IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 1 July 2024, accepted 11 July 2024, date of publication 18 July 2024, date of current version 30 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3430848

RESEARCH ARTICLE

Leveraging Transfer Learning for Hate Speech Detection in Portuguese Social Media Posts

GIL RAMOS[1], FERNANDO BATISTA[2,3], (Senior Member, IEEE), RICARDO RIBEIRO[2,3], PEDRO FIALHO[1], SÉRGIO MORO[1,4], ANTÓNIO FONSECA[1], RITA GUERRA[5], PAULA CARVALHO[3], CATARINA MARQUES[6], AND CLÁUDIA SILVA[7,8]

[1]Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, 1649-026 Lisbon, Portugal
[2]Instituto Universitário de Lisboa (ISCTE-IUL), 1649-026 Lisbon, Portugal
[3]INESC-ID, 1000-029 Lisbon, Portugal
[4]The University of Jordan, Amman 11941, Jordan
[5]ISCTE—Instituto Universitário de Lisboa and Center for Psychological Research and Social Intervention (CIS-ISCTE), 1000-029 Lisbon, Portugal
[6]ISCTE—Instituto Universitário de Lisboa and Business Research Unit (BRU-ISCTE), 1649-026 Lisbon, Portugal
[7]Interactive Technologies Institute-Laboratory of Robotics and Engineering Systems (ITI-LARSyS), 1900-319 Lisbon, Portugal
[8]Instituto Superior Tcnico (IST), 1049-001 Lisbon, Portugal

Corresponding author: Gil Ramos (gasnr@iscte-iul.pt)

This work was supported in part by the European Union under Grant CERV-2021-EQUAL (101049306). However the views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or Knowhate Project. Neither the European Union nor the Knowhate Project can be held responsible.

ABSTRACT The rapid rise of social media has brought about new ways of digital communication, along with a worrying increase in online hate speech (HS), which, in turn, has led researchers to develop several Natural Language Processing methods for its detection. Although significant strides have been made in automating HS detection, research focusing on the European Portuguese language remains scarce (as it happens in several under-resourced languages). To address this gap, we explore the efficacy of various transfer learning models, which have been shown in the literature to have better performance for this task than other Deep Learning models. We employ BERT-like models pre-trained on Portuguese text, such as BERTimbau and mDeBERTa, as well as GPT, Gemini and Mistral generative models, for the detection of HS within Portuguese online discourse. Our study relies on two annotated corpora of YouTube comments and tweets, both annotated as HS and non-HS. Our findings show that the best model for the YouTube corpus was a variant of BERTimbau retrained with European Portuguese tweets and fine-tuned for the HS task, with an F-score of 87.1% for the positive class, outperforming the baseline models by more than 20% and with a 1.8% increase compared with base BERTimbau. The best model for the Twitter corpus was GPT-3.5, with an F-score of 50.2% for the positive class. We also assess the impact of using in-domain and mixed-domain training sets, as well as the impact of providing context in generative model prompts on their performance.

INDEX TERMS Hate speech, transfer learning, transformer models, generative models, text classification.

I. INTRODUCTION

In contemporary times, the explosion of social media engagement has revolutionized the digital communication sphere, fundamentally reshaping the dynamics of self-expression and interpersonal connections [1]. Thanks to the universal presence of smartphones and internet connectivity, social media platforms are now easily accessible, enabling individuals worldwide to disseminate their thoughts and perspectives effortlessly. This democratization of expression, while fostering empowerment and meaningful dialogue, has also brought into focus a significant concern: the rapid proliferation of Hate Speech (HS) and associated transgressions [2].

There are no universally accepted and precise definitions of HS [3], however, for the purpose of this work, online HS refers to bias-motivated, derogatory language that spreads, incites, promotes, or justifies hatred, exclusion, and/or violence/aggression against a person/group because of their group membership, and can be operationalized through the following coexisting conditions [4]:

The associate editor coordinating the review of this manuscript and approving it for publication was Maria Chiara Caschera.

- HS is an intergroup phenomenon that targets social groups or individuals because of their perceived membership in certain social groups. Therefore, in this project, HS does not encompass hateful expressions that occur at the individual/interindividual level of analysis.
- The target of HS is attacked primarily because of perceived membership in a given social group, and not necessarily because of a specific behavior or action.
- HS can be expressed both directly (explicitly or overtly) and indirectly (implicitly or covertly). In the last case, the implicit meaning must be inferred.

Such discourse often precipitates psychological and emotional distress in recipients, potentially culminating in stress, anxiety, and even depression [5]. Moreover, prolonged exposure to HS can corrode societal bonds, fostering an environment with mistrust and polarization. This fragmentation exacerbates the cycle of animosity, as individuals retreat into their respective echo chambers, reinforcing preexisting biases and prejudices [6].

Numerous organizations have addressed the issue of HS on social media by implementing guidelines and policies. However, the massive volume of data generated by these platforms makes manual classification impractical. Therefore, there is a growing need to utilize Machine Learning (ML) techniques to automate the classification process, resulting in more efficient and reliable outcomes [7]. This technological shift has led to much research and development aimed at leveraging ML for HS detection, with various techniques being applied, ranging from approaches like classical ML to Ensemble Models and Deep Learning (DL) models, with promising results. With the development of Transformer-based models, such as BERT [8], there has been a paradigm shift for most NLP tasks, leading to a growing expansion in the HS detection landscape.

Despite the recommendations of the Commissioner for Human Rights of the Council of Europe for Portuguese authorities to address the increasing level of racism and HS, in response to the increase in racially motivated hate crimes and HS [9], there are still few studies that have focused on analyzing and detecting European Portuguese HS [10].

In this study, we used two different corpora created in the scope of the project *kNOwHATE: kNOwing online HATE speech*,[1] one containing YouTube comments [11], and the other consisting of tweets retrieved from Twitter (now X) [12], to apply transfer learning approaches based on several Transformer-based models pre-trained for the task of HS detection. In particular, BERTimbau-hatebr [13], mDeBERTa-hatebr [13], HATEBERTimbau [14], and the general model that that is the base of some of these pre-trained models, BERTimbau [15]. We compared the results of these models in Portuguese YouTube comments and tweets and also experimented with mixed-domain learning by training the models with data from both sources to see if it leads to a better generalization and, in turn, a better performance. In addition, we also utilized generative models such as GPT [16], Gemini [17] and Mistral [18], with and without context provided, to compare the results with the Transformer-based models.

An important point to clarify when dealing with social media data is that there is an inherent filtering process applied by the platforms, where content and users are frequently removed [19]. Despite this, datasets retrieved from social media remain valuable for HS detection tasks since a significant portion of HS is covert [20] and often escapes these filters. In addition, our datasets include messages that were later removed, ensuring that they encompass a broad range of HS, including content that was removed by the platforms' filtering mechanisms.

The remainder of this paper is organized as follows. Section II provides an overview of the relevant research literature; Section III describes the datasets used and their properties; Section IV describes all the models that were used and the setup used to run them; Section V presents the results obtained; Section VI elaborates on the results with a discussion; Section VII provides some insight on the possible causes of errors in the models; and Section VIII provides the main conclusions and provides guidance for future work.

II. RELATED WORK

The literature on HS detection is gaining increasing relevance, with the number of publications related to this subject growing annually [10]. In this section, a synopsis of the existing literature is given, classified into five categories: classical ML models, DL models, Transformer-based models, generative models, and literature on HS detection in Portuguese.

A. CLASSICAL ML MODELS

Watanabe et al. [2] used a j48graft classifier, a type of Decision Tree model, in combination with sentiment, semantic, unigram and pattern features to detect offensive speech. They concluded that the inclusion of all features improved the performance of the model. Pitropakis et al. [21] used a combination of a Support Vector Machine (SVM) with word n-grams, character n-grams, and Term Frequency-Inverse Document Frequency (TF-IDF), with word n-grams having the best performance. Saeed et al. [22] also used a combination of an SVM with character and word n-grams to obtain the best results with word n-grams. Mohapatra et al. [23] used an SVM with Word2vec features, outperforming word n-grams and TF-IDF. Shannaq et al. [24] used fine-tuned AraVec SkipGram n-grams with a SVM model outperforming other ML models. Arcila-Calderón et al. [25] used Bag-of-Words (BOW) features with an Logistic Regression (LR) model and Turki and Roy [26] used a Random Forest model with Count Vectorization as features to surpass the Bagging and Adaboost models.

[1] https://knowhate.eu/

B. DEEP LEARNING MODELS

Classical machine learning models exhibit promising results, however, they depend heavily on feature engineering for optimization, a process known to be time-consuming and reliant on human intervention. In contrast, the emergence of DL models has prompted researchers to increasingly rely on them to circumvent the aforementioned limitations.

Convolutional Neural Networks (CNNs) and Long Short Term Memories (LSTMs) are two of the most widely used DL architectures for HS detection. CNNs can capture local patterns and features in text, whereas LSTMs are adept at handling long-range dependencies. The results of using CNNs and LSTMs for HS detection are somewhat mixed, with some studies showing that CNNs outperform LSTMs [27], [28], while others have found the opposite [29], [30], [31]. For instance Karayiğit et al. [32] used Continuous Bag-of-Words (CBOW) features in conjunction with a CNN model, while Priyadarshini et al. [33] combined pre-trained GloVe embeddings with a LSTM model.

Despite these mixed results, hybrid models that combine these two types of models consistently show strong performance. These models leverage the strengths of each architecture, leading to improved results and generalizability. For example, Fazil et al. [34] employed a CNN-Bidirectional LSTM (CNN-BiLSTM) model with several filters and different kernel sizes to capture semantics relations at various windows. The encoded representation from these multiple channels passed through an attention-aware stacked 2-layer BiLSTM network, and the output was then weighted by an attention layer and further concatenated and passed via a dense layer and an output layer to classify the text.

