HateBR: Large expert annotated corpus of Brazilian Instagram comments for abusive language detection

Anonymous ACL submission

Abstract

Due to the severity of the social media abusive comments in Brazil, and the lack of research in Portuguese, this paper provides the first large-scale annotated corpus of Brazilian Instagram comments for hate speech and offensive language detection on the web and social media. The HateBR corpus was collected from Brazilian Instagram comments of political personalities and manually annotated by specialists, being composed of 7,000 documents annotated according to three different layers: a binary classification (offensive versus non-offensive comments), offense-level classes (highly, moderately, and slightly offensive messages), as well as nine hate speech targets (xenophobia, racism, homophobia, sexism, religious intolerance, partyism, apology to the dictatorship, antisemitism, and fatphobia). Each comment was annotated by three different annotators and achieved high inter-annotator agreement.

1 Introduction

In recent years, the use of social media has increased and provided several advantages for society, as virtual human interactions that enable people from anywhere to connect with anyone (Leite et al., 2020). Nevertheless, offenses and hate speech content have become pervasive on online platforms, as well as a considerable concern for government organizations around world (Zampieri et al., 2019).

Abusive language detection has attracted interest from different institutions and has become an important research topic (Pitenis et al., 2020; Zannettou et al., 2020; Çöltekin, 2020; Steimel et al., 2019; Hasanuzzaman et al., 2017; Guest et al., 2021; Schmidt and Wiegand, 2017). Moreover, although the investigation related to hate speech is undoubtedly an important research topic, the proposition of automated hate speech and offensive language approaches also have its implications for unprejudiced society concerning race, gender, religion, origin, etc. Accordingly, identifying hateful declarations may bolster security in revealing individuals harboring malicious intentions towards specific groups (Gao et al., 2017).

In Brazil, hate speech is prohibited, although the regulation is not effective due to the high difficulty of identifying, quantifying and classifying abusive comments. Figure 1 shows the occurrence of registered hate crimes in Federative units of Brazil during the 2018 year per 100,000 inhabitants[1]. Note that the chart illustrates the high incidence of hate crimes in the whole Brazilian territory, however in the the south the situation is even worse.

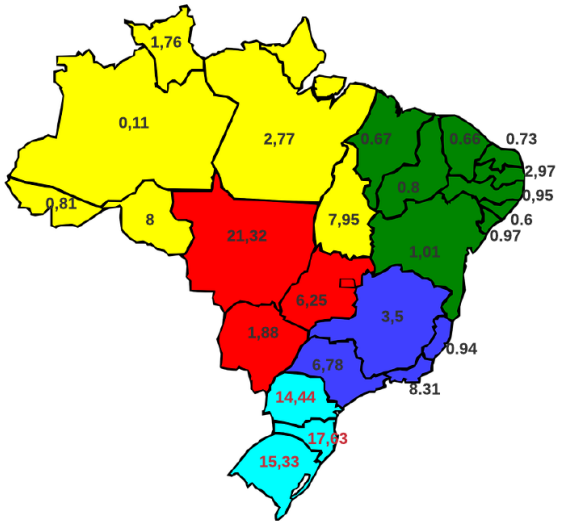

Figure 1: Hate crimes registered in Brazil during 2018. The colors indicate the Brazilian regions, in which the yellow is northern, the red is the center-west, the green is the northeast, the dark blue is the southeast, and the light blue is the south.

On the report of Mesquita (2018), the *Safernet* [2] Non-Governmental Organization, which operates in cooperation with Public Organizations in Brazil,

[1] https://www.wordshealtheworld.com/wp-content/uploads/2019/10/HATE-MAP-OF-BRAZIL-REPORT-1.pdf

[2] https://new.safernet.org.br/

as well as companies such as Google, Facebook, and Twitter, proposed a collection of data on actions that violate human rights. They state very clearly that denunciations with xenophobia content grew by 2,369.5%; apology and encouragement to crimes against life, 630.52%; Neo-Nazism, 548.4%; homophobia, 350.2%; racism, 218.2%; and religious intolerance, 145.13% [3]. In Buarque and Cretton (2021), the authors present an overall number of hate crimes registered by the Brazilian police in 2019. In total, the Brazilian police recorded a total of 12,334 hate crimes, whereby 8,979 (72.80%) were hate crimes motivated by bias based on race, 1,732 (14.04%) sexual orientation-related crimes (towards LGBTQIA+ community), 1,314 (10.65%) gender-related crimes (targeting women: femicides), 226 (1.83%) religious-related crimes, and 83 (0.67%) origin-related crimes (xenophobia).

