Natural Language Processing (2024), 1–22
doi:10.1017/nlp.2024.18

ARTICLE

Context-aware and expert data resources for Brazilian Portuguese hate speech detection

Francielle Vargas[1,2], Isabelle Carvalho[1], Thiago A. S. Pardo[1] and Fabrício Benevenuto[2]

[1]Institute of Mathematical and Computer Sciences, University of São Paulo, São Carlos, Brazil and [2]Computer Science Department, Federal University of Minas Gerais, Belo Horizonte, Brazil
Corresponding author: Francielle Vargas; Email: francielleavargas@usp.br

(Received 30 November 2022; revised 11 April 2023; accepted 16 September 2023)

Special Issue on '**Natural Language Processing Applications for Low-Resource Languages**'

Abstract

This paper provides data resources for low-resource hate speech detection. Specifically, we introduce two different data resources: (i) the HateBR 2.0 corpus, which is composed of 7,000 comments extracted from Brazilian politicians' accounts on Instagram and manually annotated a binary class (offensive versus non-offensive) and hate speech targets. It consists of an updated version of the HateBR corpus, in which highly similar and one-word comments were replaced; and (ii) the multilingual offensive lexicon (MOL), which consists of 1,000 explicit and implicit terms and expressions annotated with context information. The lexicon also comprises native-speaker translations and its cultural adaptations in English, Spanish, French, German, and Turkish. Both corpus and lexicon were annotated by three different experts and achieved high inter-annotator agreement. Lastly, we implemented baseline experiments on the proposed data resources. Results demonstrate the reliability of data outperforming baseline dataset results in Portuguese, besides presenting promising results for hate speech detection in different languages.

Keywords: hate speech; Brazilian Portuguese; low-resource languages

1. Introduction

Hate speech spreading on the web and social media is an important research problem, and its regulation is still ineffective due to the high difficulty in identifying, quantifying, and classifying hateful communication. While hateful content is found in different platforms and languages, the majority of data resources are proposed for the English language. As a result, there is a lack of research and resources for low-resource hate speech detection.

The hate speech detection literature has focused on different tasks, such as (i) automatically detecting hate speech targets such as racism (Hasanuzzaman, Dias, and Way 2017), antisemitism (Ozalp *et al.* 2020; Zannettou *et al.* 2020), religious intolerance (Ghosh Chowdhury *et al.* 2019), misogyny and sexism (Jha and Mamidi 2017; Guest *et al.* 2021), and cyberbullying (Van Hee *et al.* 2015a), (ii) filtering pages with hateful content and violence (Liu and Forss 2015), (iii) offensive language detection (Zampieri *et al.* 2019; Steimel *et al.* 2019; Vargas *et al.* 2021); (iv) toxic comment detection (Guimarães *et al.* 2020), (v) multimodal hateful content (Cao *et al.* 2022), and

HateBR 2.0 corpus: https://github.com/franciellevargas/HateBR
MOL: https://github.com/franciellevargas/MOL

https://doi.org/10.1017/nlp.2024.18 Published online by Cambridge University Press

(vi) countering hate speech in dialogue systems (Bonaldi *et al.* 2022). Finally, comprehensive surveys on automatic detection of hate speech were also proposed (Schmidt and Wiegand 2017; Fortuna and Nunes 2018; Poletto *et al.* 2021; Vidgen and Derczynski 2021).

Corroborating the particular relevance to fill the gap of data resources for low-resource languages, the possibility of proposing reliable data is essential to build better automatic applications, besides boosting the linguist diversity for hate speech research. Nevertheless, the annotation process of hate speech is intrinsically challenging, considering that what is considered offensive is strongly influenced by pragmatic (cultural) factors, and people may have different opinions on an offense. In addition, there is the presence of implicit content and sarcasm, which hides the real intention of the comment and makes the decision of the annotators confusing.

Indeed, subjective tasks (e.g., hate speech detection, sentiment and emotion analysis, sarcasms and irony detection, etc.) in natural language processing (NLP) present high complexity and a wide variety of technical challenges. Recent proposals have discussed the implications of the annotation process its impact to model hate speech phenomena (Poletto *et al.* 2021), the proposing of multilayer annotation scheme (Zampieri *et al.* 2019), target-aware hate speech annotation (as known as hate speech targets) (Basile *et al.* 2019), and the implicit-explicit distinction in the hate speech annotation process (Caselli *et al.* 2020). In fact, a well-defined annotation schema has a considerable impact on the consistency and quality of the data, as well as the performance of the derived machine-learning classifiers.

In addition, there are also different research gaps in order to address hate speech in terms of methods and data resources. For instance, most existing offensive lexicons are built by means of large crowd-sourced lexical resources, which is limited due to a wide range of irrelevant terms, resulting in high rates of false positives (Davidson, Bhattacharya, and Weber 2019). Moreover, hate speech detection methods based on large language models (LLMs) are trained on real-world data, which are known to embed a wide range of social biases (Nadeem, Bethke, and Reddy 2021; Davani *et al.* 2023). Finally, the relevant issue in the area is related to scarce data resources and methods for low-resource languages.

Toward addressing the lack of data resources for low-resource languages, this paper provides different data resources for low-resource hate speech detection. More specifically, we introduce two data resources: the HateBR 2.0 and MOL – multilingual offensive lexicon, which are both expert data resources for hate speech detection. The HateBR 2.0 consists of a large corpus of 7,000 comments extracted from Brazilian politicians' accounts on Instagram and manually annotated by experts for Brazilian Portuguese hate speech detection. It is an updated version of HateBR corpus (Vargas *et al.* 2022), in which highly similar and one-word comments were replaced. We aim to release the HateBR 2.0 version that is still more consistent and reliable. In the same settings, the MOL comprises explicit and implicit pejorative terms and expressions annotated with context information. The lexicon was also translated by native speakers taking into account cultural adaptations for low-resource languages (Turkish, German, French) and for the English and Spanish languages.

Finally, in order to support the high interhuman agreement score obtained for both data resources (HateBR 2.0 and MOL), besides assessing the reliability of annotated data, we also provide baseline experiments on the HateBR 2.0 corpus and baseline experiments using machine learning (ML) models, which embed terms and expressions extracted from the MOL in English, Spanish, and Portuguese. Results show the reliability of proposed data resources, outperforming baseline dataset results in Portuguese, besides presetting promising results for hate speech detection in different languages.

In what follows, in Section 2, we present related works. Sections 3 and 4 describe the definitions for hate speech and offensive language used in this paper. In Sections 5 and 6, we describe the HateBR 2.0 corpus and the MOL, which are expert and context-aware data resources for low-resource languages, respectively. Section 7 provides data statistics, and Section 8 baseline experiments, evaluation, and results are presented. Finally, in Section 9, the final remarks and future works are presented.

2. Related work

The hate speech detection corpora lie to user-generated public content, mostly microblog posts, and are often retrieved with a keyword-based approach and using words with a negative polarity (Poletto *et al.* 2021). In addition, hate speech detection methods, according to the literature, have focused mainly on the following approaches (Schmidt and Wiegand 2017):

(1) *Simple surface features*: Simple surface-based features such as a bag-of-words (BoW) using unigram and larger n-grams have been proposed by a wide range of authors (Chen *et al.* 2012; Xu *et al.* 2012; Waseem and Hovy 2016; Nobata *et al.* 2016). Besides that, n-gram features are also applied along with several other features (Nobata *et al.* 2016).

(2) *Word generalization*: Considering the data sparsity phenomenon, a wide variety of authors have applied such an approach for word generalization, which consists of carrying out word clustering and then using induced cluster IDs representing sets of words as additional (generalized) features (Schmidt and Wiegand 2017). A set of approaches have been most considered such as clustering (Brown *et al.* 1992), latent Dirichlet allocation (Blei, Ng, and Jordan 2003), word embeddings (Mikolov *et al.* 2013), and paragraph embeddings (Le and Mikolov 2014). Word generalization strategies have been explored by Xiang *et al.* (2012), Nobata *et al.* (2016), and Warner and Hirschberg (2012).

(3) *Sentiment analysis*: This approach assumes that semantic polarity may predict hateful and offensive messages. For example, a method proposed by Van Hee *et al.* (2015b) used a sentiment lexicon to compute the number of positive, negative, and neutral words that occur in a given comment. Examples of this approach were also proposed by Njagi *et al.* (2015), Burnap and Williams (2014), and Burnap *et al.* (2014).

(4) *Lexical resources*: In this approach, a controlled vocabulary of hateful and offensive terms and expressions is used as features to build a model of classification for hate speech and offensive language (Xiang *et al.* 2012; Burnap and Williams 2016; Vargas *et al.* 2021).

(5) *Linguistic features*: Linguistic information is surely relevant for text classification and has been explored for hate speech detection such as part-of-speech (POS), syntactical tree and dependency tuple, semantic relations, etc (Chen *et al.* 2012; Burnap and Williams 2014; Burnap *et al.* 2014; Zhong *et al.* 2016; Nobata *et al.* 2016).

(6) *Knowledge-based features*: Corroborating the fact that the hate speech detection tasks are highly context-dependent; in this approach, cultural and world knowledge-based information are used as features (Dinakar *et al.* 2012; Dadvar *et al.* 2013; Vargas *et al.* 2021).

