NoHateBrazil: A Brazilian Portuguese Text Offensiveness Analysis System

Francielle Vargas[1,2], **Isabelle Carvalho**[1], **Wolfgang S. Schmeisser-Nieto**[3]
Fabrício Benevenuto[2], **Thiago A. S. Pardo**[1]
[1]Institute of Mathematical and Computer Sciences, University of São Paulo, Brazil
[2]Computer Science Department, Federal University of Minas Gerais, Brazil
[3] Department of Linguistics, University of Barcelona, Spain
francielleavargas@usp.br, isabelle.carvalho@alumni.usp.br
wolfgang.schmeisser@ub.edu, fabricio@dcc.ufmg.br, taspardo@icmc.usp.br

Abstract

Hate speech is a surely relevant problem in Brazil. Nevertheless, its regulation is not effective due to the difficulty to identify, quantify and classify offensive comments. Here, we introduce a novel system for offensive comment analysis in Brazilian Portuguese. The system titled *NoHateBrazil*[1] recognizes explicit and implicit offensiveness in context at a fine-grained level. Specifically, we propose a framework for data collection, human annotation and machine learning models that were used to build the system. In addition, we assess the potential of our system to reflect stereotypical beliefs against marginalized groups by contrasting them with counter-stereotypes. As a result, a friendly web application was implemented, which besides presenting relevant performance, showed promising results towards mitigation of the risk of reinforcing social stereotypes. Lastly, new measures were proposed to improve the explainability of offensiveness classification and reliability of the model's predictions.

1 Introduction

The scenario of hateful comments in Brazil is severe and entails the creation of safety and fairness technologies. During the elections in 2018 and 2022, the denunciations against xenophobia content had an increase of 2,369.5%; apology and public incitement to violence and crimes against life, 630.52%, and misogyny and race-ethical, increased by 1,639% and 595%[2], respectively.

Hate speech is a particular form of offensive language that considers stereotypes to express an ideology of hate (Warner and Hirschberg, 2012; Sahoo et al., 2022; AlKhamissi et al., 2022). While systems that classify hateful content are undoubtedly relevant, these technologies are being developed with scarce consideration of their potential biases (Nadeem et al., 2021; Sap et al., 2019; Chang et al., 2019; Bordia and Bowman, 2019; Blodgett et al., 2020). These systems may discriminate against the groups they are designed to protect (Davidson et al., 2019), reflecting social stereotypes and being able to perpetuate social inequalities when propagated at scale (Davani et al., 2023).

To the best of our knowledge, no systems have attempted to analyze text offensiveness in Brazilian Portuguese. Therefore, the main contribution of this paper[3] is providing the first web system titled *NoHateBrazil* for Brazilian Portuguese offensive comments classification. The *NoHateBrazil* system receives two different inputs. The first input consists of a single comment written directly into the initial screen. The second input consists of a file in CSV format containing a set of comments. In the following outputs, three pieces of information are exhibited: (i) offensiveness categories; (ii) offensiveness overall score; and (iii) prediction reliability score, which we describe in Section 2.1.

Towards providing a reliable text offensiveness system, we focus on three strong strategies: (i) we provide a contextualized analysis of offensiveness, in which Machine Learning (ML) models recognize explicit and implicit offensive terms from a specialized lexicon annotated with context information; (ii) we propose and evaluate a framework for offensive comment detection; (iii) we evaluate the potential of our system to reflect social stereotypes through a distinctive analysis of tuples containing stereotypes versus counter-stereotype (Vargas et al., 2023). For this purpose, we used a dataset of 300 tuples containing social stereotypes versus counter-stereotypes in Brazilian Portuguese, which consists of a culturally-oriented translation from the CrowS-Pairs (Nangia et al., 2020), a benchmark fairness dataset. Finally, our system presents 88.8% of F1-Score and a low potential of reflecting social stereotypes against marginalized groups (12%).