Due to the relevance of this topic and the severity of the hate speech context in Brazil, the proposition of a reliable annotated dataset is essential to carry out experiments on automatic offensive language and hate speech detection. Nevertheless, the annotation process of abusive content is intrinsically challenging, bearing in mind that what is considered offensive is influenced by pragmatic (contextual) factors, and people may have different opinions on an offense. Moreover, another relevant contribution of a skilled and well-defined annotation schema consists of the considerable impact between the consistency and quality of the data with the performance of the derived machine learning classifiers. In NLP, subjective tasks, such as sentiment analysis, offensive language, and hate speech detection, present high complexity and a variety of challenges. Therefore, an expert annotation schema is relevant, mainly because a reliable annotation approach offers an adequate characterization of specificities that consequently improves the quality of the data labeling and the NLP tools based on them.

Accordingly, since the online abusive comments situation in Brazil is currently the biggest social, criminal and political problem, as well as the lack of research in this area for the Portuguese language, this paper describes the first large-scale annotated corpus of Brazilian Instagram comments for hate speech and offensive language detection, and provides an expert annotation schema for abusive language detection on social medias. The HateBR corpus was collected from different accounts of political personalities on Instagram. The political context was chosen due to the identification of several types of serious offenses and hate attacks in different groups. The entire annotation schema was proposed and annotated by different specialists: a linguist, a hate speech skilled, as well as machine learning engineers, and handled by accurate guidelines and training steps, in order to ensure the same understanding of the tasks, and bias minimization.

More precisely, the main contributions of this paper are:

- The first large-scale expert annotated corpus for Brazilian abusive language detection, composed of 7,000 Instagram comments annotated in three different layers (offensive versus non-offensive; offensive comments sorted into highly, moderately, and slightly offensive levels; and, nine hate speech targets: xenophobia, racism, homophobia, sexism, religious intolerance, partyism, an apology to dictatorship, antisemitism, and fatphobia).

- An expert annotation schema for hate speech and offensive language detection on social media, which is language and domain-independent (although it only has been evaluated for the Brazilian Portuguese language).

In what follows, we briefly introduce the main related work. Section 3 describes the HateBR corpus development, as well as the proposed skilled annotation schema and its evaluation. Sections 4 and 5 show the HateBR corpus statistics, as well as the final remarks, are presented.

2 Related Work

Most of hate speech and offensive language corpora are proposed for the English language (Zampieri et al., 2019; Basile et al., 2019; de Gibert et al., 2018; Fersini et al., 2018; Davidson et al., 2017; Gao and Huang, 2017; Jha and Mamidi, 2017; Golbeck et al., 2017; Waseem and Hovy, 2016). Nevertheless, Chung et al. (2019) and Ousidhoum et al. (2019) proposed corpora of Facebook and Twitter annotated data for Islamophobia, sexism, homophobia, religion intolerance and disability detection in French language. For the German language, Bretschneider and Peters (2017) provide an

[3] https://www.bbc.com/portuguese/brasil-46146756

anti-foreigner prejudice corpus with 5,836 Facebook posts hierarchically annotated for slightly and explicitly/substantially offensive language according to six targets: foreigners, government, press, community, other, and unknown. For the Greek language, Pitenis et al. (2020) and Pavlopoulos et al. (2017) supply annotated datasets of Twitter and Gazeta posts for offensive content detection. For the Slovene and Croatian languages, Ljubešić et al. (2018) describe a large-scale dataset composed of 17,000,000 posts, with 2% of abusive language on a leading media company website. In Arabic, Albadi et al. (2018) presents a new dataset with 6,136 twitter posts, which is annotated in religion intolerance subcategories. For Indonesian language, Alfina et al. (2017) and Ibrohim and Budi (2018) provide a hate speech annotated corpus from Twitter data.