(7) *Meta-information*: In this approach, meta-information (i.e., information about an utterance) is also used as a feature for the hate speech classification tasks (Xiang *et al.* 2012; Dadvar *et al.* 2013; Waseem and Hovy 2016).

(8) *Multimodal information*: Since different modalities of content also may present hateful content, in this approach, different features are explored to classify hate speech in images, video, and audio content (Boishakhi, Shill, and Alam 2021; Zhu, Lee, and Chong 2022; Thapa *et al.* 2023).

2.1 Hate speech detection for low-resource languages

While most hate speech corpora are proposed for the English language (Davidson *et al.* 2017; Gao and Huang 2017; Jha and Mamidi 2017; Golbeck *et al.* 2017; Fersini, Rosso, and Anzovino 2018; Zampieri *et al.* 2019; AlKhamissi *et al.* 2022), there are a set of proposals toward boosting low-resource hate speech technologies. For example, in French, a corpus of Facebook and Twitter annotated data for Islamophobia, sexism, homophobia, religious intolerance, and disability detection was also proposed (Chung *et al.* 2019; Ousidhoum *et al.* 2019). In Germany,

https://doi.org/10.1017/nlp.2024.18 Published online by Cambridge University Press

a new anti-foreigner prejudice corpus was proposed (Bretschneider and Peters 2017). This corpus is composed of 5,836 Facebook posts and hierarchically annotated with slightly and explicitly/substantially offensive language according to six targets: foreigners, government, press, community, other, and unknown. In Greek, an annotated corpus of Twitter and Gazeta posts for offensive content detection is also available (Pavlopoulos, Malakasiotis, and Androutsopoulos 2017; Pitenis, Zampieri, and Ranasinghe 2020). In Slovene and Croatian, a large-scale corpus composed of 17,000,000 posts, with 2 percent of hate speech on a leading media company website was built (Ljubešić *et al*. 2018). In the Arabic language, there is a corpus of 6,136 Twitter posts, which are annotated according to religion intolerance subcategories (Albadi, Kurdi, and Mishra 2018). In the Indonesian language, a hate speech annotated corpus from Twitter data was also proposed in Alfina *et al.* (2017).

2.2 Hate speech detection for the Portuguese language

For the Portuguese languages, an annotated corpus of 5,668 European and Brazilian Portuguese tweets was proposed in Fortuna *et al.* (2019). The corpus comprises two annotation levels: binary classification (hate speech versus non-hate speech) and hierarchical labels of nine direct social groups targeted by discrimination. In addition, automated methods using a hierarchy of hate were proposed. They used pretrained Glove word embedding with 300 dimensions for feature extraction and an Long Short-Term Memory (LSTM) architecture proposed in Badjatiya *et al.* (2017). The authors obtained an F1-score of 78 percent using cross-validation. Furthermore, a new specialized lexicon was also proposed specifically for European Portuguese, which may be useful to detect a broader spectrum of content referring to minorities (Fortuna *et al.* 2021).

More specifically for Brazilian Portuguese, a corpus of 1,250 comments collected from Brazilian online newspaper G1,[a] annotated with a binary class (offensive and non-offensive), and six hate speech targets (racism, sexism, homophobia, xenophobia, religious intolerance, or cursing) was provided by de Pelle and Moreira (2017). The authors provide a baseline using a support vector machine (SVM) with linear kernel and multinomial Naive Bayes (NB). The best model obtained an F1-score of 80 percent. Specifically for toxicity classification in Twitter posts for the Brazilian Portuguese language, a corpus composed of 21,000 tweets manually annotated according to seven hate speech targets—nontoxic, LGBTQ + phobia, obscene, insult, racism, misogyny, and xenophobia—was proposed in Leite *et al.* (2020), in which the BERT fine-tuning baseline was presented reaching an F1-score of 76 percent. Furthermore, the Offensive Language Identification Dataset for Brazilian Portuguese (OLID-BR) (Trajano, Bordini, and Vieira 2023) was also provided for Brazilian Portuguese, which was inspired by the original OLID in English. This corpus consists of 7,943 comments extracted from YouTube and Twitter and annotated according to different categories: health, ideology, insult, LGBTQphobia, other lifestyle, physical aspects, profanity/obscene, racism, religious intolerance, sexism, and xenophobia. The authors provide a BERT fine-tuning baseline reaching an F1-score of 77 percent. Finally, the HateBR corpus (Vargas *et al.* 2022) comprises 7,000 Instagram comments manually annotated by experts according to a binary class (offensive and non-offensive) and hate speech targets.In addiation, baseline experiments on HateBR corpus outperfromed the current state-of-the-art for Portuguese. In Table 1, we summarize the literature data resources for the Portuguese language.

3. Offensive language definition

Offensive posts include insults, threats, and messages containing any form of untargeted profanity (Zampieri *et al*. 2019). In the same settings, offensive language consists of any profanity, strongly impolite, rude, or vulgar language expressed with fighting or hateful words in order to insult a targeted individual or group (Fortuna and Nunes 2018). Accordingly, in this paper, we assume

[a] g1.globo.com

https://doi.org/10.1017/nlp.2024.18 Published online by Cambridge University Press

Table 1. Portuguese data resources

Authors	Data resources	Total	Plataform	Experiments	F-score
Trajano *et al.* (2023)	OLID-BR - corpus	7,943	Twitter, YouTube	BERT	0.77
Vargas *et al.* (2022)	HateBR - corpus	7,000	Instagram	NB, SVM	0.88
Fortuna *et al.* (2021)	MIN_PT - lexicon	155	Selected terms	None	None
Leite *et al.* (2020)	ToLD-Br - corpus	21,000	Twitter	BERT	0.76
de Pelle and Moreira (2017)	OFFCOMBR - corpus	1,250	Website comments	NB	0.81
Fortuna *et al.* (2019)	No-Name - corpus	5,668	Twitter	LSTM	0.78

Note: BERT, Bidirectional Encoder Representations from Transformers; NB, Naive Bayes.

that *offensive language consists of a type of language containing terms or expressions used with any pejorative connotation against people, institutions, or groups regardless of their social identity, which may be expressed explicitly or implicitly*.

Table 2 shows examples of offensive and non-offensive comments extracted from the HateBR 2.0 corpus. Note that **bold** indicates terms or expressions with explicit pejorative connotation and underline indicates "clues" of terms or expressions with an implicit pejorative connotation.

Table 2. Offensive and non-offensive comments extracted from the HateBR 2.0 corpus

N.	Type	Instagram comments	Translation
1	Offensive comments	Essa **besta humana** é o **câncer** do País, tem q voltar p jaula, urgentemente! E viva o Presidente Bolsonaro.	This **human beast** is the **cancer** of the country, it has to go back to the cage, urgently! And long live to President Bolsonaro.
2	Offensive comments	Pois é, deveria devolver o dinheiro aos cofres públicos do Brasil. **Canalha**.	It is means, they should refund money to the public Brazilian coffers. **Jerk**.
3	Non-offensive comments	Quem falou isso pra vc deputada? O sergio moro ta aprovado pela maioria dos brasileiros.	Who said that to you deputy? Sergio Moro is approved by the majority of Brazilians.
4	Non-offensive comments	A minoria rica é bem organizada e se reúnem secretamente. Lucram muito com essa política exclusiva deles.	The wealthy minority is well organized and meets in secret. They profit a lot from this exclusive policy of theirs.

As shown in Table 2, offensive comments comprise terms or expressions with pejorative connotations, which were expressed explicitly and implicitly. For example, in comments 1 and 2, while the term "cancer" may be used in non-pejorative contexts (e.g., he has cancer), in this comment context, it was used with a pejorative connotation. Differently, the expression "human beast" and the term "jerk" both present pejorative connotations and are mostly found in the pejorative context of use. Furthermore, in offensive comments, there are offensive terms or expressions expressed implicitly. For example, the expressions "go back to the cage" and "refund money" are clue elements that indicate terms used with pejorative connotations such as "criminal" and "thief," respectively. Finally, non-offensive comments do not present any terms or expressions used with pejorative connotations, as observed in comments 3 and 4.

4. Hate speech definition

Here, we defined hate speech *as a type of offensive language that attacks or diminishes inciting violence and hate against groups, based on specific characteristics (e.g., physical appearance, religion,*

https://doi.org/10.1017/nlp.2024.18 Published online by Cambridge University Press

descent, national or ethnic origin, sexual orientation, gender identity, etc.), and it may occur with different linguistic styles, even in subtle forms or when humor is used. Hence, hate speech is a type of offensive language used against hate targets (Fortuna and Nunes 2018). Accordingly, in order to precisely elucidate them, nine different hate speech targets are defined in this paper, which we described in detail as follows:

(1) *Antisemitism*: The definition of antisemitism adopted by the International Holocaust Remembrance Alliance (IHRA)[b] in 2016 states that “Antisemitism is a certain perception of Jews, which may be expressed as hatred towards Jews. Rhetorical and physical manifestations of antisemitism might include the targeting of the state of Israel, conceived as a Jewish collectivity,” as in the following example: *Que escroto caquético! É a velha hipocrisia judaica no mundo dos pilantras monetários. Judeu dos infernos!*
Translation: *What a cachectic asshole! It's the old Jewish hypocrisy in the world of monetary hustlers. Jew from hell!*..

