. [Online]. Available: https://www.hrw.org/news/2020/05/12/covid-19-fueling-anti-asian-racism-and-xenophobia-worldwide

[9] N. White, “Anti-asian hate speech surged by 1,662% during the pandemic, study finds,” *Independent*, Nov 2021. [Online]. Available: https://www.independent.co.uk/news/uk/home-news/anti-asian-hate-speech-covid-b1957474.html

[10] “Hate speech policy,” *Google - YouTube Help*, 2022. [Online]. Available: https://support.google.com/youtube/answer/2801939

[11] “Twitter’s Hateful conduct policy,” *Twitter Help Center*, 2022. [Online]. Available: https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

[12] “The EU Code of conduct on countering illegal hate speech online,” *European Commission - European Commission*, Oct 2021. [Online]. Available: https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

[13] “Facebook Community Standards: Hate Speech,” *Transparency Center*, Jan 2022. [Online]. Available: https://transparency.fb.com/policies/community-standards/hate-speech/

[14] M. Bhatia, T. S. Bhotia, A. Agarwal, P. Ramesh, S. Gupta, K. Shridhar, F. Laumann, and A. Dash, “One to rule them all: Towards Joint Indic Language Hate Speech Detection,” *arXiv:2109.13711*, Sep. 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2109.13711

[15] B. Perrigo, “Facebook says it’s removing more hate speech than ever before. But there’s a catch,” *Time*, Nov 2019. [Online]. Available: https://time.com/5739688/facebook-hate-speech-languages/

[16] T. Simonite, “Facebook Is Everywhere; Its Moderation Is Nowhere Close,” *Wired*, Oct 2021. [Online]. Available: https://wired.com/story/facebooks-global-reach-exceeds-linguistic-grasp

[17] S. M. Sarwar and V. Murdock, “Unsupervised Domain Adaptation for Hate Speech Detection Using a Data Augmentation Approach,” *arXiv:2107.12866*, Jul. 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2107.12866

[18] W. Yin and A. Zubiaga, “Towards generalisable hate speech detection: a review on obstacles and solutions,” *PeerJ Computer Science*, vol. 7, p. e598, 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2102.08886

[19] F. Baider and M. 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*, vol. 8, no. 2, pp. 262–287, 2020. [Online]. Available: https://doi.org/10.1075/jlac.00040.bai

[20] A. Jha and R. 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*, 2017, pp. 7–16. [Online]. Available: https://doi.org/10.18653/v1/W17-2902

[21] R. Kumar, A. K. Ojha, S. Malmasi, and M. Zampieri, "Benchmarking aggression identification in social media," in *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*, 2018, pp. 1–11. [Online]. Available: https://aclanthology.org/W18-4401

[22] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, and V. Patti, "Resources and benchmark corpora for hate speech detection: a systematic review," *Lang. Resour. Evaluation*, vol. 55, pp. 477–523, Jun. 2021. [Online]. Available: https://doi.org/10.1007/s10579-020-09502-8

[23] M. Wiegand, J. Ruppenhofer, and E. 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*. Association for Computational Linguistics, Jun. 2021, pp. 576–587. [Online]. Available: https://doi.org/10.18653/v1/2021.naacl-main.48

[24] J. H. Park, J. Shin, and P. Fung, "Reducing gender bias in abusive language detection," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pp. 2799–2804, Nov. 2018. [Online]. Available: https://doi.org/10.18653/v1/d18-1302

[25] G. Kovács, P. Alonso, and R. Saini, "Challenges of hate speech detection in social media," *SN Computer Science*, vol. 2, no. 2, pp. 1–15, 2021. [Online]. Available: https://doi.org/10.1007/s42979-021-00457-3

[26] P. Felt, E. Ringger, J. Boyd-Graber, and K. Seppi, "Making the Most of Crowdsourced Document Annotations: Confused Supervised LDA," in *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Jul. 2015, pp. 194–203. [Online]. Available: https://doi.org/10.18653/v1/K15-1020

[27] M. Sanguinetti, F. Poletto, C. Bosco, V. Patti, and M. Stranisci, "An Italian Twitter corpus of hate speech against immigrants," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. [Online]. Available: https://aclanthology.org/L18-1443