For Portuguese, Fortuna et al. (2019) adopts the definition of hate speech proposed by Fortuna and Nunes (2018), and propose a new dataset composed of 5,668 tweets in European and Brazilian Portuguese, as well as automated methods using a hierarchy of hate to identify social groups of discrimination. Moreover, de Pelle and Moreira (2017) provide a new dataset composed of 1,250 comments in Brazilian Portuguese collected from G1 Brazilian online newspaper [4], which was annotated only with a binary class: offensive and non-offensive comments.

3 HateBR Corpus Development

In this section, we describe the building of the proposed annotated corpus for online Brazilian Portuguese abusive language detection.

3.1 Proposed Approach Overview

The entire process of corpus development occurred for approximately six months, more specifically, between August 2020 to January 2021. This project was performed by different specialists (e.g., a linguist, hate speech specialists, and machine learning engineers) to ensure the reliability and quality of the annotated data. Figure 2 exhibits the step-to-step for HateBR corpus development.

As shown in Figure 2, firstly, the application domain was defined, and the political domain was the chosen one. In the second step, we defined the criteria for the selection of the Instagram accounts, which are described as follows: 500 comments

[4] https://g1.globo.com/

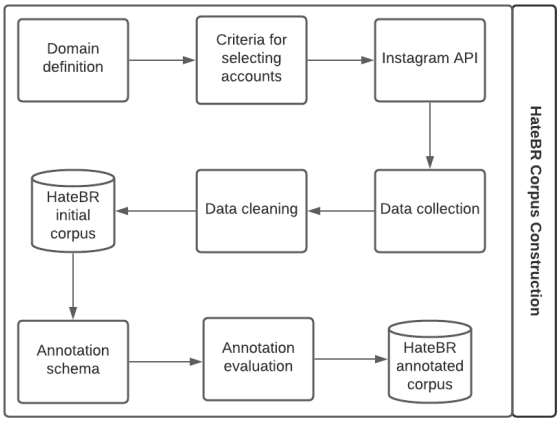

Figure 2: HateBR corpus development.

were extracted from 5 distinct posts of six different public accounts, being three liberal-party and three conservative-party accounts, whereby four were women and two were men. In the third step, we used an API to automatically extract Instagram comments using the defined criteria for selected accounts. The collected comments were published in the second half of 2019, and we selected comments across the months. For example, for the same account, 500 comments were collected from a post published in August 2019. Then, the other 500 comments were collected from a second post published in September 2019, etc. After the data collection, we remove noise, such as links, characters without semantic value, and also comments that presented only emoticons, laughs (kkk, hahah, hshshs) or that referred to other accounts (e.g., @namesomeone). We must point out that hashtags and emotions were kept.

The selection of annotators was performed using as criteria the higher levels of education (Ph.D.), as well as specialists, such as a linguist and skilled hate speech computer scientists. Moreover, we selected annotators with diversified profiles, such as distinct political orientations and colors in order to minimize bias. Finally, we began the annotation process and proposed an annotation schema, determining more precisely the offensive and hate speech classification. We also evaluated the annotation schema using inter-human annotation agreement metrics, such as Kappa (McHugh, 2012; Sim and Wright, 2005) and Fleiss (Fleiss, 1971), which obtained a high inter-annotator agreement (75% Kappa and 74% Fleiss).

Moving forward, the proposed HateBR corpus

was annotated using different levels of annotation. Firstly, each one of the 7,000 Instagram comments was annotated with the following binary classes: offensive comment (3,500) versus non-offensive comment (3,500). Moreover, offense-level classes (highly, moderately, and weakly) were also annotated for the 3,500 Instagram comments classified as offensive in the previous stage. Finally, offensive comments, which present any offenses against minority groups, received a hate speech label in nine identified hate speech targets (xenophobia, racism, homophobia, sexism, religious intolerance, partyism, apology to the dictatorship, antisemitism, and fatphobia).

3.2 Data Collection

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[5]. Each person has an account with shared photos and it is possible for others to like, comment, save and share this information. Therefore, considering that Instagram is a representative online platform for showcasing digital influencers in Brazil, as well as a powerful environment for mass media, we automatically collected Instagram comments for building the the HateBR corpus. Tables 1 and 2 show the data collection statistics.

Table 1: Data collection statistics.