(2) *Apologist for dictatorship*: According to the Brazilian Penal Code,[c] apologist for dictatorship consists of comments that incite the subversion of the political or social order, the animosity among the Armed Forces, or among these and the social classes or civil institutions, as in the following example: *Intervenção militar já !!! Acaba Supremo Tribunal Federal, não serve pra nada mesmo. . .*
Translation: *Military intervention now !!! Close the Supreme Court,[d] it is of no use at all. . .*.

(3) *Fatphobia*: Fatphobia is defined as negative attitudes based on stereotypes against people socially considered fat (Robinson, Bacon, and O'Reilly 1993), as in the following example: *Velha barriguda e bem folgada, heim? Porca rosa, Peppa!*.
Translation: *Old potbellied and very lazy, huh? Pink Nut, Peppa*[e]!.

(4) *Homophobia*: Homophobia[f] is considered an irrational fear or aversion to homosexuality, or, in other words, to lesbian, gay, and bisexual people based on prejudice, as in the following example: *Quem falou isso deve ser um global que não sai do armário :) :(e tem esse desejo :(:(nessa hora que tinha que intervir aqui e botar um merda desse no pau. . . .Dá muito o cú.*
Translation: *Whoever said that must be an artist who does not come out of the closet :) :(and has that desire :(:(at this point they should intervene here and apply the law against them. . . . They give the ass a lot.*

(5) *Partyism*: Partyism is a form of extreme hostility and prejudice against people or group based on their political orientation, which influences non-political behaviors and judgment (Westwood *et al.* 2018). In our corpus, the most relevant occurrence of hate speech consists of partyism, as in the following example: *Os petralhas colocaram sua corja em todos os lugares, não salva ninguém, que tristeza. . . Esquerda parasita* lixo.
Translation: *The petralhas*[g] *have puted their crowds everywhere, no one can be saved, how sad. They are parasite and wreckage* .

[b] The IHRA unites governments and experts to strengthen, advance, and promote Holocaust education, research, and remembrance.

[c] Brazilian Penal Code, Decree-Law No. 2,848/1940, is formed by a set of systematic rules with a punitive character. Its purpose is the application of sanctions in conjunction with discouraging the practice of crimes that threaten the social fabric.

[d] The Supreme Federal Court of Brazil is the highest court in the country on constitutional matters. There can be no appeal against its decisions.

[e] Peppa Pig is a British preschool animated television series directed and produced by Astley Baker Davies in association with Entertainment One. The show revolves around Peppa, an anthropomorphic female pig, her family, and friends.

[f] According to the European Institute for Gender Equality. https://tinyurl.com/4yca8vpm.

[g] Petralha is a deeply culture-rooted pejorative name used by conservative politicians to originally define liberal politicians associated with a specific party in Brazil.

https://doi.org/10.1017/nlp.2024.18 Published online by Cambridge University Press

(6) *Racism/racial segregation*: Racism consists of an ideology of racial domination (Wilson 1999). It presumes biological or cultural superiority of one or more racial groups, used to justify or prescribe the inferior treatment or social position(s) of other racial groups (Clair and Denis 2015). In our corpus, we found a wide range of offenses related to racial discrimination, such as "monkey" and "cheetah," as an example of racist comment: *E uma chita ela né! Opssss, uma chata.*
Translation: So, *she is a cheetah right! Opssss, a boring girl.*

(7) *Religious intolerance*: Theoretical constructs loom large in the literature on religiosity and intolerance, namely, religious fundamentalism, which is consistently associated with high levels of intolerance and prejudice against religion groups (Altemeyer 1996). For instance, observe the following comments: *Pastor dos Infernos* and *O chamado crente do demônio, né?*
Translation: *Pastor of the Church from Hell and The so-called Christian of the devil, right?*

(8) *Sexism*: Sexist behavior is mostly related to patriarchy that consists of a system of social structures that allow men to exploit women (Walby 1990). Therefore, sexism consists of hostility against self-identified people as female gender, treated them as objects of sexual satisfaction of men, reproducers labor force, and new breeders (Delphy 2000). The following example was extracted from the corpus: *Cala esse bueiro de falar merda sua vagabunda safada.*
Translation: *Shut that speaking manhole up you nasty slut*.

(9) *Xenophobia*: Xenophobia is a form of racial prejudice (Silva *et al.* 2016; Davidson *et al.* 2019), which is manifested through discriminating actions and hate against people based on their origin, as in the following example: *Ele está certo. Vai ter um monte de argentino faminto invadindo o Brazil.*
Translation: *He is right. There will be a lot of hungry Argentine people invading Brazil.*

5. HateBR 2.0 corpus

Brazil occupies the third position in the worldwide ranking of Instagram's audience with 110 million active Brazilian users, ahead of Indonesia with an audience of 93 million users.[h] On the Instagram platform, each person has an account with shared photos, and it is possible for others to like, comment, save, and share this information. Taking advantage of the fact that Instagram is an online platform with high engagement in Brazil, HateBR (Vargas *et al.* 2022) and HateBR 2.0 data collection relied on this platform.

The proposed approach for the construction of the HateBR corpus is divided into two macro-steps: (A) **data collection** and (B) **data annotation**. Data collection relied on four tasks: (first) domain definition, (second) criteria for selecting platform accounts, (third) data extraction, and (fourth) data cleaning. Data annotation relied on three tasks: (first) selection of annotators, (second) annotation schema, and (third) annotation evaluation. In this paper, we introduce an updated version of HateBR corpus (Vargas *et al.* 2022) entitled HateBR 2.0, and we followed the same methodology for data collection and data annotation, except the fact that in this version the comments were manually collected from Instagram in order to replace one-word and highly similar comments.

5.1 Data collection

We provide an updated version of HateBR corpus (Vargas *et al.* 2022) entitled HateBR 2.0. In this version, we used the original data from the HateBR and replaced a set of comments taking

[h] https://www.statista.com/

https://doi.org/10.1017/nlp.2024.18 Published online by Cambridge University Press

into account two criteria: (A) highly similar comments and (B) one-word comments, as shown in Table 3. We aim to release a corpus version that is still more consistent and reliable. For example, we aim to replace highly similar and one-word comments to improve the ability of classifiers to consistently recognize and generalize each class. In total, 17.4 percent of the corpus was replaced totaling 1,216 comments through which 911 were non-offensive and 305 were offensive, as shown in Section 7.

Table 3. Criteria for updating the HateBR corpus

Criteria	Description	Example 1	Example 2	Example 3
Criteria A	One-word comments	Criminal	Bitch	Pig
Criteria B	Highly similar comments	The best president & The best president hahahaha	Communist & @veron-ica_michelle_bachelett Communist	Love you so much & Love you so much :) =) :) =)

Observe that, as shown in Table 3, comments with only one word as “criminal” or “bitch” or “pig” were replaced by newly collected data. In addition, we also replaced highly similar comments. For instance, example 1 shows both comments “the best president” and “the best president hahahah.” In this case, the difference between both is the laugh (hahahah). In the same settings, in examples 2 and 3, the difference between comments is the account (@veronica_michelle_bachelett) and the emotions (:) =) :) =)).

Toward collecting new Instagram data, we followed the same settings used for data collection previously in HateBR corpus (Vargas *et al.* 2022). For example, in the first step—domain definition—we have chosen the political domain. In the second step—criteria for selecting accounts—the following criteria were defined to improve the representatives of the corpus in terms of localization, gender, and political orientation. Hence, 6 (six) different Brazilian politicians’ accounts were selected, of which 3 (three) accounts are from left-wing politicians and 3 (three) accounts are from right-wing politicians. In addition, we selected 4 (four) accounts from women Brazilian politicians and 2 (two) men Brazilian politicians. In the third step—data extraction—differently from the data collection for the HateBR (initial version), for the HateBR 2.0, we manually collected comments from the selected Brazilian politicians’ accounts. In the fourth step—data cleaning, we proposed an approach for data cleaning, which consists of removing noise, such as links, characters without semantic value, and also comments that presented only emoticons, laughs (kkk, hahah, hshshs), mentions (e.g., @namesomeone), or hashtags (#lulalive) without any textual content. Comments that comprise text along with hashtags or emotions were kept.

5.2 Annotation process

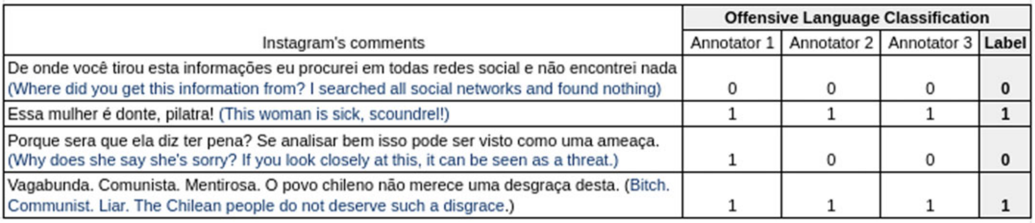
5.2.1 Selection of annotators

The first step of our annotation process consists of the selection of annotators. We argue that for hate speech tasks, only specialists with relevant levels of education should be selected due to the high complexity of tasks that tackle highly politicized domains. In addition, it is necessary to provide a strategy toward the mitigation of annotator bias such as a diverse profile of annotators. Hence, we selected annotators from different Brazilian regions (North and Southeast), and they had at least a PhD study. Furthermore, they were white and black women and were aligned with different political parties (liberal or conservative).