C. TRANSFORMER-BASED MODELS

In addition to the previously discussed DL models, the literature on HS detection has seen a notable surge of interest in Transformer-based models. These models utilize a self-attention mechanism to efficiently capture contextual dependencies in input sequences [35]. Unlike traditional DL models, Transformers do not rely on recurrent connections, making them capable of effectively handling longer sequences. For HS detection, researchers utilize pre-trained Transformer models that have been trained on large corpora of text data, such as BERT, which has already learned rich representations of language through unsupervised learning on massive text datasets. By leveraging pre-trained Transformer models and fine-tuning them for HS detection, researchers benefit from the vast amount of knowledge these models have already acquired from their pre-training phase. This approach allows the model to effectively capture the complex linguistic patterns and contextual cues indicative of HS, even when working with limited labeled data, which makes it optimal for this task, where there is often a lack of resources. As a result, Transformer-based models have become increasingly popular for HS detection tasks, with many studies employing these models and outperforming all other models from classical ML to DL [36], [37], [38], [39], [40], [41]. Given the better text representations of these models, researchers use them in addition to other classification models to tackle their classification problems. For instance [42] and [43] used BART as sentence and document representations, respectively, with DL classifiers, for fake news detection. Transformer models can also be used on their own and still achieve state-of-the-art performances [36], [38]. And authors also take advantage of these models by combining the text representations that they provide with other features, like tweet metadata [41], contextual information [37], and other handcrafted features like lexicons and emoticons [40].

D. GENERATIVE MODELS

More recently generative large language models, such as GPT, have also been solutions that have been used by researchers for HS detection. Chiu et al. [44] employed GPT-3 for the detection of sexist and racist comments on YouTube and Reddit through zero-, one-, and few-shot where example sexist/racist comments were given as context within the prompt to help the model classify the target comments. The few-shot method proved to lead to better performance. Oliveira et al. [45] performed zero-shot attempts in Brazilian Portuguese and came to the conclusion that GPT is a feasible alternative for HS detection.

E. PORTUGUESE LANGUAGE

For Portuguese, the literature on HS detection is relatively limited, with a small number of studies focusing specifically on this language. In general, the existing work has primarily focused on Brazilian Portuguese, with few addressing European Portuguese.

For European Portuguese, initial work focused primarily on constructing a hierarchically labeled dataset for HS detection, but the authors also describe the development of an initial baseline classification for the dataset, Using pre-trained word embeddings and an LSTM, they achieved a 71% micro F-score [46]. More recent studies, focused solely on the task of detecting HS, are based on BERT. Santos et al. [47] combine a Generative Adversarial Network (GAN) and a BERT-based model to obtain a 66.4% positive class F-score. Matos et al. [48] use a BERT-CNN architecture for the classification, and managed to obtain 72.1% F-score on the positive class, by also considering the annotations that had a higher inter-annotator agreement (IAA) score between them. Both of these studies used newly developed annotated datasets for European Portuguese with HS instances from YouTube and Twitter [19], [49]. Other recent works have addressed the European Portuguese variant by developing tools that can be used for the HS detection task, like foundation encoder models to expand the still very scarce ecosystem of large language models specifically developed for this language, such as the Albertina family of models [50].

Although there are several studies on HS for Brazilian Portuguese [51], [52], [53], this fact does not discard the need for further development in European Portuguese, since research has identified several differences between European and Brazilian Portuguese. First, variations exist in both contemporary language and technical vocabulary, as demonstrated by the differences in mood distribution. Brazilian Portuguese tends to prioritize the truth-value of a proposition, whereas European Portuguese are characterized by a more neutral tone [54]. Additionally, there are distinctions in the lexical, lexical-syntactic, and morpho-syntactic usages of temporal adverbials [55]. More importantly, besides these technical differences, because HS is intrinsically dependent on both the target communities and social practice (i.e., the social and historical context), existing resources and models cannot be directly transferable or easily adapted to other linguistic and pragmatic contexts [20], [56], [57]. Therefore, in this case, models developed for Brazilian Portuguese are dependent on the context of the population that uses this variation of the language and are not suited for a different social and historical context like the European Portuguese one.

This review of the existing literature on HS detection in Portuguese has revealed a relevant gap in the detection of European Portuguese HS. Current research in this domain remains in its initial stages. Consequently, there is a need for further research aimed at addressing this gap, which could also provide valuable insights and methodologies applicable to other low-resourced languages. By developing tools for a language that, as it does not yet have many resources, may be more vulnerable to the risk of discrimination and abuse, we have the potential to make a significant difference in mitigating the harmful effects of HS on the Portuguese-speaking community. It is worth mentioning that no literature was found on HS, or general encoder models, for the Portuguese variants of Portuguese-speaking African countries (PALOP countries), as expected. These variants are even more underrepresented in the literature and should be covered in future work.

The results of the review of the current literature indicate that Transformer-based models are the approach that leads to better performance, and with the current limitation in the European Portuguese HS detection literature, our main contributions in this work are as follows:

- Investigating the effectiveness of Transformer-based models for HS detection in the context of the Portuguese language and comparing domain-specific models with more general ones;
- Exploring the potential of generative models such as GPT, Gemini and Mistral, with and without the addition of context, and comparing them with BERT models;
- Filling a critical gap in the HS detection literature by focusing specifically on the European Portuguese language, addressing the lack of comprehensive research in this area.

III. DATA

This study uses two corpora containing annotated online HS messages, recently created in the scope of kNOwHATE project: a YouTube corpus consisting of YouTube comments and a Twitter (now X) corpus containing tweets retrieved from Twitter. Table 1 presents both corpora distributions. Note that the test sets were also used to calculate inter-annotator agreement.

TABLE 1. Corpora distributions.

	#messages		HS proportion	
Corpus	Train	Test	Train	Test
YouTube	23912	825	64.90%	72.06%
Twitter	21546	805	11.48%	20.62%

The YouTube corpus consists of 23912 comments collected from 88 distinct YouTube videos, covering topics and events targeting, directly or indirectly, four specific target groups: African descent, Roma, Migrants, and the LGBTQ+ communities. Initially, videos containing HS messages were selected. To broaden this selection, additional videos featured in the related section were also included, as illustrated in Fig. 1. This decision was based on the hypothesis that frequently suggested videos when watching an already HS-flagged video were more likely to attract HS. To quantify the frequency of video suggestions and to identify those most likely to contain HS, a sorted list of suggested videos was generated. Videos that appeared more than 85 times on the list were added to the dataset. After obtaining the final group of videos that were potential candidates, videos were removed from selection if they did not have a minimum number of 1000 views and 100 comments, resulting in the final 88 videos, distributed by target group as follows: Roma – 16, migrants – 19, LGBTQ+ – 24, and African descent – 29.

The Twitter corpus consists of 21546 tweets retrieved using the Twitter API published between January 1, 2021, and December 31, 2022. For the collection of relevant tweets, a list of 259 keywords associated with the four specific target groups (African descent, Roma, migrants, and the LGBTQ+ communities) was compiled, and tweets containing those keywords were collected. From the collected tweets, only those written in Portuguese were selected, resulting in a dataset predominantly consisting of Brazilian Portuguese. Therefore, to ensure geographical relevance, the tweets were further narrowed to tweets only posted in Portugal. Additionally, the entire conversation to which the tweets belonged was also retrieved, ensuring that the parent tweet of all conversations was published in Portugal. In Table 2, we display some examples of messages with HS for the different target groups for both corpora.

The corpora were manually annotated by interdisciplinary teams, consisting of four researchers for the YouTube corpus and three researchers for the Twitter corpus, all with backgrounds in language sciences and social psychology. Each annotator was tasked with annotating approximately 6000 comments on YouTube and 7000 tweets on Twitter. Additionally, a subset of comments/tweets (825 for YouTube and 805 for Twitter) was assigned to all annotators to assess IAA and annotation reliability using Krippendorff's

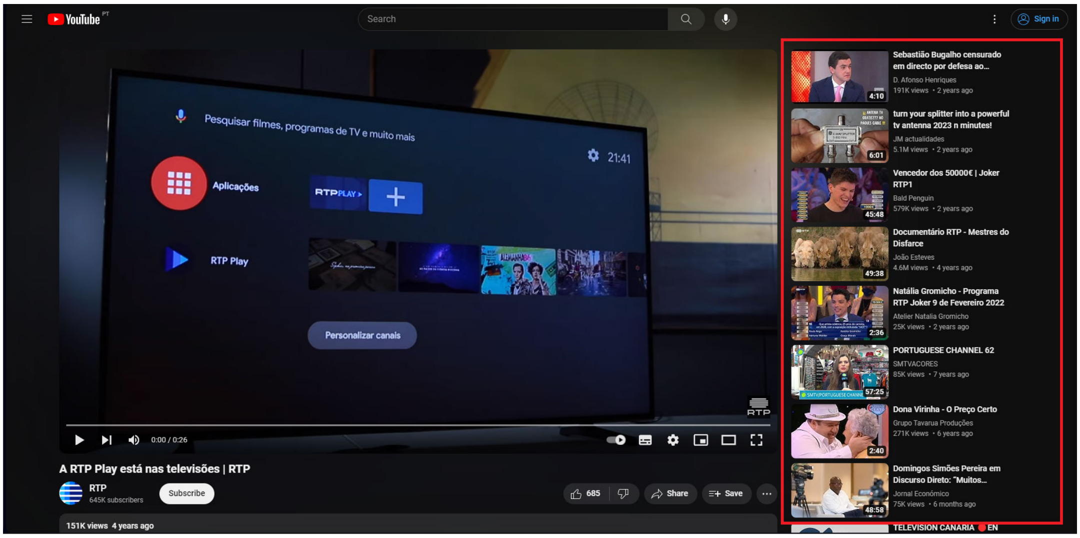

FIGURE 1. Related section shown in red.

TABLE 2. Hate Speech examples of both corpora for the different target groups.

