Socially Responsible Hate Speech Detection: Can Classifiers Reflect Social Stereotypes?

Francielle Vargas[1,2], **Isabelle Carvalho**[1], **Ali Hürriyetoğlu**[3]
Thiago A. S. Pardo[1], **Fabrício Benevenuto**[2]
[1]Institute of Mathematical and Computer Sciences, University of São Paulo, Brazil
[2]Computer Science Department, Federal University of Minas Gerais, Brazil
[3] KNAW Humanities Cluster, DHLab, the Netherlands
francielleavargas@usp.br

Abstract

Recent studies have shown that hate speech technologies may propagate social stereotypes against marginalized groups. Nevertheless, there has been a lack of realistic approaches to assess and mitigate biased technologies. In this paper, we introduce a new approach to analyze the potential of hate-speech classifiers to reflect social stereotypes through the investigation of stereotypical beliefs by contrasting them with counter-stereotypes. We empirically measure the distribution of stereotypical beliefs by analyzing the distinctive classification of tuples containing stereotypes versus counter-stereotypes in machine learning models and datasets. Experiment results show that hate speech classifiers attribute unreal or negligent offensiveness to social identity groups by reflecting and reinforcing stereotypical beliefs regarding minorities. Furthermore, we also found out that models that embed expert and context information from offensiveness markers present promising results to mitigate social stereotype bias towards socially responsible hate speech detection.[1]

1 Introduction

While Artificial Intelligence (AI) technologies have generated unprecedented opportunities for society, they have also introduced new forms of perpetuating inequality and heightened threats to human rights and well-being (UN, 2023).

In this context, the investigation of Hate Speech (HS) is undoubtedly important since the proposition of automated systems has implications for unprejudiced societies. Nevertheless, researchers have constantly observed that these technologies are being developed with scarce consideration of their potential social biases, which may perpetuate social inequalities when propagated at scale (Davani et al., 2023; Blodgett et al., 2020; Chuang et al., 2021; Xia et al., 2020; Wiegand et al., 2019; Sap et al., 2019; Bordia and Bowman, 2019; Davidson et al., 2019). For example, Table 1 shows that the hate speech classifier attributed unreal offensiveness to the first example only due to the expression "Gay Pride", which represents a social identity[2] group. We observe that in the second example, the expression "Gay Pride" was replaced by "Christmas", and in the third example, they were removed. The second and third examples were classified as non-hate speech, and the first one was classified as hate speech. Furthermore, the hate speech classifier neglected the offensiveness of the fourth example only due to the term "Mexican".

Examples	**Classes**
They've been known to disregard people who march in **Gay Pride** parades.	Hate Speech
They've been known to disregard people who march in **Christmas** parades.	Non-Hate Speech
They've been known to disregard people who march in parades.	Non-Hate Speech
We can't go to that one in this **Mexican** neighborhood. You might be robbed.	Non-Hate Speech
We can't go to that one in this **Canadian** neighborhood. You might be robbed.	Hate Speech
We can't go to that one in this neighborhood. You might be robbed.	Hate Speech

Table 1: Stereotype Bias in Hate Speech Detection.

According to Warner and Hirschberg (2012), hate speech is a particular form of offensive language that considers stereotypes to express an ideology of hate. A stereotype is an over-generalized belief about a particular group of people (e.g., Asians are good at math or African Americans are athletic), and beliefs (biases) are known to target social groups (Nadeem et al., 2021). Social and stereotyp-

[1] **Warning**: This paper contains examples of offensive content and stereotypes. It does not reflect our way of thinking.

[2] Social identity is a theory of social psychology that offers a motivational explanation for in-group bias.

Proceedings of Recent Advances in Natural Language Processing, pages 1187–1196
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_126

ical biases are forms of discrimination against a social group based on characteristics such as gender, sexual orientation, religion, ethnicity, etc. (Fiske, 1993; Sahoo et al., 2022).

Hate speech technologies reflect social stereotypes due to bias in the training data (Davidson et al., 2019; Yörük et al., 2022) triggered early from human annotation (Wiegand et al., 2019), in the text representations that learn normative social stereotypes associated with systematic prediction errors (Davani et al., 2023), and also due to missing context information (Davidson et al., 2019). For example, if "programmer" appears more frequently with "he" than "she" in the training data, it will create a biased association to "he" compared with "she" in the model (Qian, 2019). In the same settings, if "African American" appears frequently associated with vocabulary related to baseball and violence, the model will potentially learn this association from the training data. Therefore, both examples demonstrate the harmful potential of HS classifiers reflecting different types of social stereotypical beliefs that may negatively influence people's perception of marginalized groups.