[1]**Demo**: http://143.107.183.175:14581/
[2]https://new.safernet.org.br/

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

Proceedings of Recent Advances in Natural Language Processing, pages 1180–1186
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_125

2 Offensiveness Detection Framework

In this paper, we propose a new framework that encompasses data collection, human annotation, and the implementation of ML models for offensive comment detection. We used this framework to build the proposed *NoHateBrazil* web system, as shown in Figure 1.

- **Data Collection:** Given the relevance of collecting representative data, we propose a careful data collection approach composed of balanced attributes, as shown in Figure 1. Note that for each profile ***P*** from a domain ***D***, the number of comments must be balanced. For synchronous bordering, which consists of data collection during a period of time ***T***, the same number of comments must be collected for each span of time. For example, we implemented an Instagram API and collected the maximum number of 500 comments per post. We also balanced profile attributes (gender, color, political party). For data cleaning, we removed noise, such as links, and characters without semantic value, and also comments that presented only emoticons, laughs (e.g., kkk), or mentions (e.g., @fulano), without any textual content, and then applied data anonymization.

- **Annotation Process**: In spite of the enormous difficulty of automatically classifying offensive comments mainly due to ethical problems, the annotation process should be carried out by specialists (Vargas et al., 2021). As shown in Figure 1, the annotation process consists of three main stages. Firstly, the selection of expert annotators, considering their diverse profiles, such as ethnicity, gender, different political orientations, and place of origin. Secondly, the creation of a well-structured annotation schema. Lastly, evaluation metrics were applied, as Kappa and Fleiss, reaching a high inter-annotator agreement (75% Kappa and 74% Fleiss). This evaluation is fundamental to ensure data quality. The entire data collection and annotation process is described in detail in Vargas et al. (2022).

- **Context-Aware Language Models**: Large crowd-sourced lexical resources tend to include a wide range of irrelevant terms, resulting in high rates of false positives (Davidson et al., 2019). Moreover, pre-trained language models are trained on large real-world data. As a result, they are known to embody social biases (Nadeem et al., 2021). According to Davidson et al. (2019), it is possible to mitigate social bias by focusing on how context factors interact with linguistic subtleties and the definitions of offensive language. In addition, social bias decreases in magnitude when it is conditioned on particular terms and expressions that may indicate membership in negative classes. Accordingly, we assume that context information is a relevant attribute to classify offensiveness in text. Hence, we propose a computational context-aware ML model that embodies implicit and explicit offensive terms and expressions annotated manually by experts with context information. The implemented ML model, titled "B+M" is described in detail in Vargas et al. (2021). We shortly present below.

  B+M: This model uses a generated bag-of-words (BoW) from the dataset vocabulary. This model embodies labeled context information (context-dependent and context-independent) from a specialized lexicon of explicit and implicit offensive terms and expressions called *MOL* (see Section 3.1). We carried out the match with terms from MOL. Then, we assigned a weight for each term or expression labeled with context-dependent (weaker weight), and context-independent (stronger weight). According to the B+M model, the value of a term x in the document y is defined as

$$B + M_{x,y} = freq_{x,y} * weightC_x \quad (1)$$

  where $freq$ is the frequency of the term in the document, $weightC = 2$ for context-dependent terms and $weightC = 3$ when the term is context-independent.

2.1 Text Offensiveness Analysis

According to Poletto et al. (2021), Offensive language Detection (OLD) often leads to false positives when swear and offensive words occur in non-offensive contexts. Furthermore, OLD mainly presents explicit and implicit terms or expressions with pejorative connotations, and the pejorative connotation is deeply context-dependent and culturally oriented (Vargas et al., 2021).