[28] H. Al Kuwatly, M. Wich, and G. Groh, "Identifying and Measuring Annotator Bias Based on Annotators' Demographic Characteristics," in *Proceedings of the Fourth Workshop on Online*

Abuse and Harms. Association for Computational Linguistics, Nov. 2020, pp. 184–190. [Online]. Available: https://doi.org/10.18653/v1/2020.alw-1.21

[29] Z. 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*. Association for Computational Linguistics, Nov. 2016, pp. 138–142. [Online]. Available: https://doi.org/10.18653/v1/w16-5618

[30] G. Rizos, K. Hemker, and B. Schuller, "Augment to prevent: short-text data augmentation in deep learning for hate-speech classification," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. ACM, Nov. 2019, pp. 991–1000. [Online]. Available: https://doi.org/10.1145/3357384.3358040

[31] L. I. Venturott and P. M. Ciarelli, "Data Augmentation for improving Hate Speech Detection on Social Networks," in *Proceedings of the Brazilian Symposium on Multimedia and the Web*. ACM, Nov. 2020, pp. 249–252. [Online]. Available: https://doi.org/10.1145/3428658.3431760

[32] P. Shayegh, Y. Li, J. Zhang, and Q. Zhang, "Semi-supervised text classification with deep convolutional neural network using feature fusion approach," *Proceedings - 2019 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2019*, pp. 363–366, Oct. 2019. [Online]. Available: https://doi.org/10.1145/3350546.3352548

[33] R. de Pelle and V. Moreira, "Offensive comments in the brazilian web: a dataset and baseline results," in *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*, SBC. SBC, Jul. 2017. [Online]. Available: https://doi.org/10.5753/brasnam.2017.3260

[34] J. A. Leite, D. F. Silva, K. Bontcheva, and C. Scarton, "Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis," *arXiv:2010.04543*, Oct. 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2010.04543

[35] F. A. Vargas, I. Carvalho, F. R. de Góes, F. Benevenuto, and T. A. S. Pardo, "Building an Expert Annotated Corpus of Brazilian Instagram Comments for Hate Speech and Offensive Language Detection," *arXiv:2010.04543*, Mar. 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2010.04543

[36] C. J. Kennedy, G. Bacon, A. Sahn, and C. von Vacano, "Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application," 2020. [Online]. Available: https://doi.org/10.48550/ARXIV.2009.10277

[37] M. Samory, I. Sen, J. Kohne, F. Flöck, and C. Wagner, ""Call me sexist, but...": Revisiting Sexism Detection Using Psychological Scales and Adversarial Samples." in *ICWSM*, 2021, pp. 573–584. [Online]. Available: https://doi.org/10.48550/arXiv.2004.12764

[38] M. ElSherief, S. Nilizadeh, D. Nguyen, G. Vigna, and E. Belding, "Peer to peer hate: Hate speech instigators and their targets," *Proceedings of the International AAAI Conference on Web and Social Media*, 2018. [Online]. Available: https://doi.org/10.48550/ARXIV.1804.04649

[39] I. Mollas, Z. Chrysopoulou, S. Karlos, and G. Tsoumakas, "ETHOS: a multi-label hate speech detection dataset," *Complex & Intelligent Systems*, jan 2022. [Online]. Available: https://doi.org/10.1007/s40747-021-00608-2

[40] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*, vol. 11, no. 1, pp. 512–515, May 2017. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14955

[41] L. Gao and R. Huang, "Detecting Online Hate Speech Using Context Aware Models," 2017. [Online]. Available: https://doi.org/10.48550/ARXIV.1710.07395

[42] P. Röttger, B. Vidgen, D. Nguyen, Z. Waseem, H. Margetts, and J. Pierrehumbert, "HateCheck: Functional tests for hate speech detection models," 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)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 41–58. [Online]. Available: https://doi.org/10.18653/v1/2021.acl-long.4

[43] S. Alsafari and S. Sadaoui, "Semi-Supervised Self-Training of Hate and Offensive Speech from Social Media," *Applied Artificial Intelligence*, pp. 1–25, Oct. 2021. [Online]. Available: https://doi.org/10.1080/08839514.2021.1988443