5.2.2 Annotation schema

Since the offensiveness definition is a challenge and most existing definitions struggle with ambiguity, we propose an accurate definition for offensive language and hate speech through a new annotation schema, which is divided into two layers: (1) offensive language classification and (2) hate speech classification. The definitions for offensive language and hate speech used in this paper are described in detail in Sections 3 and 4, which were used to propose our annotation schema described as follows.

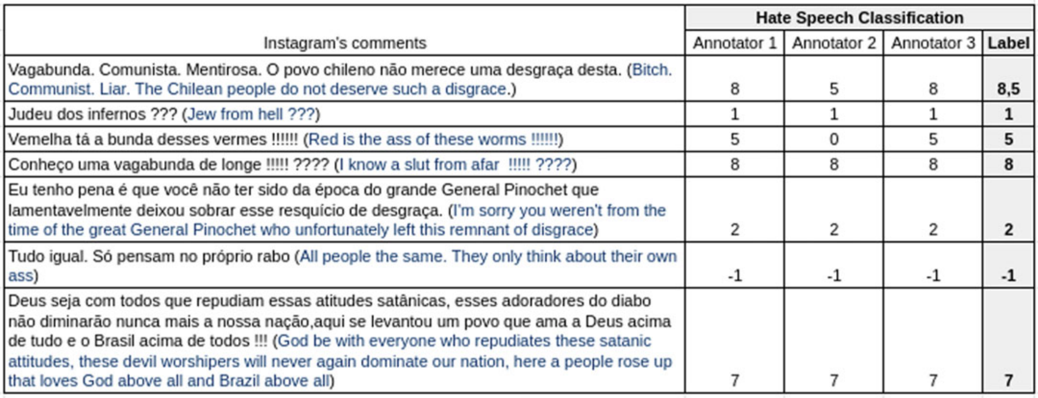
(1) **Offensive language classification**: We assume that comments that present at least one term or expression used with any pejorative connotation against people, institutions, or groups regardless of their social identity, which may be expressed explicitly or implicitly, should be classified as offensive. Otherwise, comments that have no terms or expressions with any pejorative connotation should be classified as non-offensive. Examples of offensive and non-offensive annotations are shown in Figure 1.



Instagram's comments	Offensive Language Classification			
	Annotator 1	Annotator 2	Annotator 3	**Label**
De onde você tirou esta informações eu procurei em todas redes social e não encontrei nada (Where did you get this information from? I searched all social networks and found nothing)	0	0	0	**0**
Essa mulher é donte, pilatra! (This woman is sick, scoundrel!)	1	1	1	**1**
Porque sera que ela diz ter pena? Se analisar bem isso pode ser visto como uma ameaça. (Why does she say she's sorry? If you look closely at this, it can be seen as a threat.)	1	0	0	**0**
Vagabunda. Comunista. Mentirosa. O povo chileno não merece uma desgraça desta. (Bitch. Communist. Liar. The Chilean people do not deserve such a disgrace.)	1	1	1	**1**

Figure 1. Offensive language classification: Each Instagram comment was classified according to a binary class: offensive or non-offensive. We manually balanced the classes and obtained 3,500 offensive comments labeled as (1) and 3,500 non-offensive comments labeled as (0).

(2) **Hate speech classification**: We assume that offensive comments targeted against groups based on their social identity (e.g., gender, religion, etc.) should be classified as hate speech into nine hate speech targets (xenophobia, racism, homophobia, sexism, religious intolerance, partyism, apologist for dictatorship, antisemitism, and fatphobia). On the other hand, offensive comments not targeted against groups should be classified as non-hate speech. Examples are shown in Figure 2.



Instagram's comments	Hate Speech Classification			
	Annotator 1	Annotator 2	Annotator 3	**Label**
Vagabunda. Comunista. Mentirosa. O povo chileno não merece uma desgraça desta. (Bitch. Communist. Liar. The Chilean people do not deserve such a disgrace.)	8	5	8	**8,5**
Judeu dos infernos ??? (Jew from hell ???)	1	1	1	**1**
Vemelha tá a bunda desses vermes !!!!!! (Red is the ass of these worms !!!!!!)	5	0	5	**5**
Conheço uma vagabunda de longe !!!!! ???? (I know a slut from afar !!!!! ????)	8	8	8	**8**
Eu tenho pena é que você não ter sido da época do grande General Pinochet que lamentavelmente deixou sobrar esse resquício de desgraça. (I'm sorry you weren't from the time of the great General Pinochet who unfortunately left this remnant of disgrace)	2	2	2	**2**
Tudo igual. Só pensam no próprio rabo (All people the same. They only think about their own ass)	-1	-1	-1	**-1**
Deus seja com todos que repudiam essas atitudes satânicas, esses adoradores do diabo não diminarão nunca mais a nossa nação,aqui se levantou um povo que ama a Deus acima de tudo e o Brasil acima de todos !!! (God be with everyone who repudiates these satanic attitudes, these devil worshipers will never again dominate our nation, here a people rose up that loves God above all and Brazil above all)	7	7	7	**7**

Figure 2. Hate speech classification: We identified nine hate speech targets, and we labeled them as follows: antisemitism was annotated as (1), apologist for dictatorship as (2), fatphobia as (3), homophobia as (4), partyism as (5), racism as (6), religious intolerance as (7), sexism as (8), and xenophobia as (9). It should be pointed out that a couple of comments belong to more than a target. For example, the comment *comunista, vagabunda e safada* ("shameless, communist and slut") was classified as partyism and sexism; hence it was labeled as (5,8). Offensive comments without hate speech were annotated as (−1).

Finally, we selected the final label for the HateBR 2.0 corpus considering the majority of votes for offensive language classification. For hate speech classification, we also considered the majority of votes and a linguist judged tie cases.

5.2.3 Annotation evaluation

The third step of our annotation process consists of applying metrics to evaluate the annotated data. Accordingly, we used two different evaluation metrics: Cohen's kappa (Sim and Wright 2005; McHugh 2012) and Fleiss' kappa (Fleiss 1971), as shown in Tables 4 and 5. Specifically, for the HateBR 2.0 corpus, we recalculated both Kappa and Fleiss for the updated data taking into consideration the replaced comments.

Cohen's kappa results are shown in Table 4. Notice that a high inter-annotator agreement of 75 percent was achieved for the binary class (offensive and non-offensive comments). In the same settings, as shown in Table 5, a high Fleiss kappa score was obtained of 74 percent. Therefore, the annotation evaluation results corroborate the quality of the HateBR 2.0 annotated corpus proposed in this paper.

Table 4. Cohen's kappa

Peer agreement	AB	BC	CA	Avg
Binary class (offensive × non-offensive)	0.75	0.69	0.80	**0.75**

Table 5. Fleiss' kappa

Fleiss' kappa	ABC
Binary class (offensive × non-offensive)	**0.74**

6. MOL – multilingual offensive lexicon

We argue that hate speech should be addressed as a cultural-aware research problem since it is a complex issue that deals with commonsense knowledge and normative social stereotypes. Consequently, hate speech technologies should be able to distinguish terms and expressions used with pejorative connotations according to the context of use. For instance, while the terms "cancer," "garbage," and "worms" may be used with pejorative connotations, they could be also used without any pejorative connotation (e.g., "he was cured of cancer" "the garden is full of parasites and worms" "disposal of garbage on streets").

According to linguistic studies, the pejorative connotation is used to express emotions, especially hate, anger, and frustration. In addition, it is heavily influenced by pragmatic (cultural) factors (Allan 2007; Rae 2012; Anderson and Lepore 2013; Bousfield 2013). In the same settings, swear words express the speaker's emotional state and provide a link to impoliteness and rudeness research. They are considered a type of opinion-based information that is highly confrontational, rude, or aggressive (Jay and Janschewitz 2008).

Furthermore, recent studies have shown that 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), besides the fact that pretrained language models are trained on large real-world data, consequently, they are known to embody social biases (Nadeem *et al.* 2021). In this paper, toward addressing some of these limitations, we introduce a new data resource that follows a different proposal for context recognition. The proposed resource entitled MOL – multilingual offensive

https://doi.org/10.1017/nlp.2024.18 Published online by Cambridge University Press

lexicon comprises 1,000 explicit and implicit terms and expressions used with pejorative connotations, which was manually extracted by a linguist from the HateBR corpus and annotated by 3 different experts with contextual information, reaching a high human agreement score (73 percent Kappa).

The methodology used for building the MOL comprises five steps: (i) *terms extraction*, (ii) *hate speech targets*, (iii) *context annotation*, (iv) *annotation evaluation*, and (v) *translation and cultural adaptation*. We describe in more detail each of these steps as follows.