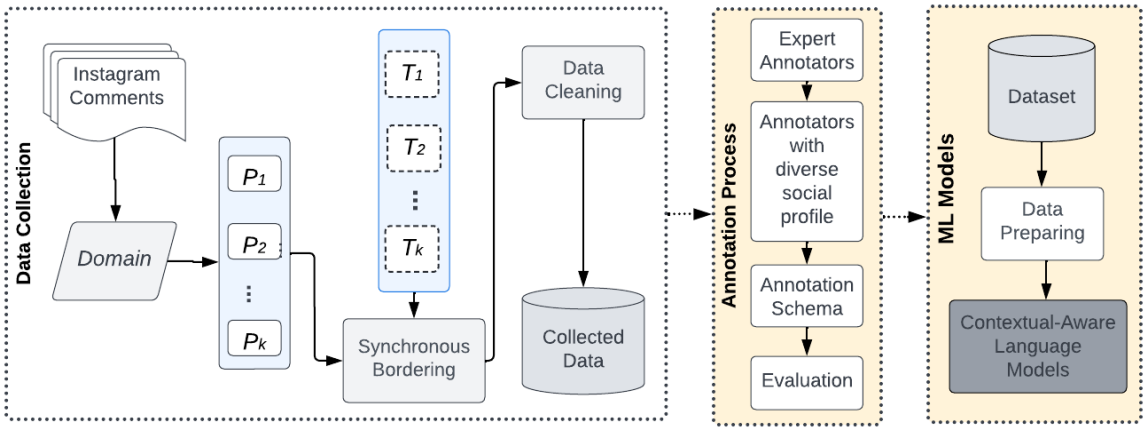

Figure 1: The proposed framework for offensive comment classification.

Corroborating the offensiveness definitions proposed by Caselli et al. (2020), our system assumes that **explicit offensiveness** consists of comments that contain explicit markers of offensiveness (e.g. comments with terms or expressions with any pejorative connotations). Conversely, **implicit offensiveness** consists of comments that contain markers of offensive content expressed implicitly. Both examples are shown in Table 1, as well as an example of a non-offensive comment. Note that bold indicates markers of implicit offensive content, and underlines explicit markers of offensiveness.

Class	Comments	Translation
Offensive	Essa besta humana é o câncer do País, tem q **voltar p jaula**, urgentemente! E viva o Presidente Bolsonaro.	This animal is the cancer of the country, it has to go **back to jail** as soon as possible! And cheers to President Bolsonaro[4]
Offensive	Pois é, deveria **devolver o dinheiro** aos cofres públicos do Brasil. Canalha.	That's right, he should **refund the money** to the public Brazilian banks. Jerk.
Non-Offensive	Quem falou isso pra vc deputada? O Sergio Moro ta aprovado pela maioria dos brasileiros.	Who said that to you, congresswoman? Sergio Moro[5] has the approval of most Brazilians.

Table 1: Offensive and non-offensive comments with explicit and implicit offensiveness.

Our system also recognizes **context information** using an offensive lexicon annotated by specialists with context information. For instance, while the terms "cancer", "garbage", and "worms" may be used with pejorative connotations, they could also be used in contexts without any pejorative connotation (e.g., "he was cured of cancer"; "the garden is full of parasites and worms"; "disposal of garbage on streets"). In this case, these terms are classified as context-dependent. Differently, the terms "hypocritical" and "ridiculous" are mostly used in contexts with pejorative connotations. Consequently, these terms are classified as context-independent.

2.1.1 Offensiveness Overall Score (OOS)

In order to present explainability for offensive comments classification at a fine-grained level, as well as to provide a more accurate prediction of offensiveness, we propose a measure titled *Offensiveness Overall Score (OOS)*. The OOS combines expert and statistical knowledge in order to classify offensive comments on three different levels: slightly, moderately, and highly. Specifically, this score consists of a scale between 0 and 100 that combines a set of parameters defined by different specialists in Vargas et al. (2022) and a probability score. In this paper, we called $score_{expert}$ the parameters provided by experts, along with the prediction probability value provided by the ML model, which we called $score_{prob}$. The OSS is defined by Equation 2.