[44] J. Li, Q. Zhu, Q. Wu, and D. Cheng, "An effective framework based on local cores for self-labeled semi-supervised classification," *Knowledge-Based Systems*, vol. 197, Jun. 2020. [Online]. Available: https://doi.org/10.1016/j.knosys.2020.105804

[45] C. Li, X. Li, and J. 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)*. Association for Computational Linguistics, Aug. 2021, pp. 5044–5053. [Online]. Available: https://doi.org/10.18653/v1/2021.acl-long.391

[46] Z. Xu, “Semantic Space-Based Self-Training for Semi-Supervised Multi-label Text Classification,” *DEIM Forum 2021*, 2021. [Online]. Available: https://proceedings-of-deim.github.io/DEIM2021/papers/E24-2.pdf

[47] S. Rosenthal, P. Atanasova, G. Karadzhov, M. Zampieri, and P. Nakov, “SOLID: A Large-Scale Semi-Supervised Dataset for Offensive Language Identification,” *arXiv:2004.14454*, no. July, pp. 915–928, 2021. [Online]. Available: https://doi.org/10.18653/v1/2021.findings-acl.80

[48] J. Chen, J. Feng, X. Sun, and Y. Liu, “Co-training semi-supervised deep learning for sentiment classification of MOOC forum posts,” *Symmetry*, vol. 12, no. 1, pp. 1–24, Dec. 2020. [Online]. Available: https://doi.org/10.3390/SYM12010008

[49] P. Karisani and N. Karisani, “Semi-supervised text classification via self-pretraining,” in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. ACM, Mar. 2021, pp. 40–48. [Online]. Available: https://doi.org/10.1145/3437963.3441814

[50] K. P. Bennett, A. Demiriz, and R. Maclin, “Exploiting unlabeled data in ensemble methods,” in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM Press, Jul. 2002, pp. 289–296. [Online]. Available: https://doi.org/10.1145/775047.775090

[51] J. Tanha, M. Van Someren, and H. Afsarmanesh, “An AdaBoost algorithm for multiclass semi-supervised learning,” *Proceedings - IEEE International Conference on Data Mining, ICDM*, pp. 1116–1121, Dec. 2012. [Online]. Available: https://doi.org/10.1109/ICDM.2012.119

[52] M. Zareapoor and K. R. Seeja, “Feature Extraction or Feature Selection for Text Classification: A Case Study on Phishing Email Detection,” *International Journal of Information Engineering and Electronic Business*, vol. 7, no. 2, pp. 60–65, Mar. 2015. [Online]. Available: https://doi.org/10.5815/ijieeb.2015.02.08

[53] K. Lee, A. Qadir, S. A. Hasan, V. Datla, A. Prakash, J. Liu, and O. Farri, “Adverse drug event detection in tweets with semi-supervised convolutional neural networks,” in *Proceedings of the 26th International Conference on World Wide Web - WWW ’17*. ACM Press, Apr. 2017, pp. 705–714. [Online]. Available: https://doi.org/10.1145/3038912.3052671

[54] W. Zhang, X. Tang, and T. Yoshida, “TESC: An approach to TExt classification using Semi-supervised Clustering,” *Knowledge-Based Systems*, vol. 75, pp. 152–160, Feb. 2015. [Online]. Available: https://doi.org/10.1016/j.knosys.2014.11.028

[55] Z. Sun, C. Fan, X. Sun, Y. Meng, F. Wu, and J. Li, “Neural Semi-supervised Learning for Text Classification Under Large-Scale Pretraining,” *arXiv:2011.08626*, Nov. 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2011.08626

[56] H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier, and D. Schwab, “FlauBERT: Unsupervised language model pre-training for French,” *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, pp. 2479–2490, Mar. 2020. [Online]. Available: https://doi.org/10.48550/arXiv.1912.05372

[57] Z. Qi, Y. Tian, and Y. Shi, “Laplacian twin support vector machine for semi-supervised classification,” *Neural networks*, vol. 35, pp. 46–53, Nov. 2012. [Online]. Available: https://doi.org/10.1016/j.neunet.2012.07.011

[58] V. N. Vapnik, *Statistical learning theory*, ser. Adaptive and Cognitive Dynamic Systems: Signal Processing, Learning, Commun ications and Control. John Wiley & Sons, Sep. 1998.