6.1 Terms extraction

In the first step, the explicit and implicit terms and expressions used with any pejorative connotation were manually identified by a linguist from the HateBR corpus (Vargas *et al.* 2022). In other words, the linguist manually identified explicit and implicit terms and expressions that presented any pejorative context of use. For instance, the terms "trash," "pig," and "bitch" are examples of explicit terms that can be used with a pejorative connotation. On the other hand, implicit terms and expressions were identified using *clues*. For example, "go back to the cage" is *clue* to identify the implicit offensive term "criminal," and "the easy woman" is a *clue* to identify the implicit offensive term "bitch." In total, 1,000 explicit and implicit pejorative terms and expressions were identified, as shown in Table 6 (see Section 7.2). Lastly, a set of terms and expressions were classified by the linguist according to a category called *deeply culture-rooted*.[i] Indeed, deeply culture-rooted terms and expressions do not make sense in other cultures; hence there are no "ideal translations." For example, the term "bolsonazi" is used in Brazil as a neologism by agglutination of the words "Bolsonaro" (former Brazilian president) with the word "Nazism." Approximately 10 percent of terms were classified as deeply culture-rooted.

6.2 Hate speech targets

In the second step, the linguist accurately identified terms and expressions used to express hate against groups based on their social identity. For example, the term "bitch" and the expression "Jews from hell" potentially are able to indicate hate speech comments against gender and Jews. For terms and expressions without any potential to indicate hate speech, the label "no-hate speech" was provided. It should be pointed out that a set of terms and expressions received more than one label. For example, the term "feminazi" received both partisan and sexist labels. There were some ambiguous cases, in which the linguist made a decision on the most suitable label. In total, 150 terms were labeled as hate speech targets, as shown in Table 7 (see Section 7.2).

6.3 Context annotation

In the third step, three different annotators classified each 1 of 1,000 identified pejorative terms and expressions according to a binary class: *context-dependent* or *context-independent*. The annotators first checked whether the term or expression was mostly found in the pejorative context of use. If yes, it was annotated as context-independent. Second, the annotators checked if the term or expression may be found in both the pejorative context of use and non-pejorative context of use. If yes, it was annotated as context-dependent. For example, the terms "wretched" and "hypocritical" are examples of terms annotated as context-independent due to the fact that both terms are mostly found in pejorative contexts of use. On the other hand, the terms "worm" and "useless" are examples of terms annotated as context-dependent given that these terms may be found in both

[i] We argue that deeply culture-rooted terms and expressions consist of a vocabulary that represents values, assumptions, and symbols of a particular culture.

https://doi.org/10.1017/nlp.2024.18 Published online by Cambridge University Press

non-pejorative and pejorative contexts of use. Lastly, in order to support the contextual annotation process, the annotators also checked the dictionary meaning for each term and expression and evaluated whether it presented or not a pejorative connotation. Then, the annotators made the decision on the best label for each term or expression also considering their world vision and expertise.

6.4 Annotation evaluation

In this fourth step, we evaluated the consistency and reliability of the contextual information annotation process using Cohen's kappa (Sim and Wright 2005; McHugh 2012). The obtained results are shown in Table 8, in which A, B, and C letters stand for the annotators, and agreement is measured by pair combination. Note that a high inter-annotator agreement of 73 percent was obtained. It is worth emphasizing that values around 70 percent are considered a substantial agreement (Landis and Koch 1977).

Table 8. Kappa score obtained for the contextual-aware offensive lexicon annotation

Metrics	AB	BC	CA	Avg
Kappa	0.72	0.60	0.87	**0.73**

6.5 Translation and cultural adaptation

The MOL terms and expressions were originally written in Brazilian Portuguese and manually translated by native speakers for five different languages: English, Spanish, French, German, and Turkish. The native speakers proposed literal translations, as well as adaptations in terms of culture. For example, the native speakers proposed translations and cultural adaptations for the 1,000 explicit and implicit terms and expressions corroborating regional and cultural aspects of the target language. Furthermore, a set of terms and expressions were categorized as "deeply culture-rooted," in which there were no suitable translations, which comprised 10 percent of the data. Finally, the contextual labels were reevaluated by native speakers, and there were no significant modifications.

7. Data statistics

7.1 HateBR 2.0 corpus

As a result, we present the HateBR 2.0 corpus statistics in Tables 9, 10, and 11. Observe that the HateBR 2.0 corpus maintained the same number of comments: 3,500 offensive and 3,500 non-offensive. Moreover, in total 17 percent of the corpus was replaced through which 305 were offensive and 911 were non-offensive. Lastly, as shown in Table 11, most offensive comments were published in posts of right-wing politicians' accounts totaling 2,099 offensive comments in contrast to 1,401 offensive comments published in posts of left-wing politicians' accounts.

Table 9. Binary class

Labels	Total
Non-offensive	3,500
Offensive	3,500
Total	7,000

https://doi.org/10.1017/nlp.2024.18 Published online by Cambridge University Press

Table 10. Replaced comments

Labels	Total
Offensive	305
Non-offensive	911
Total	1.216

Table 11. Political orientation

Political party	Non-offensive	Offensive
Left-wing politicians' accounts	1.809	1.401
Right-wing politicians' accounts	1.691	2.099
Total	3.500	3.500

7.2 MOL – multilingual offensive lexicon

We also present the MOL statistics in Tables 6, 7, 12, and 13. Observe that the specialized lexicon comprises 951 offensive terms or expressions used to explicitly express offensiveness and 49 "clues," which are used to implicitly express offensiveness, totaling 1,000 terms and expressions. In addition, 612 terms are annotated as context-independent and 388 as context-dependent. Lastly, the MOL comprises 724 terms (unigram) and 276 expressions (n-grams), and 150 terms were also annotated according to their hate speech target.

Table 6. Explicit and implicit information

Type	Total
Explicit information	951
Implicit information	49
Total	1,000

Table 7. Terms annotated with hate speech targets

Type	Total
No-hate speech	850
Partyism	69
Sexism	35
Homophobia	16
Fatphobia	9
Religious intolerance	9
Apologist for dictatorship	5
Racism	4
Antisemitism	3
Total	1,000

https://doi.org/10.1017/nlp.2024.18 Published online by Cambridge University Press

Table 12. Contextual information labels

Labels	Total
Context-independent	612
Context-dependent	388
Total	1,000

Table 13. Terms and expressions

Type	Total
Total of terms	724
Total of expressions	276
Total	1,000

8. Baseline experiments

In order to support the high interhuman agreement score obtained for both data resources (HateBR 2.0 and MOL), besides assessing the reliability of data, we implemented baseline experiments on the HateBR 2.0 corpus. We further implemented two ML-based models, which embed the terms and expressions from the MOL using corpora in English, Spanish, and Portuguese. We describe our experiments in Sections 8.1 and 8.2

8.1 Experiments on the HateBR 2.0 corpus

In our experiments, we used Python 3.6, scikit-learn,[j] pandas,[k] spaCy,[l] and Keras[m] and sliced our data in 90 percent train and 10 percent validation. We used a wide range of feature representation models and learning methods, described as follows:

The features set

We implemented text feature representation models, such as BoW (Manning and Schutze 1999) using TF-IDF (Term Frequency–Inverse Dense Frequency), Facebook word embeddings (Joulin *et al.* 2017), and mBERT (Bidirectional Encoder Representations from Transformers (Devlin *et al.* 2019). For fastText, we proposed three models (unigram, bigram, and trigram). We implemented mBERT with a maximum feature size of 500, batch size of 64, and at a 1cycle learning rate of 0.00002,1, using Keras and Ktrain. As preprocessing, we only carried out the lemmatization of corpora for BoW models. We used the scikit-learn library and *CountVectorizer* and *TfidfVectorizer*.

Learning methods

We implemented different ML methods: NB (Eyheramendy, Lewis, and Madigan 2003), SVM with linear kernel (Scholkopf and Smola 2001), fastText (Joulin *et al.* 2017), and mBERT (Devlin *et al.* 2019).

[j] https://scikit-learn.org/stable/
[k] https://pandas.pydata.org/
[l] https://spacy.io/
[m] https://keras.io/

https://doi.org/10.1017/nlp.2024.18 Published online by Cambridge University Press

8.2 Experiments using the MOL – multilingual offensive lexicon

We also performed experiments using the MOL. We implemented models that embed the lexicon and evaluated it on different corpora including the HateBR 2.0. Specifically, the following corpora were used in our experiments: the HateBR 2.0, which is a corpus of Instagram comments in Brazilian Portuguese; the OLID (Zampieri *et al.*, 2019), which is a corpus of tweets in English; and the HatEval (Basile *et al.*, 2019), which is a corpus of tweets in Spanish. As learning methods, we used SVM with linear kernel (Scholkopf and Smola 2001) and used two different feature text representations (models), called MOL and B + M (Vargas *et al.* 2021), which we describe in more detail as follows:

MOL

(Vargas *et al.* 2021): This model consists of a BoW generated using only the terms or expressions extracted from the MOL, which were used as features. For instance, for each comment, the occurrence of the MOL terms was counted. Additionally, context labels (context-independent and context-dependent) were considered to compute different weights to features. For example, terms annotated with context-independent labels were assigned a strong weight. Differently, terms annotated with context-dependent labels were assigned weak weight.

B+M

(Vargas *et al.* 2021): This model consists of a BoW, which embed the labels from the MOL. In this model, a BoW was generated using the vocabulary from all comments in the corpus. Then, we performed the match with terms in the MOL, and then we assigned a weight for terms or expressions annotated as context-dependent (weak weight) and context-independent (strong weight).