$$OOS = (score_{expert} + score_{prob}) \div 2 \quad (2)$$

As regards the $score_{expert}$, comments with at least 1 (one) MOL term annotated with the context-independent label (mol_{indep}), or at least 3 (three) MOL terms annotated with the context-dependent labels (mol_{dep}), should receive a $score_{expert}$ of 90%. In the same settings, comments that precisely present 2 (two) MOL terms annotated with

the context-dependent label (mol_{dep}), should receive a $score_{expert}$ of 60%; and comments that precisely present 1 (one) MOL term annotated with the context-dependent label (mol_{dep}), should receive a $score_{expert}$ of 30%. Algorithm 1 shows the proposed offensiveness overall score. As regards the $score_{expert}$, the prediction probability score was obtained by the ML model. Algorithm 1 describes in detail the OOS measure. Observe that the proposed OOS provides a set of machine-learned rules, besides tackling the problem of out-of-vocabulary terms.

Algorithm 1 Offensiveness Overall Score

procedure GET-OOS($prob$)
  if $mol_{indep} >= 1$ or $mol_{dep} >= 3$ **then**
    $OOS = (90 + score_{prob}) \div 2$
  end if
  if $mol_{dep} == 2$ **then**
    $OOS = (60 + score_{prob}) \div 2$
  end if
  if $mol_{dep} == 1$ **then**
    $OOS = (30 + score_{prob}) \div 2$
  end if
  if $OOS > 0$ and $OOS <= 49$ **then**
    $class = slightly\ offensive$
  end if
  if $OOS >= 50$ and $OOS <= 79$ **then**
    $class = moderately\ offensive$
  end if
  if $OOS >= 80$ and $OOS <= 100$ **then**
    $class = highly\ offensive$
  end if
  return OOS and $class$
end procedure

2.1.2 Prediction Reliability Score (PRS)

In order to provide a robust evaluation of the quality of the model's predictions for unknown sentences (unlabeled), we further provide a measure titled *Prediction Reliability Score (PRS)*. The PRS estimates a reliability scale taking into account the statistical distribution of pejorative terms and expressions from the HateBR dataset (see Section 3.1). Specifically, this measure computes a reliability score using the difference between the values obtained from a defined reliability scale, which we called $score_{gold}$, and the values provided by $score_{prob}$, which is a statistic score of the ML model. The PRS may be defined as shown in Equation 3.

$$PRS = 100 - |(score_{gold} - score_{prob})| \quad (3)$$

As regards the PRS score, two different scales for offensive comments (class 1), and non-offensive comments (class 0) were proposed, as shown in Algorithms 2 and 3, respectively.

Algorithm 2 Prediction Reliability Score (Offensive)

1: **procedure** GET-PRS($prob$)
2: **if** $mol_{indep} >= 1$ or $mol_{dep} >= 3$ **then**
3: $score_{gold} = 99\%$
4: **end if**
5: **if** $mol_{dep} == 2$ **then**
6: $score_{gold} = 90\%$
7: **end if**
8: **if** $mol_{dep} == 1$ **then**
9: $score_{gold} = 80\%$
10: **end if**
11: **if** $mol_{indep} == 0$ and $mol_{dep} == 0$ **then**
12: $score_{gold} = 10\%$
13: **end if**
14: **return** $PRS = 100 - |(score_{gold} - (score_{prob})|$
15: **end procedure**

Algorithm 3 Prediction Reliability Score (No-Offensive)

1: **procedure** GET-PRS($prob$)
2: **if** $mol_{indep} >= 1$ or $mol_{dep} >= 3$ **then**
3: **return** $score_{gold} = 10\%$
4: **end if**
5: **if** $mol_{dep} == 2$ **then**
6: **return** $score_{gold} = 80\%$
7: **end if**
8: **if** $mol_{dep} == 1$ **then**
9: **return** $score_{gold} = 90\%$
10: **end if**
11: **if** $mol_{indep} == 0$ and $mol_{dep} == 0$ **then**
12: **return** $score_{gold} = 99\%$
13: **end if**
14: **return** $PRS = 100 - |(score_{gold} - (score_{prob})|$
15: **end procedure**