Data	**Total**
Amount of comments extracted	15,000
Amount of comments removed (noise)	8,000
Final Corpus	7,000

Table 2: Instagram account profiles

Profile	**Description**
Gender	4 women and 2 man
Political	3 liberals and 3 conservative

As shown in Tables 1 and 2, in order to corroborate with our proposal of variables balancing (e.gender and political party), we collected 15,000 comments from six public Instagram accounts divided into three politicians from the liberal party, as well as three politicians from the conservative party, being four women and two men. Subsequently, we removed 8,000 noise comments, or in order words, comments that presented only emoticons laughs or accounts sign. Moreover, we selected the most popular posts for each account during the second half of 2019, being five posts for each account and 500 comments for each post.

3.3 Conceptualization

According to Post (2009), the abusive language detection tasks present a conceptual difficulty of distinguishing hateful and offensive expressions from expressions that merely denote dislike or disagreement. Therefore, bearing in mind the underlying difficulty to abusive language classification, we more accurately defined offensive comments using distinct definitions for offensive words, swear words and hate speech, which we further describe in detail in Sections 3.3.1 and 3.3.2.

3.3.1 Offensive and Swear Words

In this paper, we defined offensive and swearing as follows:

- *Offensive words*: an offense consists of a pejorative term or expression that intends to undermine or disparage any of the following social aspects: moral, appearance, physical, psychological health, sexual behavior, and orientation, intellectual, economic, religious, and political aspects.

- *Swear words*: a swear word consists of a pejorative term or expression used to convey a hateful opinion, with high aggressive value and great potential to generate negative reactions to the interlocutor.

Table 3 shows examples of offensive and swear words, well as offensive and swear expressions extracted from the proposed HateBR corpus. We describe the terms originally written in Portuguese and their translation to English.

Table 3: Offensive and swear words examples.

Type	**Term/Expression**	**Translation**
Offenses	Mentiroso	Liar
Offenses	Vagabunda	Slut
Offenses	Canalha	Scoundrel
Offenses	Desgraçado	Wretched
Offenses	Facista	Fascist
Swear words	Vai Tomar no Cú	Go Fuck Yourself
Swear words	Vergonha na Cara	Shame on you
Swear words	Vai para o Inferno	Go to Hell
Swear words	Filho da Puta	Son of a bitch
Swear words	Foda-se	Fuck it

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

3.3.2 Hate Speech

In this paper, we assume that hate speech is a language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity, or other, and it may occur with different linguistic styles, even in subtle forms or when humor is used (Fortuna and Nunes, 2018). Accordingly, we identify nine hate speech targets, which we describe in detail in what follows:

1. *Antisemitism*: The antisemitism definition adopted by the IHRA[6] in 2016 states that "Antisemitism is a certain perception of Jews, which may be expressed as hatred toward Jews. Rhetorical and physical manifestations of antisemitism are directed toward Jewish or non-Jewish individuals and/or their property, toward Jewish community institutions and religious facilities. Manifestations might include the targeting of the state of Israel, conceived as a Jewish collectivity". An example follows: *Que escroto caquético! É a velha hipocrisia judaica no mundo dos pilantras monetários. Judeu dos infernos!* ("What a cachectic scrotum! It's the old Jewish hypocrisy in the world of monetary hustlers. Jew from hell!").

2. *Apology for the Return of the Military Dictatorship*: According to the Brazilian Penal Code[7], the apology to dictatorship consists of comments that incite animosity between the armed forces, social classes, and civil institutions. An example of the apology dictatorship comment is described as follows: *Intervenção Militar já !!! Acaba Supremo Tribunal Federal, não serve pra nada mesmo...* ("Military intervention now !!! "It's over Supreme Court[8], it is of no use at all...").

3. *Fatphobia*: Robinson et al. (1993) define fatphobia as negative attitudes towards and stereotypes about fat people. An example follows: *Velha barriguda e bem folgada, heim? Porca rosa, Peppa!* ("Old bellied and very loose, huh? Pink Nut, Peppa[9]!").

4. *Homophobia*: Homophobia[10] is considered an irrational fear or aversion to homosexuality, or, in other words, to lesbian, gay and bisexual people based on prejudice. An example follows: *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.* ("Whoever said that must be a global who does not come out of the closet :(:(and has that desire :(:(at that time they had to intervene here and apply the law against them. ... It gives the ass a lot").