[59] T. Joachims, “Transductive Inference for Text Classification Using Support Vector Machines,” in *Proceedings of the Sixteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., Jun. 1999, p. 200–209. [Online]. Available: https://dl.acm.org/doi/10.5555/645528.657646

[60] Y.-F. Li, J. T. Kwok, and Z.-H. Zhou, “Semi-supervised learning using label mean,” in *Proceedings of the 26th Annual International Conference on Machine Learning*. Association for Computing Machinery, Jun. 2009, pp. 633–640. [Online]. Available: https://doi.org/10.1145/1553374.1553456

[61] Y.-F. Li and Z.-H. Zhou, “Towards making unlabeled data never hurt,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 1, pp. 175–188, Jan. 2015. [Online]. Available: https://doi.org/10.1109/TPAMI.2014.2299812

[62] Y. Ouali, C. Hudelot, and M. Tami, “An Overview of Deep Semi-Supervised Learning,” *arXiv:2006.05278*, pp. 1–43, Jun. 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2006.05278

[63] C. Cardellino, L. A. Alemany, M. Teruel, S. Villata, and S. Marro, “Convolutional ladder networks for Legal NERC and the impact of unsupervised data in better generalizations,” in *The Thirty-Second International Flairs Conference*, May 2019. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02381093

[64] A. H. Li and A. Sethy, “Semi-Supervised Learning for Text Classification by Layer Partitioning,” *ICASSP, IEEE International Conference on Acoustics, Speech and Signal*

Processing - Proceedings, pp. 6164–6168, May 2020. [Online]. Available: https://doi.org/10.1109/ICASSP40776.2020.9053565

[65] R. Xiang and S. Yin, "Semi-supervised text classification with temporal ensembling," *2021 International Conference on Computer Communication and Artificial Intelligence, CCAI 2021*, pp. 204–208, May 2021. [Online]. Available: https://doi.org/10.1109/CCAI50917.2021.9447486

[66] T. Miyato, A. M. Dai, and I. Goodfellow, "Adversarial training methods for semi-supervised text classification," *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, pp. 1–11, May 2016. [Online]. Available: https://doi.org/10.48550/arXiv.1605.07725

[67] P. Meel and D. K. Vishwakarma, "A temporal ensembling based semi-supervised ConvNet for the detection of fake news articles," *Expert Systems with Applications*, vol. 177, Sep. 2021. [Online]. Available: https://doi.org/10.1016/j.eswa.2021.115002

[68] Z. Miao, Y. Li, X. Wang, and W.-C. Tan, "Snippext: Semi-supervised opinion mining with augmented data," in *Proceedings of The Web Conference 2020*, Feb. 2020, pp. 617–628. [Online]. Available: https://doi.org/10.48550/arXiv.2002.03049

[69] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel, "MixMatch: A holistic approach to semi-supervised learning," *arXiv:1905.02249*, May 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1905.02249

[70] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-supervised learning (Chapelle, O. et al., Eds.; 2006)[Book reviews]," *IEEE Transactions on Neural Networks*, vol. 20, no. 3, pp. 542–542, Feb. 2009. [Online]. Available: https://doi.org/10.1109/TNN.2009.2015974

[71] X. J. Zhu, "Semi-supervised learning literature survey," *University of Wisconsin-Madison - Department of Computer Sciences*, Jul. 2008. [Online]. Available: https://pages.cs.wisc.edu/~jerryzhu/pub/ssl_survey.pdf

[72] W. Xu, H. Sun, C. Deng, and Y. Tan, "Variational Autoencoders for Semi-supervised Text Classification," *arXiv:1603.02514*, pp. 3358–3364, Mar. 2016. [Online]. Available: https://doi.org/10.48550/arXiv.1603.02514