As shown in Algorithm 2, **offensive comments** with at least 1 (one) MOL term annotated with the context-independent label (mol_{indep}), or at least 3 (three) MOL terms annotated with the context-dependent labels (mol_{dep}), should receive a $score_{gold}$ of 99%; and offensive comments that precisely present 2 (two) MOL terms annotated with the context-dependent labels (mol_{dep}), should receive a $score_{gold}$ of 90%; and offensive comments that precisely present 1 (one) MOL term annotated with the context-dependent label (mol_{dep}), should receive a $score_{gold}$ of 80%. Lastly, offensive comments without any MOL term should receive a $score_{gold}$ of 10%.

As shown in Algorithm 3, **non-offensive comments** with at least 1 (one) MOL term annotated with the context-independent label (mol_{indep}), or at least 3 (three) MOL terms annotated with the context-dependent labels (mol_{dep}), should receive a $score_{gold}$ of 10%; and non-offensive comments that precisely present 2 (two) MOL terms annotated with the context-dependent labels (mol_{dep}), should receive a $score_{gold}$ of 80%; and non-offensive comments that precisely present 1 (one) MOL term annotated with the context-dependent label (mol_{dep}), should receive a $score_{gold}$ of 90%. Lastly, non-offensive comments without any MOL terms should receive a $score_{gold}$ of 99%.

3 System Design

3.1 Architecture

3.1.1 Infrastructure: The web application was developed using Python version 3.9 and the following libraries: streamlit[6], unidecode[7], emoji[8], spacy[9], gensim[10] and the Brazilian Portuguese normalizer, Enelvo[11]. It was hosted on the Apache Server.

3.1.2 Machine Learning: We built a ML model using a BoW titled "B+M" and Naive Bayes algorithm. The entire experimental settings and results are described in detail in Vargas et al. (2021). Our pre-processing required (i) data cleaning (e.g. accounts, quotes, links, and emojis), (ii) lemmatization, (iii) normalization, and (iv) accent removal.

3.1.3 Data Resources: We used two different data resources: the *HateBR dataset* (Vargas et al., 2022), which consists of the first large-scale expert annotated corpus composed of 7,000 Brazilian Instagram comments; and the *MOL - Multilingual Offensive Lexicon* (Vargas et al., 2021), which consists of a context-aware offensive lexicon composed of 1,000 explicit and implicit offensive terms and expressions manually identified by a linguist and annotated in a binary-class: context-dependent and context-independent. Furthermore, both resources provide linguistic markers of nine hate speech targets (partyism, sexism, homophobia, fatphobia, religious intolerance, apology for the dictatorship, xenophobia, antisemitism and racism).

3.2 Interface

3.2.1 Inputs: As shown in Figure 2, the user may insert two types of inputs. Firstly, the user has the option to classify **only one comment by typing it directly on the interface**. Then, the user only selects the button "Enter" to obtain the classification. Secondly, the user may classify **a set of comments in CSV format by selecting the button "browser files"**. Then, the system downloads the classified data in CSV format, and an average of the obtained classification is presented on the initial screen

3.2.2 Outputs: As shown in Figure 2, after adding the inputs by the user, the system presents three different outputs: (i) **offensiveness category** (highly, moderately, or slightly); (ii) **OOS**; and (iii) **PRS**.

[6] https://streamlit.io/
[7] https://pypi.org/project/Unidecode/
[8] https://pypi.org/project/emoji/
[9] https://spacy.io/
[10] https://radimrehurek.com/gensim/
[11] https://thalesbertaglia.com/enelvo/

4 System Evaluation

4.1 Model Performance

While a direct comparison is unfair (as the authors use different datasets and methods), we present in Table 2 an analysis of our results and baselines for the Brazilian Portuguese language.