5. *Partyism*: Westwood et al. (2018) demonstrated that partyism influences behaviors and non-political judgment. According to a professor at Harvard University, "partyism" is a form of hostility and prejudice that operates across political lines (Sunstein, 2016). In our corpus, the most relevant occurrence of hate speech consists of partyism, as the following example: *Os petralhas colocaram sua corja em todos os lugares, não salva ninguém, que tristeza .. Esquerda parasita lixo.* ("The petralhas[11] put their crowds everywhere, no one may be saved, how sad. They are parasite and trash").

6. *Racism / Racial Segregation*: According to Wilson (1999), racism consists of an ideology of racial domination. In the same settings, Clair and Denis (2015) argue that racism 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. Through the process of racialization, perceived patterns of physical difference, such as skin color or eye shape, are used to differentiate groups of

[6] International Holocaust Remembrance Alliance (IHRA) unites governments and experts to strengthen, advance, and promote Holocaust education, research, and remembrance, as well as uphold the commitments of the 2000 Stockholm Declaration.

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

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

[9] *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, and her family and friends.

[10] According to European Institute for Gender Equality <https://tinyurl.com/4yca8vpm>.

[11] *Petralha* is a deep Brazilian culture rooted pejorative name used by conservative politicians to originally define liberal politicians associated to a specific political party.

people, thereby constituting them as "races"; racialization becomes racism when it involves the hierarchical and socially consequential valuation of racial groups. In our corpus, we found a wide range of offenses related to racial discrimination, such as "monkey" and "cheetah". An example of a racist comment is as follows: *E uma chita ela né! Opssss, uma chata.* ("And she is a cheetah right! Opssss, a boring girl."[12]).

7. *Religious Intolerance*: As maintained by Altemeyer and Altemeyer (1996), 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 toward out-groups. For instance, observe the following comments: *Pastor dos Infernos.* ("Pastor of the Church from Hell"), and *O chamado crente do demônio, né?* ("The so-called Christian of the devil").

8. *Sexism*: sexism behavior is mostly related to patriarchy that according to literature consists of a system of social structures that are related to each other and that allow men to exploit women. Nonetheless, Delphy (2000) complements that women are seen as objects of sexual satisfaction of men, reproducers of heirs, labor force, and new breeders. The following example was extracted from the proposed corpus: *Cala esse bueiro de falar merda sua vagabunda safada.* ("Shut that manhole to talk shit you slut and barefaced ").

9. *Xenophobia*: Oliveira (2019) describes xenophobia as a form of prejudice, which is manifested through discriminating actions and hate against foreigners. An example follows: *Ele está certo. Vai ter um monte de argentino famінto invadindo o Brazil.* ("He is right. There will be a lot of hungry Argentine people invading Brazil").

3.4 Annotation Schema

We propose a new annotation schema, which provides three different layers of annotation, determining more precisely how to classify offensive comments versus non-offensive comments, as well as offense-level classes (such as highly, moderately, and slightly offensive), and offensive comments with hate speech targets versus offensive comments without hate speech targets. Figure 3 shows the proposed annotation schema.

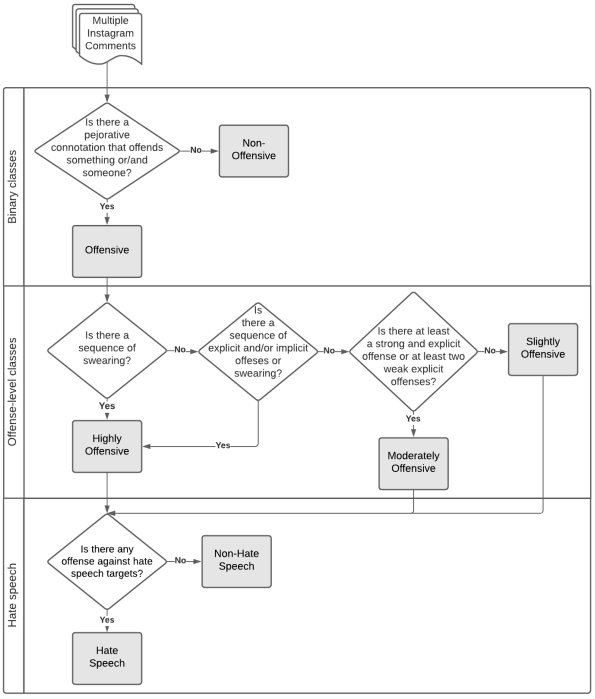

Figure 3: HateBR annotation schema.