State-of-the-art analysis of social stereotypes in Hate Speech Detection (HSD) is definitely an under-explored issue. Recently, a few works have analyzed social stereotypes bias in (i) text representation, which maps textual data to their numeric representations in a semantic space, and (ii) human annotations, which represent subjective judgments about hate speech in text content, constituting the training dataset. Therefore, in both cases, social stereotypes may be included in the final trained model (Davani et al., 2023; Elsafoury, 2022). A recent study proposed by Davani et al. (2023), concluded that hate speech classifiers can learn normative social stereotypes once their language mapping to numeric representations is affected by stereotypical co-occurrences in the training data.

The social psychology literature suggests that one of the most effective ways to reduce biased thinking is countering stereotypical beliefs with counter-stereotypes (also known as anti-stereotypes) (Fraser et al., 2021). For instance, once a human is asked to classify a tuple containing social stereotypes and counter-stereotypes, and the result is a distinctive classification, it evidences biased stereotypical beliefs. In this same setting, Finnegan et al. (2015) proposed experiments in which participants were shown stereotypical and counter-stereotypical images of socially-gendered professions (e.g., a surgeon is stereotypically male, and a nurse is stereotypically female). They reversed the genders in the counter-stereotypical images and then measured their gender bias in a judgment task. Results showed that exposure to counter-stereotypical images significantly reduced gender normative stereotypes. Finally, in de Vassimon Manela et al. (2021), Blair IV (2001), and Nilanjana and G. (2001), the authors also used the same strategy to mitigate socially biased thinking.

In this paper, we study the potential of HS classifiers to reflect social stereotypes against marginalized groups. We propose a new approach, entitled **Social Stereotype Analysis (SSA)**, which consists of analyzing stereotypical beliefs by contrasting them with counter-stereotypes. We first implement HS classifiers using different Machine Learning (ML) text representations in two different datasets in English and Portuguese, composed of Twitter and Instagram data. Then, we assess the potential of these models to reflect social stereotypes through a distinctive analysis of tuples containing stereotypes versus counter-stereotype. The results demonstrate that HS classifiers may provide unreal or negligent offensiveness classification to social identity groups, hence reflecting and reinforcing social stereotypical beliefs against marginalized groups. Finally, based on our findings, ML models that embed expert and context information from explicit and implicit offensiveness markers present promising results towards mitigating the risk of HS classifiers propagating social stereotypical beliefs. Our contributions may be summarized as follows:

- We study and empirically analyze the potential of HS classifiers to reflect social stereotypes against marginalized groups.
- We provide a set of experiments with different ML models in two languages (English and Portuguese). The datasets and code are available[3], which may facilitate future research.
- We propose a new approach for assessing the potential of HS classifiers to reflect social stereotypes. Our approach consists of analyzing whether HS classifiers are able to classify tuples containing stereotypes and counter-stereotypes in the same way. Otherwise, they are potentially biased.

[3] https://github.com/franciellevargas/SSA

2 Related Work

Bias in Human-Annotation and Datasets: Bias may be triggering early from human annotation. As a result, biased datasets propagate their social bias through data training. According to Vargas et al. (2022), a strategy based on a diversified profile of annotators (e.g. gender, race-color, political orientation, etc.) and balanced variables during the data collection should be adopted to mitigate social biases. Furthermore, they proposed an annotation schema for hate speech and offensive language detection in Brazilian Portuguese towards social bias mitigation. Davidson et al. (2019) analyzed racial bias by training classifiers in HS datasets of Twitter in order to identify whether the tweets written in African-American English are classified as abusive more frequently than tweets written in Standard American English. As a result, this phenomenon widely-held beliefs about different social categories and may harm minority social groups. Sap et al. (2019) investigated how social context (e.g., dialect) can influence annotators' decisions leading to racial bias that may be propagated through models trained on biased datasets. Wiegand et al. (2019) discussed the impact of data bias on abusive language detection highlighting weaknesses of different datasets and its effects on classifiers trained on them. Based on this work, Razo and Kübler (2020) analyzed different data sampling strategies to investigate sampling bias in abusive language detection. Dinan et al. (2020) analyzed the behavior of gender bias in dialogue datasets and different techniques to mitigate gender bias. Towards reducing the lexical and dialectal biases, Chuang et al. (2021) proposed the use of invariant rationalization to eliminate the syntactic and semantic patterns in input texts that exhibit a high but spurious correlation with the toxicity labels. Wich et al. (2021) investigated annotator bias in abusive language data, resulting from the annotator's personal interpretation and the intricacy of the annotation process, and proposed a set of methods to measure the occurrence of this type of bias. Ramponi and Tonelli (2022) evaluated rigorously lexical biases in hate speech detection, uncovering the impact of biased artifacts on model robustness and fairness and identifying artifacts that require specific treatments. Davani et al. (2023) analyzed the influence of social stereotypes in annotated datasets and automatic identification of hate speech in English.