Corpus	Target Group	Example Message
YouTube	Migrants	Isso pulhiticos merdosos, continuem a importar lixo, até Portugal deixar de ser Portugal. [*That's right shitty politicians, keep importing rubbish until Portugal stops being Portugal.*]
	Roma	Mais um bairro de ciganos onde eles é que fazem a lei.. Se o nosso belo governo lhes continuar a dar casas e dinheiro eles continuam a procriar e a encher bairros, onde depois o próprio governo não tem mão.. [*Another gypsy neighborhood where they make the law... If our beautiful government continues to give them houses and money they will continue to procreate and fill neighborhoods, where the government itself has no hand...*]
	LGBTQ+	as pessoas tem que perceber que ser "panasca" não é deixar de ser homem, é deixar de ser humano kek [*People have to realize that being "panasca"*[a] *doesn't mean stopping being a man, it means stopping being human kek*]
	African descent	Ao menos os branco de raça superior ainda criam biologia, já os pretos americanos nem sabem definir o que é biologicamente uma mulher.... Este mundo está perdido... alguém me sabe dizer se já aceitam voluntários para a missão a Marte? [*At least superior white people still create biology, while black Americans don't even know how to define what a woman is biologically.... This world is lost... can anyone tell me if they are already accepting volunteers for the mission to Mars?*]
Twitter	Migrants	Os zucas podem ofender todos os portugueses mas se a gente riposta já somos uns filhos da puta Nós tamos é cansados de ser chacota desta escória de pessoas. Não são todos, aliás, a maioria são gente boa MAS tem muito cabrão aí [*The "zucas"*[b] *may offend all Portuguese people, but if we fight back, we're already sons of bitches. We're tired of being made fun of by these scum of people. Not all of them, in fact, most of them are good people BUT there are a lot of bastards out there*]
	Roma	Ta tanto cigano no loureshopping foda se qual deles é q foi a julgamento [*There are so many gypsies in Loures Shopping, fuck which one went to trial*]
	LGBTQ+	Vai pá puta que te pariu seu paneleiro do caralho, virgem ofendida [*Go fuck yourself you fucking faggot, offended virgin.*]
	African descent	@User Se calhar são os que cometem mais ilegalidades não??!? Esta questão do coitadinho que é preto já passou de moda. Resistiu às autoridades teve o que se encontra previsto na lei, sem pôr nem tirar! [*Maybe they are the ones who commit the most illegalities, right??!? This question of the poor thing being black is out of fashion. He resisted the authorities and did what is stipulated by law, without putting in or taking away!*]

[a] Derogatory term used to refer to homosexual men; [b] Derogatory term used to refer to people from Brazil

alpha [58]. The IAA for YouTube was moderate at 0.546, whereas for Twitter was considerably lower at 0.355, indicating variations in agreement levels between the annotators across the two datasets. This IAA subset also served as the test set for model evaluation, and given the task subjectivity, only the messages that were labeled as conveying HS by at least two annotators were considered hatred content in the test sets. We did not consider the messages containing only one vote to discard unintentional errors introduced by the annotator, as the possibility that the majority of annotators making a mistake would be less likely.

IV. METHODOLOGY

In this section, we present an overview of the different models used for HS detection, along with the experimental settings used to run these models and the metrics used to evaluate the performance of each model. Fig. 2 presents the overall workflow of the experiments.

A. BASELINE

To serve as a baseline for comparison with the Transformer-based models, we employed a CNN model based on Safaya et al. [59] work with 160 convolutional filters of 5 different sizes (1, 2, 3, 4, and 5) and 32 filters for each size. We also employed an LSTM model with an initial layer comprising 128 units, followed by one dense layer with 64 units and an output layer with a softmax activation function. For both models, the embeddings used were FastText CBOW for Portuguese [60], with dimensions of 300.

B. TRANSFORMER-BASED MODELS

For the Transformer-based models, we used four different models based on BERT. The BERT base model contains an encoder with 12 layers (transformer blocks), 12 self-attention heads, and 110 million parameters. Because BERT-based models are pre-trained on large general corpora, they were fine-tuned using the domain-specific one, and there was a linear layer was added on top of the BERT architecture for the classification. For this, the [CLS] token output of the 12^{th} transformer encoder, a vector of size 768, is given as an input to a fully connected network. Subsequently, the sigmoid activation function was applied to the hidden layer to make the predictions. During training, some of the BERT weights were also updated, allowing the model to adapt to the specific characteristics of our dataset. Four different BERT-based models were used:

- BERTimbau – although developed for Brazilian Portuguese, since our work is focused on Portuguese we used BERTimbau, a pre-trained BERT model on the brWac [61] corpus;
- BERTimbau-hatebr – an already fine-tuned version of the BERTimbau model for HS with the HateBR corpus [62];
- mDeBERTa-hatebr – a fine-tuned version of mDeBERTa [63], a multilingual version of DeBERTa, which is an improved version of BERT, for HS detection using the HateBR corpus;
- HateBERTimbau – a retrained version of BERTimbau with 229103 tweets in European Portuguese associated with offensive conversations.

For the training hyperparameters of the BERT models we followed the original paper recommendations, with a batch size of 32, learning rate for Adam optimizer of 2e-5 and 3 epochs [8]. Other attempts were conducted with different parameters, also suggested by the original article, like a batch size of 16 and epochs between 1 and 5, but the used parameters proved to have better performance.

Although some of the models used were already fine-tuned on HS corpora, which was the case with the BERTimbau-hatebr and mDeBERTa-hatebr, we performed further fine-tuning of the models in our corpora, which led to better results. We did not use the previously mentioned Albertina models since at the time of our work only the large version was available, which is very resource intensive, and initial trials did not lead to a better performance.

C. GENERATIVE MODELS

In addition to the BERT-based models, we also explored three additional Transformer-based models for text generation: GPT, Gemini and Mistral. For GPT versions 3.5 and 4 were used, for Gemini version Gemini-Pro was used and the Mistral version used was Mistral-7B-Instruct-v0.3. The inclusion of Mistral in our study was due to its static nature, which addresses the issue of varying performance over time observed on other generative models that are updated over time in an opaque way [64]. Mistral ensures consistent characteristics for all users of the same version which allows for a stable benchmark against which the dynamic nature of GPT and Gemini can be compared, enhancing the robustness of our study. All runs using the generative models were conducted on April 1st, 7th, 10th and June 27th, 2024, ensuring that the results align with the versions of the models current at the time of use.

D. EXPERIMENTAL SETUP

All experiments were conducted using the computational resources of an NVIDIA RTX A6000 GPU with 48 GB of memory, housed within a dedicated machine accessed for the purposes of this study.

For all models, an initial pre-processing of the text was performed to replace all usernames with "@UserID". For the BERT models, the maximum sequence length of each text sample was set to 350 tokens to avoid overloading the GPU. Despite this limitation, a substantial number of messages did not exceed this length, with only 228 comments in the YouTube corpus and none in the Twitter corpus surpassing the threshold. This constraint did not adversely affect the model's performance. To obtain the evaluation metrics, an average of five runs was calculated and the training data was split into 80% for the training set and 20% for the validation set.

Both corpora underwent in-domain and mixed-domain assessments. In the in-domain experiments, the model was

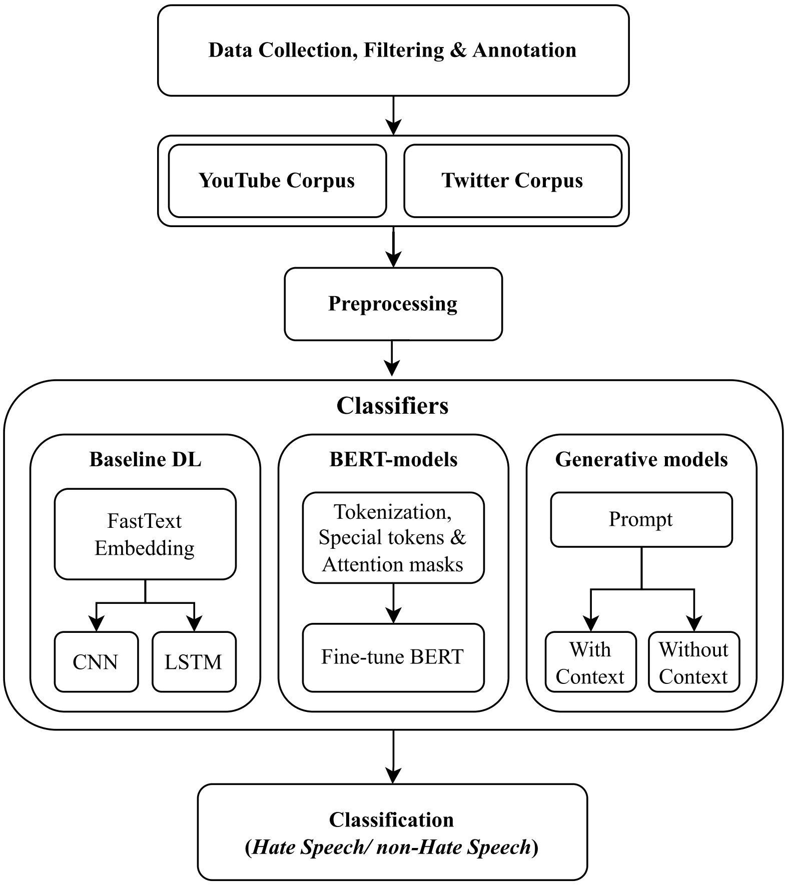

FIGURE 2. Workflow of our HS detection system.

trained exclusively on the corpus being tested. In the mixed-domain experiments, both corpora were used to train the model to determine if additional information from a different source could further enhance performance. For example, the model was trained using data from both the YouTube and Twitter corpora and then tested on the Twitter corpus test set.