Note that our annotation schema is divided into three layers: (i) binary classification, (ii) offense-level classes, and (iii) hate speech. Firstly, we annotated the corpus using binary classes: offensive or non-offensive comments. Subsequently, we selected only offensive comments obtained from the previous annotation layer and classified them into offense levels. The offense-level classification consists of three classes: highly, moderately, and slightly offensive. Finally, in the third layer, we annotated offensive comments that presented hate speech targets (one of the 9 targets that we already introduced).

Moving forward, as shown in Figure 3, we initially assume that, if in a comment there is at least one explicit or implicit offense and swearing words used with pejorative intention (considering the vocabulary described in the Multilingual Offensive Lexicon (OMITTED DUE TO DOUBLE-BLIND), which we introduce in the next paragraph), then this comment is offensive. Otherwise, if in a comment there is not an explicit or implicit offense and swearing words (according to Multilingual Offensive Lexicon), then this comment is not offensive.

[12] In Portuguese, the words that refer to "cheetah" and "boring" differ on only one letter. This is the way the speaker pretends a mistake in writing.

Comments annotated as offensive were also annotated according to three offense levels: highly, moderately, and slightly offensive. We assume that, if the comment contains at least one explicit swearing word or a sequence of explicit or implicit offense words (according to Multilingual Offensive Lexicon), then this comment is highly offensive. In the same setting, we also assume that, if the comment contains at least one explicit and strong offense word or expression (according to Multilingual Offensive Lexicon), then this comment is moderately offensive. At last, we assume that, if the comment does not present the previously defined criteria, then this comment is slightly offensive. Finally, the identification of hate speech content was accomplished according to the nine hate speech targets that were previously introduced. Tables 4, 5 and 6 show examples of the annotated comments according to the annotation layers proposed for the HateBR corpus.

Instagram's comments	Binary classes classification			
	Annotator 1	Annotator 2	Annotator 3	Class
este lixo ...	1	1	0	**1**
De onde você tirou esta informações eu procurei em todas redes social e não encontrei nada poderia informar suas fontes como link da matéria ou página do processo	0	0	0	**0**
Essa mulher é doente.pilantra!	1	1	1	**1**
Porque sera que ela diz ter pena? Se analisar bem isso pode ser visto como uma ameaça.	1	0	0	**0**
Vagabunda. Comunista. Mentirosa. O povo chileno nao merece uma desgraça desta.	1	1	1	**1**

Figure 4: Binary classification: offensive versus non-offensive comments. We labeled offensive comments as (1) and non-offensive comments as (0).

Instagram's comments	Offense-level classes classification			
	Annotator 1	Annotator 2	Annotator 3	Class
este lixo ...	2	2	3	**2**
Essa mulher é doente.pilantra!	2	3	3	**3**
Devolva o dinheiro	2	1	1	**1**
Vagabunda. Comunista. Mentirosa. O povo chileno nao merece uma desgraça desta.	3	3	3	**3**

Figure 5: Offense-level classes: slightly offensive, moderately offensive, and highly offensive. We labeled comments highly offensives as (3), comments moderately offensives as (2), and comments slightly offensives as (1).

During the annotation process, the specialists were mainly supported by two resources: a well-structured guideline, which presented the main concepts defined in Section 3.3, as well as the Multilingual Offensive Lexicon, that was cited previously [13], composed of 1,000 explicit and implicit pejorative expressions annotated with contextual information. For example, the term *"vadia"* ("slut") consists of a context-independent offensive term because it is not used in a non-offensive context. On the other hand, the term *inútil* ("useless") is a context-dependent offensive term. Note that this last term is classified as context-dependent offensive because it also may be employed in a non-offensive context, such as "this smartphone is useless" or "the process is useless for this task". Multilingual Offensive Lexicon (OMITTED DUE TO DOUBLE-BLIND) was extracted from the proposed corpus in this paper (the HateBR corpus), and each term or expression was annotated by three different annotators, obtaining a high human agreement score (73% Kappa).