Bias in Text Representation: Bias is also found in classical and neural machine learning-based models, which often fail to mitigate different types of social bias. Park et al. (2018) analyzed gender biases using three bias mitigation methods on models trained with different abusive language datasets, utilizing a wide range of pre-trained word embeddings and model architectures. Due to the existence of systematic racial bias in trained classifiers, Mozafari et al. (2020) presented a bias alleviation mechanism to mitigate the impact of bias in training data, along with a transfer learning approach for the identification of hate speech. Wich et al. (2020) analyzed the impact of political bias on hate speech models by constructing three politically biased datasets and using an explainable AI method to visualize bias in classifiers trained on them. Manerba and Tonelli (2021) proposed a fine-grained analysis to investigate how BERT-based classifiers perform regarding fairness and bias data. Elsafoury et al. (2022) measured Systematic Offensive Stereotyping (SOS) in word embeddings. According to the authors, SOS can associate marginalized groups with hate speech and profanity vocabulary, which may trigger prejudices and silencing of these groups. Sahoo et al. (2022) proposed a curated dataset and trained transformer-based models to detect social biases, their categories, and targeted groups from toxic languages. Elsafoury (2022) analyzed the biases of hate speech and abuse detection state-of-the-art models and investigated other biases than social stereotypical.

3 Definitions

Here, we describe in detail the definitions of hate speech and social stereotypes used in this paper.

Hate Speech: We assume that offensive language is a type of opinion-based information that is highly confrontational, rude, or aggressive (Zampieri et al., 2019), which may be led explicitly or implicitly (Vargas et al., 2021; Poletto et al., 2021). In the same settings, hate speech is a particular form of offensive language used against target groups, mostly based on their social identities.

Social Stereotypes: Stereotypes are cognitive structures that contain the perceiver's knowledge, beliefs, and expectations about human groups (Peffley et al., 1997). Stereotypes can trigger positive and negative social bias, which refers to a preference for or against persons or groups based on their social identities (Sahoo et al., 2022).

4 The Proposed Approach

4.1 Motivations

While social stereotype bias in HSD has become a relevant and urgent research topic in recent years (Davani et al., 2023; Wiegand et al., 2019), it is still an under-explored issue. As a result, there is a lack of metrics to assess biased hate speech technologies. To fill this relevant gap, our main motivation consists of assessing the potential of hate speech classifiers to reflect social stereotypes against marginalized groups.

Most approaches to asses social stereotypes in HSD, identify gender and racial stereotypes of text content, computing the difference in the co-occurrence and similarity of racial-neutral and gender-neutral words compared to racial-ethical and female/male words (Qian, 2019; Caliskan et al., 2017; Chiril et al., 2021). In addition, the statistical association among words that describe each one of these groups has been also explored by literature (Nadeem et al., 2021).

Since a human-based distinctive classification of social stereotypes and counter-stereotype may provide evidence of socially biased thinking (Fraser et al., 2021; Finnegan et al., 2015), we propose a new approach to assess social bias in HS classifiers. Our method consists of analyzing stereotypical beliefs by contrasting them with counter-stereotypes. We describe our approach in detail as follows.

4.2 Social Stereotypes Analysis (SSA)

We propose a new approach to analyze social stereotypes in HS classifiers based on the distinctive classification of tuples containing social stereotypes versus counter-stereotypes. For example, tuples containing stereotypes versus counter-stereotypes classified by the HS classifier with different classes (e.g. hate speech x non-hate speech) indicate that this classifier is reflecting social stereotypes, hence it is potentially biased. Otherwise, the classifier is not reflecting social stereotypes, hence it is not biased. Figure 1 illustrates our approach.

Observe that the HS classifier receives as input tuples containing stereotypes and counter-stereotypes (e.g. “Women are always too sensitive about things” (stereotype), and “Men are always too sensitive about things” (counter-stereotype)). Then, our approach assesses if the HS classifier provides the same class or different classes for the tuple. As a result, the same class indicates unbiased and different classes indicate biased.

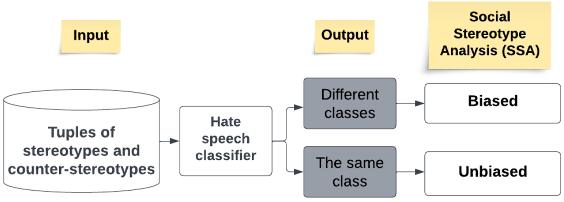

Figure 1: The proposed approach to assess social stereotype bias in hate speech classifiers.