Concerning the generative models the following prompts were used to obtain the labels for runs with and without context:

- **No context**: "Consider the comment/tweet delimited by <START> and <END> and indicate whether it is hate speech. Return only the number 1 if yes or the number 0 if not, as in a classification task <START>0<END>";
- **With context (YouTube)**: "Consider the following data about comments about a YouTube video, where <title> is the title of the video, <previous comments> corresponds to previous comments, and <comment> corresponds to the comment being analysed. Please classify the comment as being hate speech (1) or non hate-speech (0) based on the context provided. Return only the values 1 or 0, nothing more. <title>{title} <previous comments>{context} <comment>{current_comment}";
- **With context (Twitter)**: "Consider the following data about twitter conversations, where <context> corresponds to previous tweets in the conversation, and <current_comment> corresponds to the tweet being analysed. Please classify the tweet as being hate speech (1) or non hate-speech (0) based on the context provided. Return only the value 1 or 0, nothing more. <context>{context} <comment>{current_comment}".

For the context runs, each YouTube comment was supplemented with the title of the video in which the comment appeared, as well as preceding comments in the conversation if they belonged to the same thread. For tweets, in the absence of video context, only preceding tweets in the thread were provided, if available.

E. EVALUATION METRICS

The performance of the models was evaluated using three standard metrics, namely, Precision, Recall, and F-score. These metrics are mathematically defined in Equations 1, 2, and 3, respectively, where True Positive (TP) refers to the total number of correctly classified HS instances, False Positive (FP) refers to the total number of non HS instances classified as HS, True Negative (TN) refers to the total number of correctly classified non HS instances, and, finally, False Negative (FN) refers to the total number of HS instances classified as non HS.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{F-score} = \frac{2 * \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

We report the macro, weighted, and positive class scores, but when we assess the models we give more importance to the positive class F-score since it evaluates the class we want to detect and is a more balanced measure, taking into account both FP and FN.

V. RESULTS

This section is divided into two parts: the YouTube corpus results and the Twitter corpus results. We present the results of the BERT-based models for both in-domain and mixed-domain experiments, as well as the results of the generative models with and without contextual information. To ensure the statistical significance of all results presented, we conducted McNemar's test.

A. YOUTUBE

Table 3 summarises the results achieved for the in-domain and mixed-domain experiments of BERT-based models as well as the results of the generative models with and without context. The results of the in-domain experiments reveal that all BERT-based models significantly outperformed the baseline DL models by more than 20 p.p. in regards to the positive class F-score (p-value < 0.01). The best model between the BERT-based models was HateBERTimbau, with an increase in F-score of 1.5 p.p. when compared to the next best model, with all BERT models having similar performance. No significant differences where observed between the BERT-based models with an in-domain setting, with the exception of HateBERTimbau and BERTimbau (p-value < 0.05).

For the mixed-domain experiments, the models were trained with the addition of the Twitter corpora for a total of 45458 messages. The mixed-domain section of Table 3 shows that, although BERTimbau and BERTimbau-hatebr models improved their performance by 0.5% and 0.2%, this difference was not significant, and none of the models surpassed the overall best F-score obtained in the in-domain results, with the best model being again HateBERTimbau. Again, no significant differences were observed between BERT-based models in a mixed-domain setting, and there were no significant differences between in-domain and mixed-domain models, with the exception of HateBERTimbau in-domain and mDeBERTa-hatebr mixed-domain (p-value < 0.05).

Finally, the generative models section of Table 3 presents the results of all generative models. Firstly, we can see that these models have a far worse performance than the BERT models, with a decrease of almost 10 p.p. in F-score between the best models. This was confirmed by the statistical test conducted, where all generative models were significantly worse than all BERT-based models (p-value < 0.01). Comparing the runs where context about the messages was provided versus the ones where no context was provided, we see that the best result was obtained in a context setting, with GPT-3.5 achieving a 0.796 F-score, significantly different than all other generative models (p-value < 0.01), excluding GPT-4 without context. The GPT-3.5 and GPT-4 models were the only ones that improved their performance with the addition of context by 4.6 p.p. and 1.4 p.p. respectively, with only the difference observed in GPT-3.5 being significant (p-value $<$ 0.01). Both Gemini-Pro and Mistral have better performance in a no-context setting, with only the differences observed in Mistral being significant (p-value < 0.05). Although the results obtained were consistent in multiple iterations on the same day, subsequent runs in different days with identical configurations revealed differences of approximately 25 p.p. in some models. For instance, on a previous run of the GPT-4 model, we got a positive class F1 of 0.554, which marks a difference of 17.4% to the F1 presented in Table 3 of 0.728. This goes in line with the literature that shows that the behavior of the ‘‘same’’ model can change substantially in a relatively short amount of time, since these models are updated over time, in an opaque way [64], as mentioned in Section IV. This was also observed for the Twitter generative models.

B. TWITTER

For the Twitter corpus, the results were far worse, when compared with the YouTube corpus, as shown in the in-domain section of Table 4. All BERT-based models had a positive class F-score bellow 50%, with the best being again HateBERTimbau with an F-score of 47.3% (more than 3.5 p.p. above all other BERT models), although without significant differences. Among the BERT models, all significantly outperformed the baseline CNN model (p-value < 0.01), but only HateBERTimbau significantly outperformed the LSTM model (p-value < 0.01), with a 3.3 p.p. increase.

In the Twitter corpus, the addition of information to the models, by incorporating the YouTube comments in the training phase, resulted in an increase in performance, as shown in the mixed-domain section of Table 4. There was in increase of 4 p.p., 5.2 p.p., and 5.1 p.p. in BERTimbau, BERTimbau-hatebr, and mDeBERTa-hatebr models, respectively, all being statistically significant (p-value < 0.05). The previously best performing model, HateBERTimbau, did not see an increase in performance, being significantly worse than it's in-domain counterpart (p-value < 0.01). Contrary to the in-domain models, all mixed-domain models significantly outperformed both baseline models (p-value < 0.01).

Lastly, regarding the results of the generative models in the Twitter corpus, illustrated in the generative models section

TABLE 3. YouTube experiments for both BERT-based models and Generative models.

Model	Positive Class			Macro Avg			Weighted Avg		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
Baseline Models									
CNN (Baseline)	0.840	0.590	0.690	0.610	0.640	0.590	0.720	0.610	0.640
LSTM (Baseline)	0.850	0.500	0.630	0.600	0.630	0.550	0.720	0.570	0.590
BERT-based models: In-domain									
BERTimbau	0.858	0.850	0.853	0.742	0.746	0.744	0.792	0.791	0.792
BERTimbau-hatebr	0.848	0.861	0.855	0.742	0.736	0.738	0.788	0.790	0.789
mDeBERTa-hatebr	0.866	0.847	0.856	0.749	0.758	0.753	0.800	0.796	0.798
HateBERTimbau	0.863	0.879	**0.871**	0.770	0.762	0.766	0.810	0.813	0.811
BERT-based models: Mixed-domain									
BERTimbau	0.867	0.848	0.858↑	0.751	0.760	0.756	0.802	0.798	0.800
BERTimbau-hatebr	0.861	0.853	0.857↑	0.749	0.752	0.750	0.797	0.796	0.796
mDeBERTa-hatebr	0.858	0.829	0.843↓	0.729	0.740	0.734	0.785	0.779	0.781
HATEBERTimbau	0.866	0.853	**0.860**↓	0.754	0.759	0.757	0.803	0.800	0.801
Generative Models: Without context									
Gemini-Pro	0.888	0.669	**0.763**	0.685	0.727	0.680	0.782	0.691	0.706
GPT-3.5	0.873	0.658	0.750	0.665	0.704	0.659	0.750	0.669	0.686
GPT-4	0.875	0.624	0.728	0.661	0.699	0.648	0.754	0.666	0.683
Mistral-7B-Instruct-v0.3	0.895	0.616	0.729	0.675	0.716	0.657	0.770	0.672	0.688
Generative Models: With context									
Gemini-Pro	0.924	0.555	0.693↓	0.681	0.719	0.639	0.779	0.634	0.650
GPT-3.5	0.775	0.817	**0.796**↑	0.611	0.601	0.604	0.729	0.660	0.676
GPT-4	0.862	0.651	0.742↑	0.658	0.693	0.651	0.746	0.675	0.691
Mistral-7B-Instruct-v0.3	0.881	0.464	0.608↓	0.632	0.653	0.567	0.740	0.571	0.585

of Table 4, the inclusion of context did not prove to be beneficial for enhancing the performance, with every model showing significant decline with context (p-value < 0.01), excluding GPT-3.5. However, it is noteworthy that the GPT-3.5 model without context achieved the highest performance out of any model, attaining a score of 50.2%. This model was significantly superior to all in-domain and mixed-domain BERT-based models, as well as all generative models without context (p-value < 0.01), being the only one to exceed the 50% threshold.

Regarding the time performance of the models, all BERT-based models had similar performance, which is to be expected since they are all BERT versions, sharing the same number of parameters and architecture. So BERTimbau, BERTimbau-hatebr, and HateBERTimbau had a testing time of 3.97, 3.97, and 3.95 seconds, respectively, for 825 sentences. mDeBERTa-hatebr was the slowest model, taking 5.79 seconds, since this model is based on DeBERTa-V3 which has 184 million parameters instead of the 110 million present on the other models. For the generative models, their testing time is dependent on the company that provides them, seeing that they control the number of requests allowed. For instance the free version of Gemini is limited to 15 requests per minute (RPM) which accounts for a 55 minute testing time for the same 825 sentences. GPT-3.5 and GPT-4 took 5m19s and 5m57s respectively, and Mistral took 33m17s, although Mistral was tested in a different machine, with the use of Ollama[2] for resource optimization, so it is not possible to draw direct comparisons.