[13] Available at (OMITTED DUE TO DOUBLE-BLIND)

Instagram's comments	Hate speech classification			
	Annotator 1	Annotator 2	Annotator 3	Class
Vagabunda. Comunista. Mentirosa. O povo chileno nao merece uma desgraça desta.	8	5	8	**5;8**
Judeu dos infernos ??	1	1	1	**1**
Vermelha tá a bunda desses vermes !!!!!!	5	5	5	**5**
Conheco uma vagabunda de longe!!! ??????	8	8	8	**8**
Na terra em que essa mulher africana é bonita as trans são mulheres. Surrealismo clássico.	4	4	6	**4**
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!!!	7	7	7	**7**
Eu tenho pena é de você não ter sido da época do grande General Pinochet que lamentavelmente deixou sobrar este resquício de desgraça.	2	2	2	**2**

Figure 6: Hate speech: we identify nine hate speech targets: antisemitism, apology to the dictatorship, fatphobia, homophobia, partyism, racism, religious intolerance, sexism, and xenophobia. We labeled antisemitism as (1), apology to the 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). We must point out that a couple of comments belongs to more than a group. For example, the comment *comunista vagabunda e safada* ("communist slut and shameless") was classified as partyism and sexism, therefore, it was labeled as (5;8).

.

3.5 Annotators Background

Due to the degree of complexity of abusive language detection tasks, mainly because it involves a highly politicized domain, we decided to select specialists at higher levels of education. Moreover, in order to minimize bias, we selected annotators from different political orientations, as well as different colors, as shown in Table 4.

Table 4: Annotator profiles.

Profile	Description
Education	PhD or PhD candidate
Gender	Feminine
Political	Liberal and Conservative
Color	White and Black

3.6 Evaluation

The annotation was carried out by three different specialists. Each comment was annotated by each one in order to guarantee the reliability of the annotation process. Moreover, we computed inter-annotator agreement, using two different metrics:

Kappa (McHugh, 2012; Sim and Wright, 2005) and Fleiss (Fleiss, 1971). Table 5 and 6 shown the obtained results.

Additionally, we performed two evaluation steps for the hate speech target classification. Firstly, the comments annotated with hate speech targets by at least two annotators were immediately selected. Subsequently, the comments annotated with hate speech targets labels by only one annotator were submitted on a new checking step, where an experienced linguist decided if that label was applicable or not.

Table 5: Cohen's kappa.

Peer Agreement	AB	BC	CA	AVG
Binary classification	0.76	0.72	0.76	**0.75**
Offense-level classes	0.46	0.44	0.50	**0.47**

Table 6: Fleiss' kappa.

Fleiss' kappa	ABC
Binary classification	**0.74**
Fine-grained offenses	**0.46**

Kappa values range from 0 to 1, and there are possible interpretations of these values (Landis and Koch, 1977). Table 5 shows the obtained results using Cohen's kappa. Note that we obtained a high inter-annotator agreement for binary classes 75%, and 47% for offense-level classes.

In the same settings, the Fleiss evaluation measure is an extension of Cohen's kappa for the case where there are more than two annotators (or methods). This means that Fleiss' kappa is applied for a wide variety of annotators that provide categorical ratings, such as binary or nominal scale, for a fixed number of items (Fleiss, 1971).

4 HateBR Dataset Statistics

As a result, we present the dataset statistics for the proposed HateBR annotated corpus. Tables 7, 8 and 9 shonw the results.

Table 7: Binary classification.

Classes	Total
Non-Offensive	3,500
Offensive	3,500
Total	7,000

Overall, the HateBR corpus is composed of 7,000 document-level annotated. Firstly, the corpus was annotated using a binary classification: 3,500 offensive comments versus 3,500 non-offensive comments. Moreover, 3,500 comments identified as offensive were also classified according to offense-level classes, being 1,678 slightly offensives, 1,044 moderately offensive, and 778 highly offensive.

Table 8: Offensive-level Classes.

Offense-level Classes	Total
Slightly Offensive	1,678
Moderately Offensive	1,044
Highly Offensive	778
Total	3,500

Furthermore, offensive comments were also categorized according to the nine discrimination groups in order to identify hate speech targets, as shown in Table 9.

Table 9: Hate speech targets.