[73] M. Cheng, S. Nazarian, and P. Bogdan, "VRoC: Variational Autoencoder-aided Multi-task Rumor Classifier Based on Text," *The Web Conference 2020 - Proceedings of the World Wide Web Conference - WWW '20*, pp. 2892–2898, Jan. 2020. [Online]. Available: https://doi.org/10.1145/3366423.3380054

[74] J. Qian, M. ElSherief, E. Belding, and W. Y. Wang, "Hierarchical CVAE for fine-grained hate speech classification," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pp. 3550–3559, Nov. 2020. [Online]. Available: https://doi.org/10.18653/v1/d18-1391

[75] K. K. Thekumparampil, C. Wang, S. Oh, and L.-J. Li, "Attention-based Graph Neural Network for Semi-supervised Learning," *arXiv:1803.03735*, Mar. 2018. [Online]. Available: https://doi.org/10.48550/arXiv.1803.03735

[76] A. Benamira, B. Devillers, E. Lesot, A. K. Ray, M. Saadi, and F. D. Malliaros, "Semi-Supervised Learning and Graph Neural Networks for Fake News Detection," in *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. ACM, Aug. 2019, pp. 568–569. [Online]. Available: https://doi.org/10.1145/3341161.3342958

[77] J. Huang, N. Tao, H. Chen, Q. Deng, W. Wang, and J. Wang, "Semi-supervised Text Classification Based On Graph Attention Neural Networks," in *2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD)*, Jun. 2021, pp. 325–330. [Online]. Available: https://doi.org/10.1109/ICAIBD51990.2021.9459003

[78] P. Mishra, M. Del Tredici, H. Yannakoudakis, and E. Shutova, "Abusive language detection with graph convolutional networks," *arXiv:1904.04073*, Apr. 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1904.04073

[79] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv:1609.02907*, Sep. 2016. [Online]. Available: https://doi.org/10.48550/arXiv.1609.02907

[80] A. G. D'Sa, I. Illina, D. Fohr, D. Klakow, and D. Ruiter, "Label Propagation-Based Semi-Supervised Learning for Hate Speech Classification," in *Proceedings of the First Workshop on Insights from Negative Results in NLP*. Association for Computational Linguistics, Nov. 2020, pp. 54–59. [Online]. Available: https://doi.org/10.18653/v1/2020.insights-1.8

[81] M. Papadaki, "Data Augmentation Techniques for Legal Text Analytics," Master's thesis, Athens University of Economics and Business, Oct. 2017. [Online]. Available: http://nlp.cs.aueb.gr/theses.html

[82] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, vol. 6, no. 1, Jul. 2019. [Online]. Available: https://doi.org/10.1186/s40537-019-0197-0

[83] S. T. Luu, K. Van Nguyen, and N. L.-T. Nguyen, “Empirical Study of Text Augmentation on Social Media Text in Vietnamese,” *arXiv:2009.12319*, Sep. 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2009.12319

[84] M. Juuti, T. Gröndahl, A. Flanagan, and N. Asokan, “A little goes a long way: Improving toxic language classification despite data scarcity,” *arXiv:2009.12344*, pp. 2991–3009, Sep. 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.findings-emnlp.269

[85] A. V. Mosolova, V. V. Fomin, and I. Y. Bondarenko, “Text augmentation for neural networks,” in *Supplementary Proceedings of the Seventh International Conference on Analysis of Images, Social Networks and Texts (AIST 2018)*, vol. 2268. CEUR Workshop Proceedings, Jul. 2018, pp. 104–109. [Online]. Available: http://ceur-ws.org/Vol-2268/

[86] C. Rastogi, N. Mofid, and F.-I. Hsiao, “Can We Achieve More with Less? Exploring Data Augmentation for Toxic Comment Classification,” *arXiv:2007.00875*, Jul. 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2007.00875

[87] D. R. Beddiar, M. S. Jahan, and M. Oussalah, “Data expansion using back translation and paraphrasing for hate speech detection,” *Online Social Networks and Media*, vol. 24, p. 100153, 2021. [Online]. Available: https://doi.org/10.1016/j.osnem.2021.100153