8.3 Evaluation and results

We evaluated the ML models using precision, recall, and F1-score measures. We presented the results for each class involved and the arithmetic average. The results are shown in Tables 14 and 15.

HateBR 2.0 corpus

The results of experiments on the HateBR 2.0 corpus are shown in Table 14. Observe that we implemented six different models (TF-IDF (SVM), TF-IDF (NB), fastText-unigram, fastText-bigram, fastText-trigram, and mBERT) and three different learning methods (BoW, word embeddings, and transformers). The best performance was obtained using mBert (85 percent of F1-score). Surprisingly, the models TF-IDF (SVM) obtained relevant performance (84 percent of F1-score) similar to results obtained in fastText-unigram (84 percent of F1-score). The worst performance was obtained using TF-IDF (NB) (77 percent of F1-score). Finally, despite the fact that corpus comparison is a generally challenging task in NLP, we propose a comparison analysis between Portuguese corpora. We compared the HateBR 2.0 corpus and our baselines with current corpora and literature baselines for European and Brazilian Portuguese. First, the interhuman agreement obtained in the HateBR 2.0 corpus overcame the other Portuguese data resources (see Table 1). In addition, our corpus presents a balanced class (3.500 offensive and 3.500 non-offensive), in contrast to the other Portuguese corpora, in which the classes are unbalanced. Analyzing the current baseline on Portuguese datasets, the results obtained on HateBR 2.0 corroborate the initial premise that an expert annotated corpus and accurate definitions may provide better performance for automatic ML classifiers. Our baseline experiments on the HateBR 2.0

https://doi.org/10.1017/nlp.2024.18 Published online by Cambridge University Press

Table 14. Baseline experiments on the HateBR 2.0 corpus

Type	Models	Class	Precision	Recall	F1-score
Bag-of-words	TF-IDF (SVM)	0	0.84	0.84	0.85
		1	0.83	0.83	0.83
		avg	0.84	0.84	0.84
	TF-IDF (NB)	0	0.75	0.85	0.80
		1	0.81	0.70	0.75
		avg	0.78	0.78	0.77
Word embeddings	fastText-unigram	0	0.82	0.82	0.84
		1	0.87	0.83	0.85
		avg	0.84	0.84	0.84
	fastText-bigram	0	0.76	0.90	0.82
		1	0.89	0.74	0.80
		avg	0.82	0.81	0.81
	fastText-trigram	0	0.75	0.90	0.81
		1	0.88	0.71	0.78
		avg	0.81	0.80	0.80
Transformers	mBert	0	0.87	0.85	0.86
		1	0.84	0.86	0.85
		avg	0.85	0.85	**0.85**

Note: TF-IDF, Term Frequency–Inverse Dense Frequency; SVM, support vector machine; NB, Naive Bayes; mBert, Bidirectional Encoder Representations from Transformers.

overcame the current baseline datasets for the Portuguese language, as shown in Tables 1 and 14, reaching 85 percent of F1-score by fine-tuned mBert.

Multilingual Offensive Lexicon

The results of experiments using the MOL are shown in Table 15 and in Figures 3, 4, and 5. Notice that the experiments relied on corpora in three different languages (English, Spanish, and Portuguese) using models that embed the specialized lexicon. In general, the obtained results are highly satisfactory for both models that embed the lexicon entitled MOL and B + M on the corpus in English, Spanish, and Portuguese. For instance, the MOL model presented better performance for Spanish (82 percent) and Portuguese (88 percent), and the B + M model presented better performance for English (73 percent). The best performance was obtained using MOL on the corpus in Portuguese, and the worst performance was obtained using the same model on the corpus in English (72 percent). It should be pointed out that we are not proposing a corpora comparison here, we aim to present baseline experiments on hate speech data in different languages, in which the implemented models embed the MOL lexicon. Therefore, the experiments demonstrate that models that embed the proposed MOL lexicon are promising for hate speech detection in different languages.

https://doi.org/10.1017/nlp.2024.18 Published online by Cambridge University Press

Table 15. Baseline experiments using the MOL – multilingual offensive lexicon

Models	Dataset	Class	Precision	Recall	F1-score
MOL	English	0	0.66	0.86	0.75
		1	0.83	0.60	0.70
		avg	0.74	0.73	0.72
	Spanish	0	0.76	0.85	0.80
		1	0.87	0.79	0.83
		avg	0.82	0.82	**0.82**
	Portuguese	0	0.84	0.94	0.89
		1	0.92	0.81	0.86
		avg	0.88	0.87	**0.88**
B+M	English	0	0.67	0.84	0.75
		1	0.82	0.63	0.71
		avg	0.74	0.74	**0.73**
	Spanish	0	0.70	0.90	0.79
		1	0.90	0.70	0.79
		avg	0.80	0.80	0.79
	Portuguese	0	0.81	0.88	0.84
		1	0.86	0.77	0.81
		avg	0.83	0.83	0.83

In addition, we evaluated the prediction errors of models that embed the MOL, as shown in Figure 3. Based on the receiver operating characteristic (ROC) curves analysis, the MOL model presented more unsuccessful predictions compared to the B + M model on the tree corpus in English, Spanish, and Portuguese. Therefore, even though the MOL model presents the best performance in terms of F-score, it also presents wrong predictions. Hence, the B + M is the best model in terms of successful predictions. We also observed that the extraction of pejorative terms and their context of use are factors relevant toward maximizing successful prediction and better performance for the B + M model. Hence, in future work, an automatic extractor of pejorative terms taking into account their context of use should be developed in order to maximize successful prediction and performance for the B + M model. Finally, as shown in Figures 4 and 5, the overall results of the confusion matrix present an expected low rate of false positives and false negatives. For Spanish, the rate of false positives and false negatives was lower compared to English and Portuguese. We show examples of misclassification cases in Table 16.

Note that, as shown in Table 16, the results of false negative and false positive cases are mostly composed of verbs and terms highly ambiguous. For example, “coxinha” in Brazilian Portuguese is very often pejoratively referred to people or groups with social privilege. While we propose the MOL lexicon as a solution toward tackling these challenges, which also showed relevant performance in terms of precision, it is still necessary to address some limitations of our approach including the limitation related to vocabulary coverage.

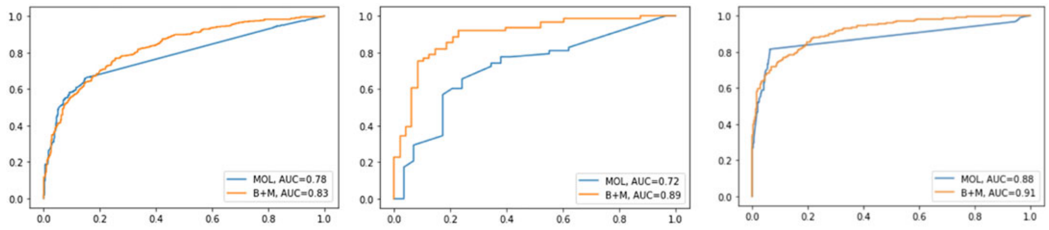


Figure 3. ROC curves for the MOL and B + M models. We evaluated these models on the OLID corpus of English tweets (left), the HatEval corpus of Spanish tweets (center), and the HateBR 2.0 of Brazilian Portuguese Instagram comments (right).

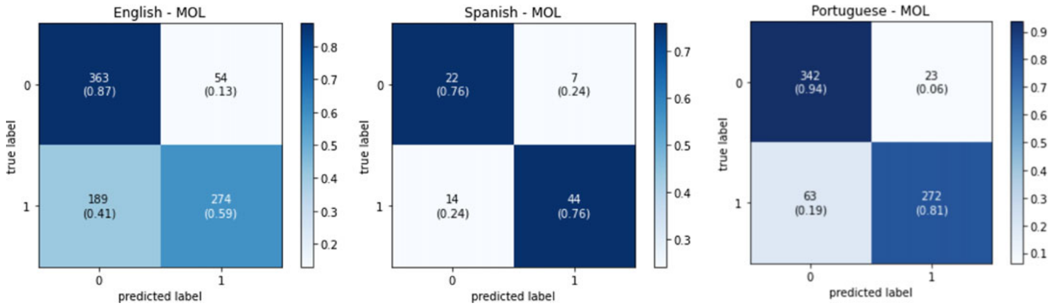


Figure 4. Confusion matrix for the MOL model. Specifically, we implemented the MOL on the OLID corpus of English tweets (left), the HatEval corpus of Spanish tweets (center), and the HateBR 2.0 corpus of Brazilian Portuguese Instagram comments (right).

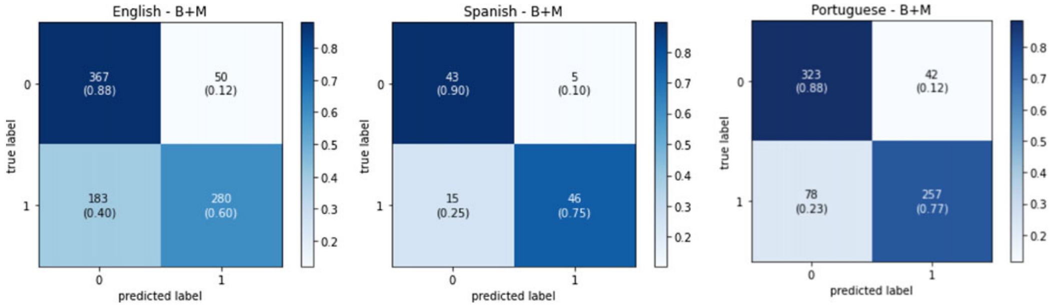


Figure 5. Confusion matrix for the B + M model. Specifically, We implemented the B + M on the OLID corpus of English tweets (left), the HatEval corpus of Spanish tweets (center), and the HateBR 2.0 corpus of Brazilian Portuguese Instagram comments (right).