VI. DISCUSSION

Regarding the overall results of the models employed, we can conclude that BERT-based models are more effective for the HS detection task, when compared to generative models and other DL models. This finding aligns well with existing literature [65] and was, to some extent, anticipated, as BERT-based models underwent a fine-tuning process with the used datasets, whereas the generative models were not optimised for our data. Despite this, for the Twitter corpus, GPT-3.5 with the no-context prompt managed to obtain the best result out of all the models. A possible explanation for the surprising results in the Twitter corpus, where all models struggled to even break the 50% positive class F-score mark, could be the low IAA recorded, that showed the annotators had differing views on what constituted HS in this corpus. This divergence of annotations could have impacted the BERT models in the fine-tuning phase, which led to the poor performance. Although the performance of the generative models was not great on its own, they managed to match, and even outperform the BERT models with GPT-3.5. The generative models were also in more agreement between them, with an IAA of 0.542 in their predictions, greater than the 0.355 obtained by the annotators. The disparity observed between the performance obtained in the YouTube and Twitter corpora could also be explained by the differences in discourse style and linguistic characteristics inherent to each platform. Twitter, because of its character limit and fast-paced nature, often has condensed and cryptic language, that can make interpreting and detecting HS more challenging compared to the relatively more verbose and explicit language typically found in YouTube comments. Finally, the prevalence of HS messages on each corpora can also be an explanation for the

[2] https://ollama.com

TABLE 4. Twitter experiments for both BERT-based models and Generative models.

Model	Positive Class			Macro Avg			Weighted Avg		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
Baseline Models									
CNN	0.290	0.750	0.420	0.580	0.600	0.510	0.730	0.530	0.560
LSTM	0.300	0.800	0.440	0.590	0.620	0.520	0.750	0.530	0.560
Transformer Models: In-domain									
BERTimbau	0.511	0.371	0.429	0.679	0.639	0.652	0.778	0.796	0.784
BERTimbau-hatebr	0.494	0.395	0.438	0.672	0.645	0.655	0.777	0.792	0.782
mDeBERTa-hatebr	0.507	0.375	0.431	0.678	0.640	0.653	0.778	0796.	0.784
HATEBERTimbau	0.497	0.454	**0.473**	0.679	0.667	0.672	0.786	0.791	0.788
Transformer Models: Mixed-domain									
BERTimbau	0.422	0.528	0.469↑	0.645	0.670	0.654	0.777	0.754	0.763
BERTimbau-hatebr	0.440	0.542	**0.486**↑	0.657	0.682	0.666	0.784	0.754	0.763
mDeBERTa-hatebr	0.440	0.534	0.482↑	0.656	0.678	0.664	0.783	0.763	0.772
HATEBERTimbau	0.358	0.630	0.456↓	0.619	0.667	0.619	0.772	0.689	0.715
Generative Models: without context									
Gemini-Pro	0.388	0.614	0.476	0.635	0.681	0.643	0.774	0.737	0.751
GPT-3.5	0.400	0.675	**0.502**	0.649	0.706	0.656	0.799	0.720	0.743
GPT-4	0.389	0.705	0.501	0.646	0.708	0.649	0.797	0.711	0.735
Mistral-7b-v0.3	0.371	0.578	0.452	0.621	0.662	0.628	0.768	0.711	0.731
Generative Models: with context									
Gemini-Pro	0.468	0.398	0.430↓	0.659	0.640	0.648	0.758	0.779	0.766
GPT-3.5	0.293	0.729	0.418↓	0.589	0.636	0.546	0.806	0.597	0.633
GPT-4	0.350	0.645	**0.453**↓	0.616	0.667	0.613	0.772	0.680	0.707
Mistral-7b-v0.3	0.314	0.590	0.410↓	0.588	0.628	0.580	0.749	0.650	0.681

difference in results, since the Twitter corpus has a much lower prevalence of HS messages (11.48%) when compared to the YouTube one (64.9%), and although BERT-based models are generally not as susceptible to the quantity of data as traditional ML models, having more data for fine-tuning could still improve the performance of BERT models. This is even more relevant when the task requires domain-specific knowledge and when the dataset is highly diverse or complex, which is the case. Additionally, the standard deviations for the BERT models' results were around 0.002 to 0.012 for the YouTube results and 0.002 to 0.035 for the Twitter results. These higher standard deviations for the Twitter results indicate greater variability in model performance, which can be attributed to the low IAA and the challenging nature of the Twitter data already discussed. In contrast, the lower standard deviations for the YouTube results suggest more consistent model performance in the YouTube corpus.

Upon examining the BERT-based models employed, we can see that HateBERTimbau exhibited the best overall performance for both corpora. This model was retrained with task-relevant data and further fine-tuned with our specific corpora. This model, alongside BERTimbau-hatebr, that was already fine-tuned for the downstream HS detection task and further fine-tuned on our corpora, had the best performances, outperforming both multilingual mDeBERTa and general BERTimbau. The achieved results were expected and are in line with the literature [66], [67]. These models are domain- and task-specific, making them more adept at capturing the intricacies of Portuguese HS language and context. This specialization results in improved performance compared to more general BERT models. In addition, we can see that by retraining the model on European Portuguese we have a better performance than using a model trained on Brazilian Portuguese, confirming the need for European Portuguese resources, even in the presence of Brazilian Portuguese ones. Regarding the mixed-domain tests, the results obtained do not conclusively demonstrate that incorporating information from a different context than the target domain enhances model performance. Although when training the models with both YouTube and Twitter data we observed significant improved performance in the Twitter test set, this improvement was not observed in the YouTube test set. This discrepancy may be attributed to the Twitter data, which, as we already mentioned, may lack sufficient information for effective training due to its low IAA and unique language characteristic. Such specificities may have hindered the model's ability to learn discriminative features relevant to the YouTube domain, thereby limiting its performance. We believe that the inclusion of diverse data sources can enrich the model's understanding and generalization capabilities across domains, as shown in literature [30], however the quality of the data needs to be assured.

Finally, concerning the generative models, GPT-3.5 outperformed Gemini-Pro, GPT-4 and Mistral for the HS detection task, and although their results fell short of the results of BERT models, they still proved to be a viable alternative for HS detection, having fairly good results and outperforming both baseline DL models. In addition, these models were not fine-tuned with our datasets, so there is still room for improvement. When running the models it was apparent that prompt design matters a lot in this type of setting, with different prompts leading to varying results, which is consistent with the literature [68]. Lastly, our experiments

TABLE 5. Examples of misclassifications of our models.

Corpus	Type of misclassification	Message
YouTube	False Positive	Imagina a seguinte situação... Um angolano chega em Portugal e leva corona vírus... Agora pensem na reação de vcs. Sem mais... No cu dos outros é refresco ne [*Imagine the following situation... An Angolan arrives in Portugal and carries the coronavirus... Now think about your reaction. Nothing more... In other people's asses it's refreshment, right?*]
		Cala te vai lavar boca.... brancos e pretos fazem merda tb [*Shut up, wash your mouth.... white and black people do shit too*]
		Os ciganos são muito bem vindes em Portugal. Eles tem cultura, educação e valorizam os povos nativos. Carrega ciganos! Saudações do Brasil. [*Gypsies are very welcome in Portugal. They have culture, education and value native people. Go gypsies! Greetings from Brazil.*]
		"Deles" quem meu?... Tás a falar de quem? Os portugueses são respeitados e bem vindos em Angola mas é claro que uma minoria residual de gente estúpida e mal formada existe em todo o lado. Isso não quer dizer com certeza que a maioria dos angolanos é contra o povo português. Longe disso. [*"Theirs" who man?... Who are you talking about? The Portuguese are respected and welcome in Angola but it is clear that a residual minority of stupid and poorly educated people exists everywhere. This does not mean for sure that the majority of Angolans are against the Portuguese people. Far from it.*]
YouTube	False Negative	Força gonçalo [*Go gonçalo*]
		@User és grande. [*@User you are great.*]
		olá amiga Marta. obrigado eu pela paciência em comentar sempre as publicações e é um gosto perceber que concordo sempre contigo, e este não é exceção. cumprimentos e beijinhos [*Hello friend Marta. Thank you for your patience in always commenting on posts and it's a pleasure to know that I always agree with you, and this one is no exception. greetings and kisses*]
		Mais um comentário excelente. Parece tudo demasiado óbvio e até parece surreal alguém não entender ainda esta visão que é descrita pelo nosso amigo. Obrigado [*Another excellent comment. It all seems too obvious and it even seems surreal that someone still doesn't understand this vision described by our friend. Thanks*]
Twitter	False Positive	então a diferença entre um manifestante e um terrorista continua a ser a cor da pele? Asking for a friend que não acredita em racismo sistémico e privilégio branco. diz que é moda agora. [*So the difference between a protester and a terrorist is still skin color? Asking for a friend who doesn't believe in systemic racism and white privilege. says it's fashionable now.*]
		Ventura convidou dois ciganos, para provar que não é anti-cigano. Um dos ciganos já confessou não ser cigano. A outra é cigana mas: "Não me revejo nas coisas deles, como não trabalhar, falta de higiene, viver em barracas". Em suma, revê-se apenas no discurso xenófobo vs ciganos. [*Ventura invited two gypsies, to prove that he is not anti-gypsy. One of the gypsies has already confessed that he is not a gypsy. The other is a gypsy but: "I don't see myself in their things, like not working, lack of hygiene, living in tents". In short, he sees himself only in the xenophobic vs gypsy discourse.*]
		O Marcelo ser contra os Direitos LGBT é mentira! Tenham vergonha na cara [*Marcelo being against LGBT Rights is a lie! Have shame on your face*]
		E xenofobia também. Tugas no seu melhor [*And xenophobia too. Tugas at its best*]
Twitter	False Negative	@User Não há paciência.. está gente não tem remédio, dão me cabo dos nervos [*@User There is no patience.. there is no remedy for this people, they make me nervous*]
		@User "Pois dos bichos tratam mas das pessoas não" [*@User "Because they take care of animals, but not of people"*]
		@User Respeito o hate [*@User I respect the hate*]
		@User quando eles cumprirem o estado de direito deixo [*@User when they comply with the rule of law I leave it*]

with adding context to the prompts of these models (as they were not fine-tuned) showed that there were improvements in GPT-3.5 and 4 in the YouTube corpus. In all other runs, the addition of context decreased performance, which appears to contradict the literature [44], where context typically enhances performance. However, it has been demonstrated that while GPT-3.5 benefits from context, other models may not [65]. Our findings align with this observation. For the generative models, the standard deviations were between 0.003 and 0.025 for Gemini and between 0.001 and 0.007 for

GPT models. These relatively low standard deviations for the GPT models indicate more consistent performance across runs, while the higher standard deviations for the Gemini model suggest more variability.