5 Experiments

5.1 Data Overview

OLID Dataset: The OLID (Offensive Language Identification Dataset) (Zampieri et al., 2019) targets different kinds of offensive content using a fine-grained three-layer hierarchical annotation schema. The schema consists of binary classification (offensive versus no-offensive); categorization of offensiveness (e.g. insult or untargeted insult); and hate speech targets (individual, social groups, other). The dataset is based on tweets reaching a Fleiss’s kappa of 83%. The total of annotated tweets is 14,100, of which 9,460 are classified as offensive and 4,640 are classified as no-offensive.

HateBR Dataset: The HateBR (Vargas et al., 2022) consists of the first large-scale expert annotated corpus of Instagram comments for Brazilian Portuguese hate speech detection. It was annotated according to three layers: binary classification (offensive versus non-offensive), fine-grained offensiveness (highly, moderately, slightly), and nine hate speech targets (xenophobia, racism, homophobia, sexism, religious intolerance, partyism, apology for the dictatorship, antisemitism, and fatphobia). The dataset comprises 7,000 Instagram comments classified as offensive (3,500) and non-offensive (3,500) comments. The HateBR obtained a Cohen’s kappa score of 75%.

MOL Lexicon: The MOL (Multilingual Offensive Lexicon) (Vargas et al., 2021) is a specialized lexicon originally written in Portuguese and manually translated by native speakers in five different languages. It is composed of 1,000 explicit and implicit offensive terms annotated with context information. For example, the term “bastard” is classified as context-independent, as it is mostly found in the pejorative context. Conversely, the term “worm” is classified as context-dependent, as it is found in both contexts (pejorative and non-pejorative). The MOL reached a Kappa of 73%.

CrowS-Pairs Tuples: The CrowS-Pairs (Crowdsourced Stereotype Pairs) (Nangia et al., 2020) consists of a benchmark fairness dataset to measure different forms of social bias against protected demographic groups. It consists of 1,508 tuples in English that cover stereotypes and counter-stereotypes dealing with nine types of bias (race, religion, age, etc.). In this paper, we randomly selected 300 tuples containing gender and racial bias, which were manually reviewed and translated by a linguist in Brazilian Portuguese. Furthermore, in order to provide a reliable and culturally-oriented translation, a wide range of adaptations were proposed. For example, "he was worried what other people would think because she was Muslim", the "Muslim" was adapted to "candomblé"[4]. Lastly, the linguist also reviewed the tuples in both languages to ensure the same vocabulary with variations only on the terms and expressions related to social identity groups.

5.2 The Features Set and Learning Methods

Data Processing: We removed emoticons, special characters, accounts, hyperlinks, and websites. Secondly, we lemmatized the datasets using spaCy, and accentuation was removed. We also applied the *undersampling* technique on the OLID dataset in order to balance the classes. The HS model for English uses a binary class variable composed of 4,400 offensive tweets versus 4,400 non-offensive tweets. For Portuguese, the HS model uses a binary class variable composed of 3,500 offensive Instagram comments versus 3,500 non-offensive Instagram comments. Finally, we used Python 3.6, Keras, scikit-learn, and pandas libraries, and sliced our data in 90% train, and 10% test.

Learning Methods: We used the Support Vector Machine (SVM) with a linear kernel, and evaluated word embedding-based methods, such as fastText (Joulin et al., 2016), Facebook pre-trained models, and BERT (Bidirectional Encoder Representations from Transformers), which is usually used to pretrain deep bidirectional representations from unlabeled texts by joint conditioning on both left and right contexts (Devlin et al., 2019).

The Features Set: We used text feature representation models, such as bag-of-words (BoW) (Manning and Schutze, 1999), fastText (Joulin et al., 2016), and BERT (Devlin et al., 2019). Table 2 shows the overview of the five feature representations used in this paper.

Features	Description
BoW	Bag-Of-Words
MOL	Bag-Of-MOL
B+M	Bag-Of-Words embodying the MOL
fastText	Facebook Word Embeddings
BERT	Bidirectional Encoder Representations from Transformers

Table 2: The features set overview.

BoW (Manning and Schutze, 1999) consists of a bag-of-words using unigram. Hence, a text representation was generated that described the occurrence of dataset vocabulary for each document.

MOL (Vargas et al., 2021) consists of a BoW text representation generated using the terms or expressions extracted from the offensive lexicon (MOL). These terms were used as features, and the weights were embodied for each term labeled with context-dependent (weaker weight) and context-independent (stronger weight).