Table 16. Examples of misclassification cases

	MOL		B+M	
Language	False negative	False positive	False negative	False positive
English	Predator	Make a scene	Abandon	Activate
	Sewer rat	Beat	Abhorrent	Appear
Spanish	Apestoso (stinky)	Acabar (finish)	Aborrecer (annoy)	Absurda (absurd)
	Bruja (wich)	Astillar (splinter)	Aborto (abortion)	Soletita (alone)
Portuguese	Coxinha (privileged group/person)	Família (family)	Aberração (aberration)	Africano (african)
	Macaca (racism)	Veneno (poison)	Avarento (miser)	Aliança (alliance)

Note: MOL, multilingual offensive lexicon.

https://doi.org/10.1017/nlp.2024.18 Published online by Cambridge University Press

9. Final remarks and future works

This paper provides context-aware and expert data resources for low-resource hate speech detection. Specifically, we introduced the HateBR 2.0, a large-scale expert annotated corpus for Brazilian Portuguese hate speech detection, and a new specialized lexicon manually extracted from this corpus, which was annotated with contextual information. It was also translated and culturally adapted by native speakers for English, Spanish, French, German, and Turkish. The HateBR 2.0 corpus consists of an updated version of HateBR, in which highly similar and one-word comments were replaced in order to improve its consistency and reliability. The proposed specialized lexicon consists of a context-aware offensive lexicon called MOL – multilingual offensive lexicon. It was extracted manually by a linguist and annotated by three different expert annotators according to a binary class: context-dependent and context-independent. A high human annotation agreement was obtained for both corpus and lexicon (75 percent and 73 percent Kappa, respectively). Baseline experiments were implemented on the proposed data resources, and results outperformed the current baseline hate speech dataset results for the Portuguese language reaching 85 percent at F1-score. Lastly, the obtained results demonstrate that models that embed our specialized lexicon are promising for hate speech detection in different languages. As future works, we aim to investigate methods to predict the pejorative connotation of terms and expressions according to their context of use.

Acknowledgments. The authors are grateful to Zohar Rabinovich (USA), Laura Quispe (Peru), Diallo Mohamed (Burkina Faso/Africa), Rebeka Mengel (German), and Ali Hürriyetoğlu (Turkey) for providing native-speakers translations and its cultural adaptations for the MOL and Fabiana Góes for helping us with annotation.

Competing interests. The author(s) declare none.

References

Albadi N., Kurdi M. and Mishra S. (2018). Are they our brothers? Analysis and detection of religious hate speech in the Arabic twittersphere. In *Proceedings of the 10th International Conference on Advances in Social Networks Analysis and Mining*, Barcelona, Spain, pp. 69–76.

Alfina I., Mulia R., Fanany M. I. and Ekanata Y. (2017). Hate speech detection in the Indonesian language: A dataset and preliminary study. In *Proceedings of the 9th International Conference on Advanced Computer Science and Information*, Bali, Indonesia, pp. 233–238.

AlKhamissi B., Ladhak F., Iyer S., Stoyanov V., Kozareva Z., Li X., Fung P., Mathias L., Celikyilmaz A. and Diab M. (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*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics, pp. 2109–2120.

Allan K. (2007). The pragmatics of connotation. *Journal of Pragmatics* **39**(6), 1047–1057.

Altemeyer B. (1996). *The Authoritarian Specter*. Cambridge, MA: Harvard University Press. pp. vii, 632.

Anderson L. and Lepore E. (2013). What did you call me? Slurs as prohibited words. *Analytic Philosophy* **54**(3), 350–363.

Badjatiya P., Gupta S., Gupta M. and Varma V. (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, Geneva, Swiss, pp. 759–760.

Basile V., Bosco C., Fersini E., Nozza D., Patti V., Rangel Pardo F. M., Rosso P. and Sanguinetti M. (2019). SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, Minneapolis, USA, pp. 54–63.

Blei D. M., Ng A. Y. and Jordan M. I. (2003). Latent dirichlet allocation. *Machine Learning Research* **3**(0), 993–1022.

Boishakhi F. T., Shill P. C. and Alam M. G. R. (2021). Multi-modal hate speech detection using machine learning. In *2021 IEEE International Conference on Big Data (Big Data)*, Held online, pp. 4496–4499.

Bonaldi H., Dellantonio S., Tekiroğlu S. S. and Guerini M. (2022). Human-machine collaboration approaches to build a dialogue dataset for hate speech countering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates, pp. 8031–8049.

Bousfield D. (2013). Face in conflict. *Journal of Language Aggression and Conflict* **1**(1), 37–57.

Bretschneider U. and Peters R. (2017). Detecting offensive statements towards foreigners in social media. In *Proceedings of the 50th Hawaii International Conference on System Sciences*, Hawaii, USA, pp. 2213–2222.

https://doi.org/10.1017/nlp.2024.18 Published online by Cambridge University Press

Brown P. F., Della Pietra V. J., deSouza P. V., Lai J. C. and Mercer R. L. (1992). Class-based *n*-gram models of natural language. *Computational Linguistics* **18**(4), 467–480.

Burnap P. and Williams M. L. (2014). Hate speech, machine classification and statistical modelling of information flows on Twitter: interpretation and communication for policy decision making. In *Internet, Policy and Politics Conference*, Oxford, United Kingdom, pp. 1–18.

Burnap P. and Williams M. L. (2016). Us and them: Identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data Science* **5**(11), 2–15.

Burnap P., Williams M. L., Sloan L., Rana O., Housley W., Edwards A., Knight V., Procter R. and Voss A. (2014). Tweeting the terror: Modelling the social media reaction to the woolwich terrorist attack. *Social Network Analysis Mining* **4**(206), 1–14.

Cao R., Lee R. K.-W., Chong W.-H. and Jiang J. (2022). Prompting for multimodal hateful meme classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates. pp. 321–332.

Caselli T., Basile V., Mitrović J., Kartoziya I. and Granitzer M. (2020). I feel offended, don't be abusive! Implicit/explicit messages in offensive and abusive language. In *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France, pp. 6193–6202.

Chen Y., Zhou Y., Zhu S. and Xu H. (2012). Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing*, Amsterdam, Netherlands, pp. 71–80.

Chung Y.-L., Kuzmenko E., Tekiroglu S. S. and Guerini M. (2019). CONAN–COunter NArratives through Nichesourcing: A multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, pp. 2819–2829.

Clair M. and Denis J. S. (2015). Sociology of racism. *The International Encyclopedia of the Social and Behavioral Sciences* **19**, 857–863.

Dadvar M., Trieschnigg D., Ordelman R. and de Jong F. (2013). Improving cyberbullying detection with user context. In Serdyukov P., Braslavski P., Kuznetsov S. O., Kamps J., Rüger S., Agichtein E., Segalovich I. and Yilmaz E., (eds), *Advances in Information Retrieval*. Berlin, Heidelberg, Berlin Heidelberg: Springer, pp. 693–696.

Davani A. M., Atari M., Kennedy B. and Dehghani M. (2023). Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics* **11**, 300–319.

Davidson T., Bhattacharya D. and Weber I. (2019). Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the 3rd Workshop on Abusive Language Online*, Florence, Italy, pp. 25–35.

Davidson T., Warmsley D., Macy M. W. and Weber I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International Conference on Web and Social Media*, Quebec, Canada, pp. 512–515.

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

Delphy C. (2000). Théories du patriarcat. In *Dictionaire critique du féminisme*. Presses Universitaires France, pp. 141–146.

Devlin J., Chang M.-W., Lee K. and Toutanova K. (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, Minnesota,* USA, pp. 4171–4186.

Dinakar K., Jones B., Havasi C., Lieberman H. and Picard R. (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems* **2**(3), 1–30.

Eyheramendy S., Lewis D. D. and Madigan D. (2003). On the Naive Bayes model for text categorization. In *Proceedings of the 9th International Workshop on Artificial Intelligence and Statistics*, Florida, USA, pp. 93–100.

Fersini E., Rosso P. and Anzovino M. (2018). Overview of the task on automatic misogyny identification. In *Proceedings of the 3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages co-located with 34th Conference of the Spanish Society for Natural Language Processing*, Sevilla, Spain, pp. 214–228.

Fleiss J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin* **76**(5), 378–382.

Fortuna P., Cortez V., Sozinho Ramalho M. and Pérez-Mayos L. (2021). MIN_PT: An European Portuguese lexicon for minorities related terms. In *Proceedings of the 5th Workshop on Online Abuse and Harms*, Held Online, pp. 76–80.

Fortuna P. and Nunes S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys* **51**(4), 1–30.

Fortuna, P., Rocha da Silva, J., Soler-Company, J., Wanner, L., and Nunes, S. (2019). A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the 3rd Workshop on Abusive Language Online*, Florence, Italy, pp. 94–104.