VII. ERROR ANALYSIS

To gain insights into the performance of our models, we conducted an error analysis, examining instances of false positives (FP) and false negatives (FN) in the predictions. Notably, we can observe in Table 5 that many FP instances contained counter-speech instances, that commonly have words associated with HS, leading to misclassifications. For example, the comment "Shut up, wash your mouth.... white and black people do shit too" was classified as HS probably because of the inclusion of the term "black" and the negative connotation of the message, despite this being an instance of counter-speech, where the intention was to battle hate. Looking at the other examples, we see the same phenomena happening with other terms like "racism", "gypsies", "Angola", and "LGBT", that appear inserted in an aggressive message, where the intent is to combat HS. While these findings may suggest that the models rely heavily on lexical clues, it is important to note that the misclassified messages closely resemble HS messages in their structure and wording. Thus, while lexical cues play a role, the misclassifications may also stem from the nuanced similarity between these messages and actual instances of HS. Similarly, FN instances often required additional context to discern the presence of HS, particularly for covert forms. In these cases, the absence of explicit HS language made it challenging for the models to accurately identify the underlying harmful intent. For instance, the message "@User you are great." lacked overtly discriminatory language but probably implied support to a previously derogatory sentiment toward a specific group, illustrating the nuanced nature of covert HS.

VIII. CONCLUSION

In this work, we investigated the performance of various transfer learning models in identifying HS in European Portuguese online discourse in a YouTube corpus and a Twitter corpus. Specifically, we compared different BERT-based models – BERTimbau, BERTimbau-hatebr, mDeBERTa-hatebr, and HateBERTimbau – along with four generative models – GPT-4, GPT-3.5, Gemini-Pro and Mistral-7B-Instruct-v0.3. HateBERTimbau achieved the best positive class F-score with 87.1% for the YouTube corpus, surpassing the baseline scores by more than 20 p.p., and GPT-3.5 achieved the best performance for the Twitter corpus with a positive class F-score of 50.2%, with an increase of 6.2 p.p. compared to the baseline. We showed that the incorporation of mixed-domain data for the training of the models has the potential to improve performance, significantly increasing the performance of BERT models in the Twitter corpus, by training them with the Twitter and YouTube corpus simultaneously. In order to achieve this, it is necessary to ensure the quality of the data, since none of the models had an improvement in performance when the Twitter data was incorporated, which may be caused by the low IAA between annotators, potentially adding noise to the models. For the generative models, they had a worse performance when compared with the BERT models in the YouTube corpus, but since there was no fine-tuning done, and the models did not learn from the annotations of the training data – they made predictions based on their representations of HS, that may not be aligned with our definition. This can also be the reason why in the Twitter corpus they managed to outperform the BERT models, because they where not exposed to the possible noisy data with low IAA.

Overall, our study contributes to understanding the effectiveness of different transfer learning models for HS detection, in general, and in European Portuguese online discourse, specifically. Our findings suggest that BERT-based models fine-tuned for the HS detection task have better performance than general BERT models not fine-tuned for a downstream task, and that models retrained on European Portuguese are more effective in identifying HS in European Portuguese than models trained on only Brazilian Portuguese. Regarding the error analysis, we found that some of the messages mislabeled as non-HS did not have sufficient context to be able to be classified as HS. This underscores the necessity for additional context provided by preceding messages. Additionally, some of the messages mislabeled as HS were in fact counter-speech attempts or messages containing words that are often used in HS messages, which further confirms the need to provide some context to the models in order to accurately predict HS. To overcame this limitation, future work could focus on incorporating context alongside target messages to better inform the models, especially the BERT-based ones; distinguishing between overt and covert hate speech may also lead to better representations of the different types of HS and improve classification accuracy; and, finally, pre-fine-tuning generative models with training data to align with annotation criteria. We also believe that for future work, multi-class detection attempts should be made, especially in detecting HS directed at different target groups, such as those present in our datasets. Furthermore, recent studies have explored network immunization after detection in various ways. Either by proactive approaches [69], tree-based approaches [70], community-based approaches [71], or real-time approaches [72], they aim to stop the propagation of problematic content in networks. We consider this a very promising avenue for application in the HS detection space. Future work should combine both tasks: HS detection and network immunization, to not only identify forms of HS but also to effectively mitigate their spread within online communities. This integrated approach could enhance the overall effectiveness of HS management and contribute to creating safer online environments.

While our study has provided valuable insights concerning the effectiveness of different transfer learning models for HS detection, it is important to acknowledge some limitations. Specifically, our corpora were annotated by a small number of annotators, ranging from three to four individuals, each with distinct backgrounds. This variability among annotators may introduce considerable data variance.

ACKNOWLEDGMENT

This work was supported in part by the European Union under Grant CERV-2021-EQUAL (101049306). However the views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or Knowhate Project. Neither the European Union nor the Knowhate Project can be held responsible.

REFERENCES

[1] Statista. (Aug. 2023). *Number of Social Media Users Worldwide From 2017 to 2027*. [Online]. Available: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

[2] H. Watanabe, M. Bouazizi, and T. Ohtsuki, "Hate speech on Twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection," *IEEE Access*, vol. 6, pp. 13825–13835, 2018. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8292838

[3] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, and V. Patti, "Resources and benchmark corpora for hate speech detection: A systematic review," *Lang. Resour. Eval.*, vol. 55, no. 2, pp. 477–523, Jun. 2021, doi: 10.1007/s10579-020-09502-8.

[4] P. Carvalho and R. Guerra, "D3.2/D3.3 annotation guidelines OHS & OCS," Knowhate Project—CERV-2021-EQUAL (101049306), Private/Sensitive—Limited Under the Conditions of the Grant Agreement, ISCTE-Inst. Univ. de Lisboa, Lisbon, Portugal, Tech. Rep., May 2023.

[5] B. M. Tynes, M. T. Giang, D. R. Williams, and G. N. Thompson, "Online racial discrimination and psychological adjustment among adolescents," *J. Adolescent Health*, vol. 43, no. 6, pp. 565–569, Dec. 2008.

[6] Media Smarts. (2021). *Impact of Online Hate*. [Online]. Available: https://mediasmarts.ca/online-hate/impact-online-hate

[7] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, P. S. Yu, and L. He, "A survey on text classification: From traditional to deep learning," *ACM Trans. Intell. Syst. Technol.*, vol. 13, no. 2, pp. 1–41, Apr. 2022, doi: 10.1145/3495162.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, MI, USA: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[9] Council of Europe. (2021). *Portugal Should Act More Resolutely to Tackle Racism and Continue Efforts to Combat Violence Against Women*. [Online]. Available: https://www.coe.int/en/web/commissioner/-/portugal-should-act-more-resolutely-to-tackle-racism-and-continue-efforts-to-combat-violence-against-women

[10] M. S. Jahan and M. Oussalah, "A systematic review of hate speech automatic detection using natural language processing," *Neurocomputing*, vol. 546, Aug. 2023, Art. no. 126232. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231223003557

[11] C. Carvalho, R. Guerra, C. Marques, R. Sarroeira, F. Batista, R. RIbeiro, S. Moro, and C. Silva, "Unpacking online hate speech in Portuguese social media: A social-psychological and linguistic-discursive approach," 2024.

[12] R. Guerra, C. Carvalho, C. Marques, R. Sarroeira, F. Batista, R. RIbeiro, S. Moro, and C. Silva, "Counter-speech: A social-psychological and linguistic approach to counter online hate," 2024.

[13] R. Chaves Rodrigues, M. Tanti, and R. Agerri. (Mar. 2023). *Evaluation of Portuguese Language Models*. [Online]. Available: https://github.com/ruanchaves/eplm

[14] B. C. Matos, "Automatic hate speech detection in Portuguese social media text," M.S. thesis, Inst. Superior Técnico, Lisbon, Portugal, Nov. 2022.

[15] F. Souza, R. Nogueira, and R. Lotufo, "BERTimbau: Pretrained BERT models for Brazilian Portuguese," in *Proc. 9th Brazilian Conf. Intell. Syst.*, Rio Grande do Sul, Brazil, Oct. 2020, pp. 403–417.

[16] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., Dec. 2020, pp. 1–25. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[17] G. Team, R. Anil, S. Borgeaud, Y. Wu, J. B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, and K. Millican, "Gemini: A family of highly capable multimodal models," 2023, *arXiv:2312.11805*.

[18] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. Le Scao, T. Lavril, T. Wang, T. Lacroix, and W. El Sayed, "Mistral 7B," 2023, *arXiv:2310.06825*.

[19] P. Carvalho, B. Matos, R. Santos, F. Batista, and R. Ribeiro, "Hate speech dynamics against African descent, Roma and LGBTQ+ communities in Portugal," in *Proc. 13th Conf. Lang. Resour. Eval.*, 2022, pp. 2362–2370.

[20] F. Baider, "Covert hate speech, conspiracy theory and anti-semitism: Linguistic analysis versus legal judgement," *Int. J. Semiotics Law-Revue Internationale de Sémiotique Juridique*, vol. 35, no. 6, pp. 2347–2371, Dec. 2022.

[21] N. Pitropakis, K. Kokot, D. Gkatzia, R. Ludwiniak, A. Mylonas, and M. Kandias, "Monitoring users' behavior: Anti-immigration speech detection on Twitter," *Mach. Learn. Knowl. Extraction*, vol. 2, no. 3, pp. 192–215, Aug. 2020.