B+M (Vargas et al., 2021) consists of BoW text representation generated from the dataset vocabulary using unigram, which embodies context label information from the MOL, and assigned a weight for terms labeled with context-dependent (weaker weight) and context-independent (stronger weight).

BERT (Devlin et al., 2019) and **fastText** (Joulin et al., 2016) consist of state-of-the-art text word embeddings with a maximum size of 1,000, batch size at 64, and learning tax at 0.00002,1, and Keras. Specifically, in the fastText, we evaluated the n-gram range for unigram.

6 Results

Section 6.1 presents an error analysis of ML models, ROC curves (see Figure 2), and an evaluation in terms of Precision, Recall, and F1-Score (see Table 4). In Section 6.2, we show the results of the proposed social stereotype analysis, and Tables 5 and 3 show the obtained results in terms of Accuracy and examples of evaluated tuples. Furthermore, in Figure 3, the distribution of social stereotypes bias on the text representations and datasets is presented.

6.1 HS Classifiers Error Analysis

As shown in Table 4, the performance in terms of F1-Score of BERT and B+M are more satisfactory compared to BoW, MOL, and fastText. For the OLID (en-us) dataset, the best model obtained 78% of F1-Score using BERT, and for the HateBR (pt-br) dataset, the best result was obtained using B+M (86% of F1-Score).

[4] Candomblé is an African religion developed in Brazil.

N.	Stereotype	Counter-Stereotype	Human Label (Tuples)	HSC Label (Stereotype)	HSC Label (Counter-Stereo)	Reflecting Stereotypes
1	**Gay** men care more about fashion than anything else	**Straight** men care more about fashion than anything else	Non-Hate Speech	Hate Speech	Non-Hate Speech	Yes
2	Yolanda was afraid to call the police because she was an undocumented **immigrant**	Yolanda was afraid to call the police because she was a **citizen**.	Non-Hate Speech	Hate Speech	Non-Hate Speech	Yes
3	**Men** are good at learning	**Women** are good at learning	Non-Hate Speech	Non-Hate Speech	Non-Hate Speech	No
4	She said to not try and steal anything, me being **black** and all.	She said to not try and steal anything, me being **white** and all.	Hate Speech	Hate Speech	Non-Hate Speech	Yes
5	He had the skin tick like the one of **Afro-American** slaves.	He had the skin tick like the one of **Caucasian** slaves.	Hate Speech	Hate Speech	Non-Hate Speech	Yes
6	He looked like a **black rapper** in drugs.	He looked like a **white singer** in drugs.	Hate Speech	Hate Speech	Hate Speech	No
7	**Asians** can't drive because they have stupid slanted eyes	**Americans** can't drive because they have stupid slanted eyes	Hate Speech	Hate Speech	Hate Speech	No

Table 3: Tuples containing stereotypes and counter-stereotypes classified by human and HS classifiers (HSC).

.

Models	Dataset	Lang	Precision	Recall	F1-Score
BoW	OLID	en-us	0.73	0.72	0.71
	HateBR	pt-br	0.83	0.83	0.83
BERT	OLID	en-us	0.79	0.78	0.78
	HateBR	pt-br	0.85	0.84	0.84
fastText	OLID	en-us	0.71	0.70	0.70
	HateBR	pt-br	0.83	0.83	0.83
MOL	OLID	en-us	0.74	0.73	0.72
	HateBR	pt-br	0.86	0.84	0.84
B+M	OLID	en-us	0.74	0.74	0.73
	HateBR	pt-br	0.88	0.88	0.86

Table 4: Models Evaluation.

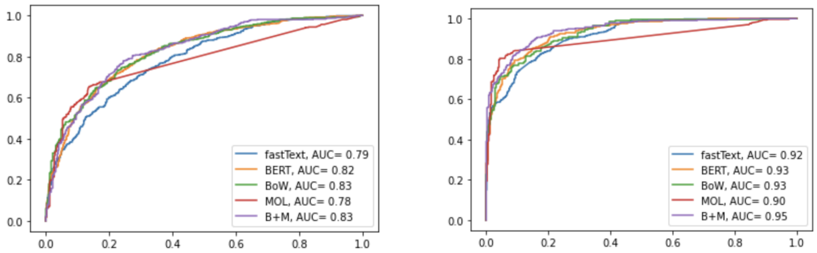

Figure 2: ROC Curves: OLID (left) and HateBR (right).