[88] B. Li, Y. Hou, and W. Che, “Data augmentation approaches in natural language processing: A survey,” *AI Open*, vol. 3, pp. 71–90, 2022. [Online]. Available: https://doi.org/10.1016/j.aiopen.2022.03.001

[89] R. Liu, G. Xu, and S. Vosoughi, “Enhanced Offensive Language Detection Through Data Augmentation,” *arXiv:2012.02954*, Dec. 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2012.02954

[90] M. Z. Alksasbeh, B. A. Alqaralleh, T. Abukhalil, A. Abukaraki, T. Al Rawashdeh, and M. Al-Jaafreh, “Smart detection of offensive words in social media using the soundex algorithm and permuterm index,” *Int. J. Electr. Comput. Eng. (IJECE)*, vol. 11, no. 5, pp. 4431–4438, Oct. 2021. [Online]. Available: https://doi.org/10.11591/ijece.v11i5.pp4431-4438

[91] Y. Belinkov and Y. Bisk, “Synthetic and natural noise both break neural machine translation,” *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, pp. 1–13, Nov. 2017. [Online]. Available: https://doi.org/10.48550/arXiv.1711.02173

[92] Z. Yuan and Y. Wen, “A new semi-supervised inductive transfer learning framework: Co-Transfer,” *arXiv:2108.07930*, Aug. 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2108.07930

[93] M. Kang, A. K. Biswas, D.-c. Kim, and J. Gao, “Semi-supervised Discriminative Transfer Learning in Cross-language Text Classification,” in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. IEEE, Dec. 2019, pp. 1031–1038. [Online]. Available: https://doi.org/10.1109/ICMLA.2019.00174

[94] C. Abderrouaf and M. Oussalah, “On Online Hate Speech Detection. Effects of Negated Data Construction,” *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*, pp. 5595–5602, Dec. 2019. [Online]. Available: https://doi.org/10.1109/BigData47090.2019.9006336

[95] M.-A. Rizoiu, T. Wang, G. Ferraro, and H. Suominen, “Transfer Learning for Hate Speech Detection in Social Media,” *arXiv:1906.03829*, Jun. 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1906.03829

[96] L. Stappen, F. Brunn, and B. Schuller, “Cross-lingual Zero- and Few-shot Hate Speech Detection Utilising Frozen Transformer Language Models and AXEL,” *arXiv:2004.13850*, Apr. 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2004.13850

[97] M. Mozafari, R. Farahbakhsh, and N. Crespi, “A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media,” *Studies in Computational Intelligence*, vol. 881 SCI, pp. 928–940, Oct. 2020. [Online]. Available: https://doi.org/10.1007/978-3-030-36687-2_77

[98] R. Gupta, S. Sahu, C. Espy-wilson, and S. Narayanan, “Semi-supervised and transfer learning approaches for low resource sentiment classification,” *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5109–5113, Jun. 2018. [Online]. Available: https://doi.org/10.48550/arXiv.1806.02863

[99] M. A. Bashar, R. Nayak, K. Luong, and T. Balasubramaniam, “Progressive domain adaptation for detecting hate speech on social media with small training set and its application to COVID-19 concerned posts,” *Social Network Analysis and Mining*, vol. 11, no. 1, pp. 1–18, Jul. 2021. [Online]. Available: https://doi.org/10.1007/s13278-021-00780-w

[100] J. Schäfer and B. Burtenshaw, “Offence in dialogues: A corpus-based study,” in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*. INCOMA Ltd., Sep. 2019, pp. 1085–1093. [Online]. Available: https://doi.org/10.26615/978-954-452-056-4_125

[101] R. Kumar, A. N. Reganti, A. Bhatia, and T. Maheshwari, “Aggression-annotated corpus of Hindi-English code-mixed data,” *arXiv:1803.09402*, Mar. 2018. [Online]. Available: https://doi.org/10.48550/arXiv.1803.09402

[102] T. Fitzsimons, “Nearly 1 in 5 hate crimes motivated by anti-LGBTQ bias, FBI finds,” *NBC News*, Nov 2019. [Online]. Available: https://www.nbcnews.com/feature/nbc-out/nearly-1-5-hate-crimes-motivated-anti-lgbtq-bias-fbi-n1080891