[22] R. Saeed, H. Afzal, S. A. Rauf, and N. Iltaf, "Detection of offensive language and ITS severity for low resource language," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 22, no. 6, pp. 1–27, Jun. 2023.

[23] S. K. Mohapatra, S. Prasad, D. K. Bebarta, T. K. Das, K. Srinivasan, and Y.-C. Hu, "Automatic hate speech detection in english-odia code mixed social media data using machine learning techniques," *Appl. Sci.*, vol. 11, no. 18, p. 8575, Sep. 2021.

[24] F. Shannaq, B. Hammo, H. Faris, and P. A. Castillo-Valdivieso, "Offensive language detection in Arabic social networks using evolutionary-based classifiers learned from fine-tuned embeddings," *IEEE Access*, vol. 10, pp. 75018–75039, 2022.

[25] C. Arcila-Calderón, J. J. Amores, P. Sánchez-Holgado, and D. Blanco-Herrero, "Using shallow and deep learning to automatically detect hate motivated by gender and sexual orientation on Twitter in Spanish," *Multimodal Technol. Interact.*, vol. 5, no. 10, p. 63, Oct. 2021.

[26] T. Turki and S. S. Roy, "Novel hate speech detection using word cloud visualization and ensemble learning coupled with count vectorizer," *Appl. Sci.*, vol. 12, no. 13, p. 6611, Jun. 2022.

[27] P. K. Roy, A. K. Tripathy, T. K. Das, and X.-Z. Gao, "A framework for hate speech detection using deep convolutional neural network," *IEEE Access*, vol. 8, pp. 204951–204962, 2020.

[28] A. T. Kabakus, "Towards the importance of the type of deep neural network and employment of pre-trained word vectors for toxicity detection: An experimental study," *J. Web Eng.*, vol. 20, pp. 2243–2268, Nov. 2021.

[29] F. E. Ayo, O. Folorunso, F. T. Ibharalu, and I. A. Osinuga, "Hate speech detection in Twitter using hybrid embeddings and improved cuckoo search-based neural networks," *Int. J. Intell. Comput. Cybern.*, vol. 13, no. 4, pp. 485–525, Nov. 2020.

[30] S. Dascălu and F. Hristea, "Towards a benchmarking system for comparing automatic hate speech detection with an intelligent baseline proposal," *Mathematics*, vol. 10, no. 6, p. 945, Mar. 2022.

[31] H. Madhu, S. Satapara, S. Modha, T. Mandl, and P. Majumder, "Detecting offensive speech in conversational code-mixed dialogue on social media: A contextual dataset and benchmark experiments," *Exp. Syst. Appl.*, vol. 215, Apr. 2023, Art. no. 119342.

[32] H. Karayigit, Ç. Aci, and A. Akdagli, "Detecting abusive Instagram comments in Turkish using convolutional neural network and machine learning methods," *Exp. Syst. Appl.*, vol. 174, Jul. 2021, Art. no. 114802.

[33] I. Priyadarshini, S. Sahu, and R. Kumar, "A transfer learning approach for detecting offensive and hate speech on social media platforms," *Multimedia Tools Appl.*, vol. 82, no. 18, pp. 27473–27499, Jul. 2023.

[34] M. Fazil, S. Khan, B. M. Albahlal, R. M. Alotaibi, T. Siddiqui, and M. A. Shah, "Attentional multi-channel convolution with bidirectional LSTM cell toward hate speech prediction," *IEEE Access*, vol. 11, pp. 16801–16811, 2023.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Long Beach, CA, USA: Curran Associates, 2017, pp. 1–9.

[36] J. M. Molero, J. Pérez-Martín, A. Rodrigo, and A. Peñas, "Offensive language detection in Spanish social media: Testing from bag-of-words to transformers models," *IEEE Access*, vol. 11, pp. 95639–95652, 2023.

[37] J. M. Pérez, F. M. Luque, D. Zayat, M. Kondratzky, A. Moro, P. S. Serrati, J. Zajac, P. Miguel, N. Debandi, A. Gravano, and V. Cotik, "Assessing the impact of contextual information in hate speech detection," *IEEE Access*, vol. 11, pp. 30575–30590, 2023.

[38] F. Rodríguez-Sánchez, J. Carrillo-de-Albornoz, and L. Plaza, "Automatic classification of sexism in social networks: An empirical study on Twitter data," *IEEE Access*, vol. 8, pp. 219563–219576, 2020.

[39] M. R. Awal, R. K. Lee, E. Tanwar, T. Garg, and T. Chakraborty, "Model-agnostic meta-learning for multilingual hate speech detection," *IEEE Trans. Computat. Social Syst.*, vol. 11, no. 1, pp. 1086–1095, Feb. 2024.

[40] M. Bhardwaj, M. Sundriyal, M. Bedi, M. S. Akhtar, and T. Chakraborty, "HostileNet: Multilabel hostile post detection in Hindi," *IEEE Trans. Computat. Social Syst.*, vol. 11, no. 2, pp. 1842–1852, Apr. 2024.

[41] M. Casavantes, M. E. Aragón, L. C. González, and M. Montes-y-Gómez, "Leveraging posts' and authors' metadata to spot several forms of abusive comments in Twitter," *J. Intell. Inf. Syst.*, vol. 61, no. 2, pp. 519–539, Oct. 2023.

[42] C.-O. Truică and E.-S. Apostol, "It's all in the embedding! Fake news detection using document embeddings," *Mathematics*, vol. 11, no. 3, p. 508, Jan. 2023, doi: 10.3390/math11030508.

[43] C.-O. Truica, E.-S. Apostol, and A. Paschke. (2022). *Awakened At Check-that! 2022: Fake News Detection Using Bilstm and Sentence Transformer*. [Online]. Available: https://publica.fraunhofer.de/handle/publica/425268

[44] K.-L. Chiu, A. Collins, and R. Alexander, "Detecting hate speech with GPT-3," 2022, *arXiv:2103.12407*.

[45] A. S. Oliveira, T. C. Cecote, P. H. L. Silva, J. C. Gertrudes, V. L. S. Freitas, and E. J. S. Luz, "How good is ChatGPT for detecting hate speech in Portuguese?" in *Proc. Anais do 14th Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, Sep. 2023, pp. 94–103.

[46] P. Fortuna, J. Rocha da Silva, J. Soler-Company, L. Wanner, and S. Nunes, "A hierarchically-labeled Portuguese hate speech dataset," in *Proc. 3rd Workshop Abusive Lang. Online*, S. T. Roberts, J. Tetreault, V. Prabhakaran, and Z. Waseem, Eds., Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 94–104. [Online]. Available: https://aclanthology.org/W19-3510

[47] R. B. Santos, B. C. Matos, P. Carvalho, F. Batista, and R. Ribeiro, "Semi-supervised annotation of Portuguese hate speech across social media domains," in *Proc. 11th Symp. Lang., Appl. Technol.* (Open Access Series in Informatics (OASIcs)), vol. 104, J. A. Cordeiro, M. J. A. Pereira, N. F. Rodrigues, and S. A. Pais, Eds., Dagstuhl, Germany: Schloss Dagstuhl-Leibniz-Zentrum Für Informatik, 2022, pp. 1–14. [Online]. Available: https://drops.dagstuhl.de/entities/document/10.4230/OASIcs.SLATE.2022.11

[48] B. C. Matos, R. B. Santos, P. Carvalho, R. Ribeiro, and F. Batista, "Comparing different approaches for detecting hate speech in online Portuguese comments," in *Proc. 11th Symp. Lang., Appl. Technol.* (Open Access Series in Informatics (OASIcs)), vol. 104, J. A. Cordeiro, M. J. A. Pereira, N. F. Rodrigues, and S. A. Pais, Eds., Dagstuhl, Germany: Schloss Dagstuhl-Leibniz-Zentrum Für Informatik, 2022, pp. 1–12, doi: 10.4230/OASIcs.SLATE.2022.10.

[49] 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," *J. Lang. Aggression Conflict*, Jun. 2023. [Online]. Available: https://www.jbe-platform.com/content/journals/10.1075/jlac.00085.car

[50] R. Santos, J. Rodrigues, L. Gomes, J. Silva, A. Branco, H. L. Cardoso, T. F. Osório, and B. Leite, "Fostering the ecosystem of open neural encoders for Portuguese with albertina PT* family," 2024, *arXiv:2403.01897*.

[51] A. A. Firmino, C. de Souza Baptista, and A. C. de Paiva, "Improving hate speech detection using cross-lingual learning," *Exp. Syst. Appl.*, vol. 235, Jan. 2024, Art. no. 121115. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417423016172

[52] A. Silva and N. Roman, "Hate speech detection in Portuguese with naïve bayes, SVM, mlp and logistic regression," in *Proc. Anais do 17th Encontro Nacional de Inteligência Artificial e Computacional*. Porto Alegre, RS, Brasil: SBC, 2020, pp. 1–12. [Online]. Available: https://sol.sbc.org.br/index.php/eniac/article/view/12112

[53] A. A. Firmino, C. S. de Baptista, and A. C. de Paiva, "Using cross lingual learning for detecting hate speech in Portuguese," in *Database and Expert Systems Applications*, C. Strauss, G. Kotsis, A. M. Tjoa, and I. Khalil, Eds., Cham, Switzerland: Springer, 2021, pp. 170–175.

[54] R. Marques, "On the system of mood in European and Brazilian Portuguese," *J. Portuguese Linguistics*, vol. 3, no. 1, pp. 89–109, Jun. 2004.

[55] T. Móia and A. T. Alves, "Differences between European and Brazilian Portuguese in the use of temporal adverbials," *J. Portuguese Linguistics*, vol. 3, no. 1, p. 37, Jun. 2004.