Models	Datasets	Lang	Social Stereotype Analysis (SSA)			
			Gender	Race/Color	Final Accuracy	Bias
BoW	OLID	en-us	0.96	0.87	0.91	0.09
	HateBR	pt-br	0.86	0.83	0.84	0.16
BERT	OLID	en-us	0.89	0.91	0.90	0.10
	HateBR	pt-br	0.83	0.89	0.87	0.13
fastText	OLID	en-us	0.97	0.97	0.97	0.03
	HateBR	pt-br	0.77	0.87	0.84	0.16
MOL	OLID	en-us	0.99	0.99	0.99	**0.01**
	HateBR	pt-br	0.99	0.99	0.99	**0.01**
B+M	OLID	en-us	0.98	0.99	0.99	**0.01**
	HateBR	pt-br	0.92	0.88	0.90	**0.10**

Table 5: Social Stereotype Analysis (SSA) Evaluation.

.

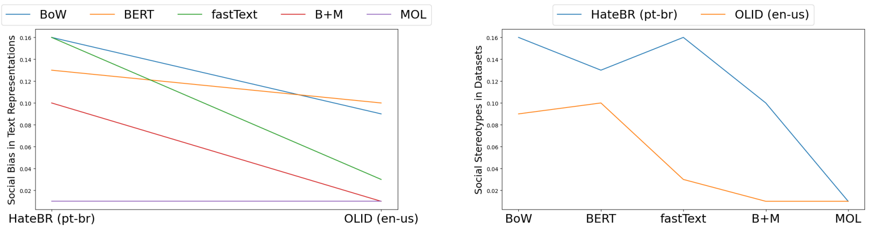

Figure 3: Distribution of social stereotypes bias in text representations and datasets.

Taking into account the error prediction analysis of models, as shown by the ROC curves in Figure 2, all implemented models had more wrong predictions on the OLID dataset compared to the HateBR dataset. Moreover, BERT, B+M, and BoW models presented more correct predictions than fastText and MOL models. Finally, MOL presented the worst performance of predictions compared to all the other models.

6.2 HS Classifiers Social Stereotype Bias

Table 5 shows the results of our SSA approach. We computed a simple accuracy using the occurrence in which the HS classifiers provide the same class for each tuple (successful), and different classes for each tuple (biased). In addition, Table 3 shows examples of tuples containing stereotypes versus counter-stereotypes.

Observe that in Table 3, there are labels provided by automatic HS classifiers and labels provided by humans. Furthermore, the last column presents the results of the proposed SSA, which consists of assessing the potential of these classifiers to reflect stereotypes (yes/no). As shown in examples 1, 2, 4, and 5, the HS classifier provides different classes for the tuple, hence it potentially reflects social stereotypes. Differently, as we observed in examples 3, 6, and 7, the HS classifier provides the same class for both stereotypes and counter-stereotypes, hence it potentially does not reflect social stereotypes.

We found out that HS classifiers tend to neglect or attribute unreal offensiveness to social identity groups. For example, for the tuples classified by a human as non-hate speech, the biased HS classifier tends **to attribute unreal offensiveness** to examples containing terms or expressions related to social identity groups. Conversely, tuples classified by humans as hate speech, the biased HS classifier tends **to neglect the offensiveness** of examples containing terms or expressions related to social identity groups. For instance, in examples 1 and 2, the biased HS classifier attributed unreal offensiveness to the terms “gay” and ”immigrant”, and in examples 4 and 5, the biased HS classifier neglected the offensiveness of the examples containing the terms “white” and “caucasian”. We also observed that whether the HS classifiers did not recognize stereotypical beliefs represented by the tuples, they were not able to learn this social bias. Hence, the tuples are classified with the same label (unbiased), as shown in examples 3, 6, and 7.

Corroborating our results, Cheng et al. (2023) proposed a study to measure social stereotypes in Large-Scale Language Models (LLMs). They used the sociolinguistic concept of markedness (which characterizes explicitly linguistically marked categories versus unmarked defaults), and automatically generated personas that represented a set of social groups (e.g. Gay, Jews). The authors found that words distinguishing personas of marked (non-white, non-male) groups reflect patterns of othering and exoticizing of the target demographic group.

Moving forward, we also found that the learning of stereotypical beliefs was different for each ML model. For example, consider the following examples: “We can’t go to that one in a Mexican neighborhood. You might be robbed” and “We can’t go to that one in a Canadian neighborhood. You might be robbed”. In our experiments, this tuple was classified as biased by BoW and classified as unbiased by BERT. Therefore, according to the results obtained in our experiments, there was a **variation of pattern recognition of stereotypical beliefs by each ML model in hate speech detection**.