Authors	Datasets	Methods	Fscore
NoHateBrazil	**HateBR**	**NB**	**88.8**
de Pelle and Moreira (2017)	OFFCOMBR	NB	81.0
Leite et al. (2020)	ToLD-Br	BERT	76.0
Trajano et al. (2023)	OLID-BR	BERT	77.0

Table 2: Offensive language detection baselines for Brazilian Portuguese: Results.

As shown in Table 2, even though it only offers an idea of the general performance of the approaches, the *NoHateBrazil* system presents higher performance for the offensive language detection task in Brazilian Portuguese.

4.2 Social Stereotype Analysis (SSA)

Recent studies show that hate speech technologies may propagate societal stereotypes against marginalized groups (Davani et al., 2023; Davidson et al., 2019). Hence, we also assess the potential of our system to reflect gender and racial stereotypes. We use a metric called **Social Stereotype Analysis (SSA)** (Vargas et al., 2023). The SSA measures the potential of hate speech classifiers to reflect social stereotypes by the distinctive classifying of tuples containing stereotypes versus counter-stereotypes. For example, "They've been known to disregard people who march in Gay Pride parades" and "They've been known to disregard people who march in Christmas parades.", the system should classify both with the same label. Otherwise, the system potentially reflects stereotypes against the LGBTQIA+ group. Specifically, we assess a set of 300 tuples containing gender and racial stereotypes versus counter-stereotypes in Brazilian Portuguese[12]. Results are shown in Table 3.

Tuples	Total	Accuracy
300	600	**88.0**

Table 3: SSA Evaluation.

As shown in Table 3, we classified 300 tuples (600 comments), in which **12%** of tuples were classified with different labels by our system.