[56] F. Baider, "Pragmatics lost: Overview, synthesis and proposition in defining online hate speech," *Pragmatics Soc.*, vol. 11, no. 2, pp. 196–218, Jul. 2020. [Online]. Available: https://www.jbe-platform.com/content/journals/10.1075/ps.20004.bai

[57] M. Pohjonen and S. Udupa, "Extreme speech online: An anthropological critique of hate speech debates," *Int. J. Commun.*, vol. 11, p. 19, Mar. 2017. [Online]. Available: https://ijoc.org/index.php/ijoc/article/view/5843

[58] G. Marzi, M. Balzano, and D. Marchiori, "K-alpha calculator–Krippendorff's alpha calculator: A user-friendly tool for computing Krippendorff's alpha inter-rater reliability coefficient," *MethodsX*, vol. 12, Jun. 2024, Art. no. 102545.

[59] A. Safaya, M. Abdullatif, and D. Yuret, "KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media," in *Proc. 14th Workshop Semantic Eval.*, A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, and E. Shutova, Eds., 2020, pp. 2054–2059. [Online]. Available: https://aclanthology.org/2020.semeval-1.271

[60] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning word vectors for 157 languages," in *Proc. Int. Conf. Lang. Resour. Eval.*, 2018, pp. 1–5.

[61] J. A. Wagner Filho, R. Wilkens, M. Idiart, and A. Villavicencio, "The brWaC corpus: A new open resource for Brazilian Portuguese," in *Proc. 11th Int. Conf. Lang. Resour. Eval.*, N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, Eds., Miyazaki, Japan: European Language Resources Association (ELRA), May 2018, pp. 1–6. [Online]. Available: https://aclanthology.org/L18-1686

[62] F. Vargas, I. Carvalho, F. R. de Góes, T. Pardo, and F. Benevenuto, "HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection," in *Proc. 13th Lang. Resour. Eval. Conf.*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, Eds., Marseille, France: European Language Resources Association, Jun. 2022, pp. 7174–7183. [Online]. Available: https://aclanthology.org/2022.lrec-1.777

[63] P. He, J. Gao, and W. Chen, "DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing," 2021, *arXiv:2111.09543*.

[64] L. Chen, M. Zaharia, and J. Zou, "How is ChatGPT's behavior changing over time?" 2023, *arXiv:2307.09009*.

[65] G. Assis, A. Amorim, J. Carvalho, D. de Oliveira, D. Vianna, and A. Paes, "Exploring Portuguese hate speech detection in low-resource settings: Lightly tuning encoder models or in-context learning of large models?" in *Proc. 16th Int. Conf. Comput. Process. Portuguese*, P. Gamallo, D. Claro, A. Teixeira, L. Real, M. Garcia, H. G. Oliveira, and R. Amaro, Eds., Santiago de Compostela, Galicia/Spain: Association for Computational Lingustics, Mar. 2024, pp. 301–311. [Online]. Available: https://aclanthology.org/2024.propor-1.31

[66] F. M. Plaza-del-Arco, M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia, "Comparing pre-trained language models for Spanish hate speech detection," *Exp. Syst. Appl.*, vol. 166, Mar. 2021, Art. no. 114120.

[67] J. A. Benítez-Andrades, Á. González-Jiménez, Á. López-Brea, J. Aveleira-Mata, J.-M. Alija-Pérez, and M. T. García-Ordás, "Detecting racism and xenophobia using deep learning models on Twitter data: CNN, LSTM and BERT," *PeerJ Comput. Sci.*, vol. 8, p. e906, Mar. 2022.

[68] L. Li, L. Fan, S. Atreja, and L. Hemphill, "'HOT' ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media," *ACM Trans. Web*, vol. 18, no. 2, pp. 1–36, 2023.

[69] A. Petrescu, C.-O. Truică, E.-S. Apostol, and P. Karras, ''Sparse shield: Social network immunization vs. harmful speech,'' in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manag.* New York, NY, USA: Association for Computing Machinery, Oct. 2021, pp. 1426–1436, doi: 10.1145/3459637.3482481.

[70] C.-O. Truică, E.-S. Apostol, R.-C. Nicolescu, and P. Karras, ''MCWDST: A minimum-cost weighted directed spanning tree algorithm for real-time fake news mitigation in social media,'' *IEEE Access*, vol. 11, pp. 125861–125873, 2023.

[71] E. S. Apostol, Ö. Coban, and C.-O. Truica, ''CONTAIN: A community-based algorithm for network immunization,'' *Eng. Sci. Technol., Int. J.*, vol. 55, Jul. 2024, Art. no. 101728. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2215098624001149

[72] E.-S. Apostol, C.-O. Truică, and A. Paschke, ''ContCommRTD: A distributed content-based misinformation-aware community detection system for real-time disaster reporting,'' 2023, *arXiv:2301.12984*.

GIL RAMOS received the B.S. degree in sports sciences from the Faculty of Human Kinetics, Lisbon, in 2021, and the M.S. degree in physical education. He is currently pursuing the M.S. degree in data science with Lisbon University Institute (ISCTE-IUL). Since 2023, he has been a Research Assistant with the ISTAR—Information Sciences and Technologies and Architecture Research Center, Lisbon. He has previous experience in scientific research with two publications, while attending the M.S. degree. His research interest includes text mining.

FERNANDO BATISTA (Senior Member, IEEE) received the Ph.D. degree in computer science and engineering from the Instituto Superior Ténico (IST), in 2011. From 2011 to 2015, he was a member of the LxMLS Technical Staff. He is currently an Associate Professor with the ISCTE—University Institute of Lisbon and an Integrated Researcher with INESC-ID, Lisbon. Since 2016, he has been a member of the organization team of the LxMLS—Lisbon Machine Learning Summer School. His current research interests include spoken and written natural language processing, machine learning, and text mining for social media. Since 2008, he has been a member of the ISCA Speech.

RICARDO RIBEIRO is an Associate Professor with the ISCTE–Instituto Universitrio de Lisboa (ISCTE-IUL), where he is currently a Coordinator of the artificial intelligence scientific area, and the Director of the M.Sc. studies in computer engineering. He is also an Integrated Researcher at INESC-ID, Lisbon, working on human language technologies. He has participated in several European and nationally-funded projects. He was the Human Language Technologies INESC-ID Team Coordinator at RAGE (2015–2019) European-funded project and a principal investigator of a Ministry of National Defence-funded project on information extraction from text. He has participated in several scientific events, such as IJCAI, ICASSP, LREC, and Interspeech, either as an organizer or as a member of the program committee. His current research interests include high-level information extraction from unrestricted text, speech, or music, and improving machine-learning techniques using domain-related information. He was an editor of a book on the computational processing of Portuguese.

PEDRO FIALHO is currently a Research Assistant with the ISTAR Research Center (formerly INESC-ID), he has been involved in various research projects related to natural language processing, mostly focused on the integration of speech technologies and embodied virtual assistants and their application in health and education research problems. His Ph.D. studies were focused on models to assess the semantic similarity between sentences, for which he contributed an approach based on formal semantic representations. His work has received 231 Google Scholar (H-index of 7) citations.

SÉRGIO MORO is a Full Professor with ISCTE-IUL, where is currently the Vice President of the Scientific Council. He is also the former Deputy Director with the ISTAR Research Center, where he coordinated the Information Systems Group. He is the Deputy Director of the Department of Information Science and Technology and the Director of the Master in Data Science. He is an Interdisciplinary Data Scientist who envisions the development of innovative predictive systems through data science approaches in distinct domains including managerial and social sciences. His work has received more than 2900 Scopus (H-index of 27) and more than 5400 Google Scholar (H-index of 34) citations.

ANTÓNIO FONSECA received the degree in electronics and telecommunications from IST, Lisbon, and the master's and Ph.D. degrees in complex systems sciences from ISCTE/IUL and Faculdade de Ciências de Lisboa. He is currently an Engineer and Complexity Sciences Researcher. He is also an Assistant Professor with ISCTE-Sintra and an Integrated Researcher with ISTAR/ISCTE-IUL, Lisbon. He worked in several private companies as an Engineer before joining the public service as an IT Expert. His main research interests include complex social systems, information dynamics and modeling, and simulation of complex systems.

RITA GUERRA received the Ph.D. degree. She is currently a Senior Researcher with CIS-ISCTE. She coordinates research projects and consortiums supported by several competitive grants from the European Commission (Citizens, Equality, and Rights and Values Programme; European Fund for the Integration of non-EU immigrants). Her research interests include prejudice and discrimination, extremism, and online hate speech. She is an Elected Member of the CIS-ISCTE Scientific Commission, a Coordinator of the Thematic Line Global Governance of the Sociodigital Laboratory for Public Policy, and the Vice President of the Scientific Council of Iscte_Conhecimento e Inovação. She is also an Associate Editor of *Journal of Applied Social Psychology*.

PAULA CARVALHO is currently a Researcher with INESC-ID. Her research interests include corpus linguistics, computational linguistics, and natural language processing. She has been involved in several interdisciplinary research projects related to the development of language resources and methods, and their application to digital humanities and social sciences research problems. Her main focus revolves around tackling information disorder, including hate speech and prevalent on social media platforms.

CATARINA MARQUES received the Ph.D. degree in quantitative methods with a specialization in statistics and data analysis. She is currently an Associate Professor with ISCTE—Instituto Universitáriode Lisboa and an Integrated Researcher of BRU-ISCTE. She has participated in several international scientific conferences as a scientific committee member. Her main research interest includes structural equation modeling, and more recently she has used her experience and competencies to research in data science.

CLÁUDIA SILVA is currently an Invited Assistant Professor with the Department of Computer Science and Engineering, Instituto Superior Técnico (IST), University of Lisbon, and an Integrated Research Fellow with the Interactive Technologies Institute/Laboratory of Robotics and Engineering Systems (ITI-LARSyS). Her research interests include transdisciplinary and practice-based. Recently, she has been interested in exploring interactive digital counter-narratives to combat hate speech.