[103] H. Park and I. Lyshyn, “L.G.B.T. people are more likely to be targets of hate crimes than any other minority group,” *The New York Times*, Jun 2016. [Online]. Available: https://www.nytimes.com/interactive/2016/06/16/us/hate-crimes-against-lgbt.html

[104] W. Ronan, “New FBI Hate Crimes Report Shows Increases in Anti-LGBTQ Attacks,” *Human Rights Campaign*, Nov 2020. [Online]. Available: https://www.hrc.org/press-releases/new-fbi-hate-crimes-report-shows-increases-in-anti-lgbtq-attacks

[105] S. Benesch, D. Ruths, K. P. Dillon, H. M. Saleem, and L. Wright, “Counterspeech on twitter: A field study. dangerous speech project,” 2016. [Online]. Available: https://dangerousspeech.org/counterspeech-on-twitter-a-field-study/

[106] P. Carvalho, B. Matos, R. Santos, F. Batista, and R. Ribeiro, “Hate Speech Dynamics Against African descent, Roma and LGBTQI Communities in Portugal,” *LREC*, 2022. [Online]. Available: https://aclanthology.org/2022.lrec-1.253

[107] K. Shah, H. Patel, D. Sanghvi, and M. Shah, “A comparative analysis of logistic regression, random forest and knn models for the text classification,” *Augmented Human Research*, vol. 5, no. 1, pp. 1–16, 2020. [Online]. Available: https://doi.org/10.1007/s41133-020-00032-0

[108] C. Breazzano, D. Croce, and R. 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, Nov. 2021. [Online]. Available: http://ceur-ws.org/Vol-3015/

[109] R. Ayari, “NLP: Word Embedding Techniques Demystified,” 2020. [Online]. Available: https://towardsdatascience.com/nlp-embedding-techniques-51b7e6ec9f92

[110] F. Souza, R. Nogueira, and R. Lotufo, “BERTimbau: Pretrained BERT models for Brazilian Portuguese,” in *Brazilian Conference on Intelligent Systems*. Springer, 2020, pp. 403–417. [Online]. Available: https://doi.org/10.1007/978-3-030-61377-8_28

[111] T. Caselli, V. Basile, J. Mitrović, and M. Granitzer, “HateBERT: Retraining BERT for Abusive Language Detection in English,” *arXiv preprint arXiv:2010.12472*, 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2010.12472

[112] 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*, vol. 12, pp. 2825–2830, 2011. [Online]. Available: https://scikit-learn.org/

[113] M. Kazijevs, F. A. Akyelken, and M. D. Samad, “Mining social media data to predict covid-19 case counts,” in *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*. IEEE, 2022, pp. 104–111. [Online]. Available: https://doi.org/10.1109/ICHI54592.2022.00027

[114] L. Ajallouda, K. Najmani, A. Zellou *et al.*, “Doc2vec, sbert, infersent, and use which embedding technique for noun phrases?” in *2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*. IEEE, 2022, pp. 1–5. [Online]. Available: https://doi.org/10.1109/IRASET52964.2022.9738300

[115] S. Modha, P. Majumder, and T. Mandl, “An empirical evaluation of text representation schemes to filter the social media stream,” *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 34, no. 3, pp. 499–525, 2022. [Online]. Available: https://doi.org/10.1080/0952813X.2021.1907792

[116] M. Miliani, G. Giorgi, I. Rama, G. Anselmi, and G. E. Lebani, “DANKMEMES@ EVALITA 2020: The Memeing of Life: Memes, Multimodality and Politics,” in *EVALITA*, 2020. [Online]. Available: http://ceur-ws.org/Vol-2765/paper174.pdf

[117] A.-M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis, “Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior,” in *Twelfth International AAAI Conference on Web and Social Media*. arXiv, 2018. [Online]. Available: https://doi.org/10.48550/arXiv.1802.00393

[118] D. Santos and A. Simões, “Portuguese-English word alignment: some experiments,” *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*, May 2008. [Online]. Available: https://hdl.handle.net/1822/14511