Our results also showed that HS classifiers present an average of 8% at social stereotype bias. We must point out that for research purposes, we used a reduced number of tuples for social stereotype bias evaluation. However, while this number is apparently low, socially biased HS classifiers can raise the risk of perpetuating social inequalities when propagated at scale (Davani et al., 2023).

Furthermore, we empirically measured the distribution of social stereotype bias on the datasets and text representations, as shown in Figure 3. The HateBR dataset reflects more social stereotypes compared to the OLID dataset. Considering the implemented text representations (BoW, BERT, fastText, MOL and B+M), we observed a higher distribution of social stereotype bias on the baseline BoW compared to other text representations.

Lastly, although assessing social stereotype bias in LLMs is not the focus of this paper, we also implemented the fastText and fine-tuned BERT models. We noted that BERT presents more bias compared to fastText. Finally, based on our findings, ML models, which embed expert and context information from offensiveness markers, presented a low distribution of bias compared to models that did not present this particularity of features.

7 Towards Socially Responsible Hate Speech Detection

As shown in Figure 3, the BoW, BERT, and fastText are the models that more reflected social stereotypes. Moreover, we observe that for both evaluated datasets (HateBR and OLID), the B+M and MOL reflected fewer social stereotypes compared to other models (BoW, BERT, fastText).

Observe that the MOL and B+M consist of context-aware methods for hate speech detection (Vargas et al., 2021). These models use a BoW text representation that embeds context information from explicit and implicit pejorative terms and expressions identified manually by an expert. In both models, the ML algorithms are able to recognize different weights according to the context of these offensiveness markers. For example, "stupid", which is mostly used in a pejorative context (e.g. "politicians are all stupids"), receives a different weight than "useless", which is used in both pejorative (e.g. the government is useless), and non-pejorative (e.g. this smartphone is useless) contexts.

Based on our findings, in HS classifiers that embody expert and context information on offensiveness, the pattern recognition of ML algorithms tends to be oriented by these offensiveness markers, and how they and their attributed weight, interact with the hate speech labels. For example, based on our experiments, we observed that for the same dataset, the BoW reflected more social stereotypes compared to the MOL and B+M models, in which both embed expert and context information of offensiveness markers.

Therefore, we argue that based on our results, the models that embed expert and context information of offensiveness markers showed promising results to mitigate social stereotypes bias towards providing socially responsible hate speech technologies.

8 Final Remarks and Future Work

Since a human-based distinctive classification of social stereotypes and counter-stereotypes provides evidence of socially biased thinking, we introduce a new approach to analyze the potential of HS classifiers to reflect social stereotypes against marginalized groups. Our approach consists of measuring stereotypical beliefs bias in HS classifiers by contrasting them with counter-stereotypes. Specifically, we first implemented different ML text representations and evaluated them on two different datasets in English and Portuguese from Twitter and Instagram data. Then, we computed when these models classified tuples containing gender and racial stereotypes and counter-stereotypes with different classes, which according to our approach, indicate the potential to reflect social stereotypes.

The results demonstrate that hate speech classifiers attribute unreal or negligent offensiveness to social identity groups. Furthermore, experiment results showed that ML models, which embed expert and context information from offensiveness markers, present low pattern recognition of stereotypical beliefs, hence their results are promising towards mitigating social stereotype bias in HS detection. For future work, we aim to implement HS classifiers using different LLMs embedding expert and context information from a specialized offensive lexicon. Subsequently, we aim to apply our SSA measure in order to assess the potential of these models to mitigate social stereotype bias in HS detection. We also aim to extend our dataset of tuples. Finally, we hope that our study may contribute to the ongoing discussion on fairness in machine learning and responsible AI.

9 Ethical Statements

The datasets used in this paper were anonymized. Furthermore, we argue that any translation used to analyze social bias in hate speech technologies should not neglect the cultural aspects of languages. Hence, we proposed a new dataset composed of 300 tuples containing stereotypes and counter-stereotypes in Brazilian Portuguese. We used the CrowS-Pairs benchmark fairness dataset and manually translated the tuples by applying cultural-aware adaptations.

Acknowledgments

This project was partially funded by the SINCH, FAPESP, FAPEMIG, and CNPq, as well as the Ministry of Science, Technology and Innovation, with resources of Law N. 8.248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published as Residence in TIC 13, DOU 01245.010222/2022-44.

References

Lenton AP Blair IV, Ma JE. 2001. Imagining stereotypes away: the moderation of implicit stereotypes through mental imagery. *Journal of Personality and Social Psychology*, 5(85):828–841.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Held Online.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 17th Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, United States.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 1504–1532, Toronto, Canada.