[12] https://github.com/franciellevargas/SSA/tree/main/tuples/pt-br

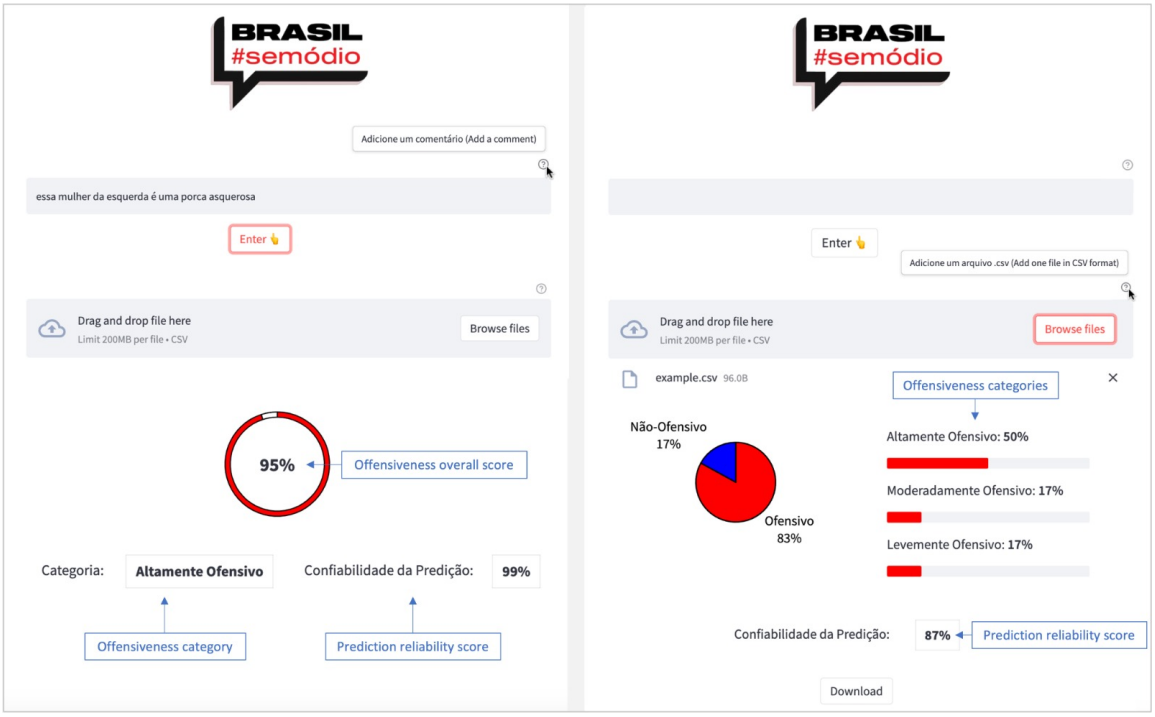

Figure 2: *NoHateBrazil* web system - input and output interfaces

.

4.3 OOS and PRS Measures

Lastly, we evaluated both proposed measures (OOS and PRS) using human evaluation[13]. In order to evaluate the OOS, we manually collected 90 new comments from Instagram divided equally among highly, moderately, and slightly offensive. For the PRS evaluation, we also collected 60 more news comments from Instagram divided equally between offensive and non-offensive comments. We followed the annotation scheme proposed by Vargas et al. (2022). Subsequently, we evaluated the predicted class compared with the human-proposed labels. Results are shown in Table 4.

Measure	Total	Accuracy
OOS	90	**70.0**
PRS	60	**89.0**

Table 4: OOS and PRS Evaluation Results.

Note that the OOS presented an accuracy of 70%, corroborating the study proposed by Vargas et al. (2022), that claim that the fine-grained offensiveness is a complex task. The PRS obtained an accuracy of 89%, highlighting the capability of our ML model to efficiently classify offensive comments.

[13] https://github.com/franciellevargas/HateBR/tree/main/NoHateBrazil/evaluation

5 Final Remarks

This paper introduces the first system for text offensiveness analysis in Brazilian Portuguese. The *NoHateBrazil* web system recognizes explicit and implicit offensiveness in context at a fine-grained level. We proposed a friendly design and robust architecture, resulting in a high system performance, besides promising results towards mitigation of the risk of perpetuating social stereotypes against marginalized groups. We also provided a robust framework for offensive comment classification, which encompasses data collection, human annotation, and ML models. Finally, two new measures were proposed to improve the explainability of offensiveness classification at a fine-grained level and the reliability of the model's predictions.

Acknowledgements

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

Badr AlKhamissi, Faisal Ladhak, Srinivasan Iyer, Veselin Stoyanov, Zornitsa Kozareva, Xian Li, Pascale Fung, Lambert Mathias, Asli Celikyilmaz, and Mona Diab. 2022. ToKen: Task decomposition and knowledge infusion for few-shot hate speech detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2109–2120, Abu Dhabi, United Arab Emirates.

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.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France.

Kai-Wei Chang, Vinodkumar Prabhakaran, and Vicente Ordonez. 2019. Bias and fairness in natural language processing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, Hong Kong, China.

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.

Rogers de Pelle and Viviane Moreira. 2017. Offensive comments in the Brazilian web: A dataset and baseline results. In *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*, pages 510–519, Rio Grande do Sul, Brazil.

João Augusto Leite, Diego Silva, Kalina Bontcheva, and Carolina Scarton. 2020. Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924, Suzhou, China.

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.

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.

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.

Douglas Trajano, Rafael Bordini, and Renata Vieira. 2023. Olid-br: offensive language identification dataset for brazilian portuguese. *Language Resources & Evaluation*, 1:1–25.

Francielle Vargas, Isabelle Carvalho, Ali Hürriyetoğlu, Thiago A. S. Pardo, and Fabrício Benevenuto. 2023. Socially responsible hate speech detection: Can classifiers reflect social stereotypes? In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, Varna, Bulgaria.

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 Rodrigues de Góes, Isabelle Carvalho, Fabrício Benevenuto, and Thiago Pardo. 2021. Contextual-lexicon approach for abusive language detection. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 1438–1447, 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.