Gao L. and Huang R. (2017). Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, Varna, Bulgaria, pp. 260–266.

Ghosh Chowdhury A., Didolkar A., Sawhney R. and Shah R. R. (2019). ARHNet - Leveraging community interaction for detection of religious hate speech in Arabic. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, Florence, Italy, pp. 273–280.

https://doi.org/10.1017/nlp.2024.18 Published online by Cambridge University Press

Golbeck J., Ashktorab Z., Banjo R. O., Berlinger A., Bhagwan S., Buntain C., Cheakalos P., Geller A. A., Gergory Q., Gnanasekaran R. K., Gunasekaran R. R., Hoffman K. M., Hottle J., Jienjitlert V., Khare S., Lau R., Martindale M. J., Naik S., Nixon H. L., Ramachandran P., Rogers K. M., Rogers L., Sarin M. S., Shahane G., Thanki J., Vengataraman P., Wan Z. and Wu D. M. (2017). A large labeled corpus for online harassment research. In *Proceedings of the 9th ACM Web Science Conference*, New York, USA, pp. 229–233.

Guest E., Vidgen B., Mittos A., Sastry N., Tyson G. and Margetts H. (2021). An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, Held Online, pp. 1336–1350.

Guimarães S. S., Reis J. C. S., Ribeiro F. N. and Benevenuto F. (2020). Characterizing toxicity on facebook comments in Brazil. In *Proceedings of the 26th Brazilian Symposium on Multimedia and the Web*, Maranhão, Brazil, pp. 253–260.

Hasanuzzaman M., Dias G. and Way A. (2017). Demographic word embeddings for racism detection on Twitter. In *Proceedings of the 8th International Joint Conference on Natural Language Processing*, Taipei, Taiwan, pp. 926–936.

Jay T. and Janschewitz K. (2008). The pragmatics of swearing. *Journal of Politeness Research - Language Behaviour Culture* **4**(2), 267–288.

Jha A. and Mamidi R. (2017). When does a compliment become sexist? Analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the 2nd Workshop on NLP and Computational Social Science*, Vancouver, Canada, pp. 7–16.

Joulin A., Grave E., Bojanowski P. and Mikolov T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Valencia, Spain, pp. 427–341

Landis J. R. and Koch G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* **33**(1), 159–174.

Le Q. and Mikolov T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, Beijing, China, pp. 1188–1196.

Leite J. A., Silva D., Bontcheva K. and Scarton C. (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*, Suzhou, China, pp. 914–924.

Liu S. and Forss T. (2015). Text classification models for web content filtering and online safety. In *Proceedings of the 15th IEEE International Conference on Data Mining Workshop*, New Jersey, USA, pp. 961–968.

Ljubešić N., Erjavec T. and Fišer D. (2018). Datasets of Slovene and Croatian moderated news comments. In *Proceedings of the 2nd Workshop on Abusive Language Online*, Brussels, Belgium, pp. 124–131.

Manning C. and Schutze H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press. ISBN 0262133601.

McHugh M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica* **22**(3), 276–282.

Mikolov T., Chen K., Corrado G. and Dean J. (2013). Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations*, Arizona, USA.

Nadeem M., Bethke A. and Reddy S. (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*, Held Online, pp. 5356–5371.

Njagi D., Zuping Z., Hanyurwimfura D. and Long J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering* **10**, pp. 215–230.

Nobata C., Tetreault J., Thomas A., Mehdad Y. and Chang Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, Republic and Canton of Geneva, CHE, pp. 145–153.

Ousidhoum N., Lin Z., Zhang H., Song Y. and Yeung D.-Y. (2019). Multilingual and multi-aspect hate speech analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, China, pp. 4675–4684.

Ozalp S., Williams M. L., Burnap P., Liu H. and Mostafa M. (2020). Antisemitism on Twitter: Collective efficacy and the role of community organisations in challenging online hate speech. *Social Media + Society* **6**(2), 01–20.

Pavlopoulos J., Malakasiotis P. and Androutsopoulos I. (2017). Deep learning for user comment moderation. In *Proceedings of the 1st Workshop on Abusive Language Online*, British Columbia, Canada, pp. 25–35.

Pitenis Z., Zampieri M. and Ranasinghe T. (2020). Offensive language identification in Greek. In *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France, pp. 5113–5119.

Poletto F., Basile V., Sanguinetti M., Bosco C. and Patti V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation* **55**(3), 477–523.

Rae L. (2012). Beyond belief: Pragmatics in hate speech and pornography. In *Speech and Harm: Controversies Over Free Speech*, pp. 72–93.

Robinson B. B. E., Bacon L. C. and O'reilly J. (1993). Fat phobia: Measuring, understanding, and changing anti-fat attitudes. *International Journal of Eating Disorders* **14**(4), 467–480.

https://doi.org/10.1017/nlp.2024.18 Published online by Cambridge University Press

Schmidt A. and Wiegand M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the 5th International Workshop on Natural Language Processing for Social Media*, Valencia, Spain, pp. 1–10.

Scholkopf B. and Smola A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge: MIT Press.

Silva L., Modal M., Correa D., Benevenuto F. and Weber I. (2016). Analyzing the targets of hate in online social media. In *Proceedings of the 10th International AAAI Conference on Web- and Social Media,* Cologne, Germany, pp. 687–690.

Sim J. and Wright C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy* **85**(3), 257–268.

Steimel K., Dakota D., Chen Y. and Kübler S. (2019). Investigating multilingual abusive language detection: A cautionary tale. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, Varna, Bulgaria, pp. 1151–1160.

Thapa S., Jafri F., Hürriyetoğlu A., Vargas F., Lee R. K.-W. and Naseem U. (2023). Multimodal hate speech event detection - shared task 4, CASE 2023. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text*, Varna, Bulgaria, pp. 151–159.

Trajano D., Bordini R. and Vieira R. (2023). OLID-BR: Offensive language identification dataset for Brazilian Portuguese. *Language Resources & Evaluation*. https://doi.org/10.1007/s10579-023-09657-0.

Van Hee C., Lefever E., Verhoeven B., Mennes J., Desmet B., De Pauw G., Daelemans W. and Hoste V. (2015a). Detection and fine-grained classification of cyberbullying events. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, Hissar, Bulgaria, pp. 672–680.

Van Hee C., Lefever E., Verhoeven B., Mennes J., Desmet B., De Pauw G., Daelemans W. and Hoste V. (2015b). Detection and fine-grained classification of cyberbullying events. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, Hissar, Bulgaria, pp. 672–680.

Vargas F., Carvalho I., Rodrigues de Góes F., Pardo T. and Benevenuto F. (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*, Marseille, France, pp. 7174–7183.

Vargas F., Goes G., Carvalho I., Benevenuto F. and Pardo T. (2021). Contextual-lexicon approach for abusive language detection. In *Proceedings of the Recent Advances in Natural Language Processing*, Held Online, pp. 1438–1447.

Vidgen B. and Derczynski L. (2021). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE* **15**(12), 1–32.

Walby S. (1990). *Theorizing Patriarchy*. Oxford: Basil Blackwell. Volume 1. p. 229. ISBN: 9780631147688.

Warner W. and Hirschberg J. (2012). Detecting hate speech on the world wide web. In *Proceedings of the 2nd Workshop on Language in Social Media*, Montréal, Canada, pp. 19–26.

Waseem Z. and Hovy D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics - Student Research Workshop*, California, USA, pp. 88–93.

Westwood S. J., Iyengar S., Walgrave S., Leonisio R., Miller L. and Strijbis O. (2018). The tie that divides: Cross-national evidence of the primacy of partyism. *European Journal of Political Research* **57**(2), 333–354.

Wilson W. (1999). *The Bridge over the Racial Divide: Rising Inequality and Coalition Politics.* Aaron Wildavsky forum for public policy. University of California Press, p. 163. ISBN: 9780520229297.

Xiang G., Fan B., Wang L., Hong J. and Rose C. (2012). Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, New York, United States, 1980-1984.

Xu J.-M., Jun K.-S., Zhu X. and Bellmore A. (2012). Learning from bullying traces in social media. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Montréal, Canada, pp. 656–666.

Zampieri M., Malmasi S., Nakov P., Rosenthal S., Farra N. and Kumar R. (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*, Minnesota, USA, pp. 1415–1420.

Zannettou S., Finkelstein J., Bradlyn B. and Blackburn J. (2020). A quantitative approach to understanding online antisemitism. In *Proceedings of the 14th International AAAI Conference on Web and Social Media*, Held Online, pp. 786–797.

Zhong H., Li H., Squicciarini A., Rajtmajer S., Griffin C., Miller D. and Caragea C. (2016). Content-driven detection of cyberbullying on the Instagram social network. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, New York, United States, pp. 3952–3958.

Zhu J., Lee R. K.-W. and Chong W. H. (2022). Multimodal zero-shot hateful meme detection. In *14th ACM Web Science Conference*, Barcelona Spain, pp. 382–389.

Cite this article: Vargas F, Carvalho I, Pardo TAS and Benevenuto F. Context-aware and expert data resources for Brazilian Portuguese hate speech detection. *Natural Language Processing* https://doi.org/10.1017/nlp.2024.18