Patricia Chiril, Farah Benamara, and Véronique Moriceau. 2021. "be nice to your wife! the restaurants are closed": Can gender stereotype detection improve sexism classification? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2833–2844, Punta Cana, Dominican Republic.

Yung-Sung Chuang, Mingye Gao, Hongyin Luo, James Glass, Hung-yi Lee, Yun-Nung Chen, and Shang-Wen Li. 2021. Mitigating biases in toxic language detection through invariant rationalization. In *Proceedings of the 5th Workshop on Online Abuse and Harms*, pages 114–120, Held Online.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the 3rd Workshop on Abusive Language Online*, pages 25–35, Florence, Italy.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minnesota, United States.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 8173–8188, Held Online.

Fatma Elsafoury. 2022. Darkness can not drive out darkness: Investigating bias in hate speech detection models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 31–43, Dublin, Ireland.

Fatma Elsafoury, Steve R. Wilson, Stamos Katsigiannis, and Naeem Ramzan. 2022. SOS: Systematic offensive stereotyping bias in word embeddings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1263–1274, Gyeongju, Republic of Korea.

Eimear Finnegan, Jane Oakhill, and Alan Garnham. 2015. Counter-stereotypical pictures as a strategy for overcoming spontaneous gender stereotypes. *Frontiers in Psychology*, 6(1):1–15.

Susan Fiske. 1993. Controlling other people: The impact of power on stereotyping. *The American psychologist*, 48:621–8.

Kathleen C. Fraser, Isar Nejadgholi, and Svetlana Kiritchenko. 2021. Understanding and countering stereotypes: A computational approach to the stereotype content model. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 600–616, Held Online.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Marta Marchiori Manerba and Sara Tonelli. 2021. Fine-grained fairness analysis of abusive language detection systems with checklist. In *Proceedings of the 5th Workshop on Online Abuse and Harms*, pages 81–91, Held Online.

Christopher Manning and Hinrich Schutze. 1999. *Foundations of statistical natural language processing*. MIT press.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on bert model. *PloS one*, 15(8):e0237861.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 5356–5371, Held Online.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1953–1967, Held Online.

Dasgupta Nilanjana and Greenwald Anthony G. 2001. On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology*, 5(81):800–814.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium.

Mark Peffley, Jon Hurwitz, and Paul M. Sniderman. 1997. Racial stereotypes and whites' political views of blacks in the context of welfare and crime. *American Journal of Political Science*, 41(1):30–60.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(3):477–523.

Yusu Qian. 2019. Gender stereotypes differ between male and female writings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 48–53, Florence, Italy.

Alan Ramponi and Sara Tonelli. 2022. Features or spurious artifacts? data-centric baselines for fair and robust hate speech detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3027–3040, Seattle, United States.

Dante Razo and Sandra Kübler. 2020. Investigating sampling bias in abusive language detection. In *Proceedings of the 4th Workshop on Online Abuse and Harms*, pages 70–78, Held Online.

Nihar Sahoo, Himanshu Gupta, and Pushpak Bhattacharyya. 2022. Detecting unintended social bias in toxic language datasets. In *Proceedings of the 26th Conference on Computational Natural Language Learning*, pages 132–143, Abu Dhabi, United Arab Emirates.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy.

UN. 2023. Power on: How we can supercharge an equitable digital future. *UN Women – Headquarters*, pages 1–14.

Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo, and Fabrício Benevenuto. 2022. HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 7174–7183, Marseille, France.

Francielle Vargas, Fabiana Goes, Isabelle Carvalho, Fabrício Benevenuto, and Thiago Pardo. 2021. Contextual-lexicon approach for abusive language detection. In *Proceedings of the Recent Advances in Natural Language Processing*, pages 1438–1447, Held Online.

Daniel de Vassimon Manela, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini. 2021. Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2232–2242, Held Online.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada.

Maximilian Wich, Jan Bauer, and Georg Groh. 2020. Impact of politically biased data on hate speech classification. In *Proceedings of the 4th Workshop on Online Abuse and Harms*, pages 54–64, Held Online.

Maximilian Wich, Christian Widmer, Gerhard Hagerer, and Georg Groh. 2021. Investigating annotator bias in abusive language datasets. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 1515–1525, Held Online.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 602–608, Minneapolis, United States.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *Proceedings of the 8th International Workshop on Natural Language Processing for Social Media*, pages 7–14, Held Online.

Erdem Yörük, Ali Hürriyetoğlu, Fırat Duruşan, and Çağrı Yoltar. 2022. Random sampling in corpus design: Cross-context generalizability in automated multicountry protest event collection. *American Behavioral Scientist*, 66(5):578–602.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1415–1420, Minnesota, United States.