Hate Speech Targets	Total
Partyism	496
Sexism	97
Religion Intolerance	47
Apology to Dictatorship	32
Fat Phobia	27
Homophobia	17
Racism	8
Antisemitism	2
Xenophobia	1

5 Final Remarks

Since the online abusive comments situation in Brazil is currently the biggest research, social and criminal problem, this paper provides the first large-scale expert annotated corpus of Brazilian Portuguese Instagram comments for online abusive language detection. The proposed corpus consists of 7,000 documents annotated with three different layers. The first layer consists of comments annotated as offensive versus non-offensive. In the second layer, offensive comments were annotated according to the following offense-level classes: slightly, moderately, and highly offensive. In the third layer, offensive comments were also classified according to nine hate speech targets, which identify discriminatory content, such as xenophobia, racism, homophobia, sexism, religious intolerance, partyism, an apology to the dictatorship, antisemitism, and fatphobia. We evaluate the proposed annotation schema, and a high human annotation agreement was obtained (75% Kappa and 74% Fleiss).

References

N. Albadi, M. Kurdi, and S. Mishra. 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*, pages 69–76, Barcelona, Spain.

Ika Alfina, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. 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*, pages 233–238, Bali, Indonesia.

Robert A Altemeyer and Bob Altemeyer. 1996. *The authoritarian specter*. Harvard University Press.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 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*, pages 54–63, Minnesota, USA.

Uwe Bretschneider and Ralf Peters. 2017. Detecting offensive statements towards foreigners in social media. In *Proceedings of the 50th Hawaii International Conference on System Sciences*, pages 2213–2222, Hawaii, USA.

Beatriz Buarque and Marcio Cretton. 2021. Hate map of brazil: Insights and recommendations for policy. *Words Heal the World*, 1:1–67.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 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*, pages 2819–2829, Florence, Italy.

Matthew Clair and Jeffrey S Denis. 2015. Sociology of racism. *The International Encyclopedia of the Social and Behavioral Sciences*, 19:857–863.

Çağrı Çöltekin. 2020. A corpus of Turkish offensive language on social media. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6174–6184, Marseille, France.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International Conference on Web and Social Media*, pages 512–515, Quebec, Canada.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online*, pages 11–20, Brussels, Belgium.

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.

Christine Delphy. 2000. Théories du patriarcat. In *Dictionaire critique du féminisme*, pages 141–146. Presses Universitaires France.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 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*, pages 214–228, Sevilla, Spain.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

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

Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. 2019. A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the 3rd Workshop on Abusive Language Online*, pages 94–104, Florence, Italy.

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

Lei Gao, Alexis Kuppersmith, and Ruihong Huang. 2017. Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. In *Proceedings of the 8th International Joint Conference on Natural Language Processing*, pages 774–782, Taipei, Taiwan.

Jennifer Golbeck, Zahra Ashktorab, Rashad O. Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A. Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, Kelly M. Hoffman, Jenny Hottle, Vichita Jienjitlert, Shivika Khare, Ryan Lau, Marianna J. Martindale, Shalmali Naik, Heather L. Nixon, Piyush Ramachandran, Kristine M. Rogers, Lisa Rogers, Meghna Sardana Sarin, Gaurav Shahane, Jayanee Thanki, Priyanka Vengataraman, Zijian Wan, and Derek Michael Wu. 2017. A large labeled corpus for online harassment research. In *Proceedings of the 9th ACM Web Science Conference*, pages 229–233, New York, USA.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 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*, pages 1336–1350, Online.

Mohammed Hasanuzzaman, Gaël Dias, and Andy Way. 2017. Demographic word embeddings for racism detection on Twitter. In *Proceedings of the 8th International Joint Conference on Natural Language Processing*, pages 926–936, Taipei, Taiwan.

Muhammad Okky Ibrohim and Indra Budi. 2018. A dataset and preliminaries study for abusive language detection in Indonesian social media. *Procedia Computer Science*, 135:222–229.

Akshita Jha and Radhika Mamidi. 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*, pages 7–16, Vancouver, Canada.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Joao Augusto Leite, Diego F. 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.

Nikola Ljubešić, Tomaž Erjavec, and Darja Fišer. 2018. Datasets of Slovene and Croatian moderated news comments. In *Proceedings of the 2nd Workshop on Abusive Language Online*, pages 124–131, Brussels, Belgium.

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

Lígia Mesquita. 2018. Denúncias de discurso de ódio online dispararam no 2º turno das eleições. *BBC*, 1.

Leila Maria Oliveira. 2019. *Imigrantes, xenofobia e racismo: uma análise de conflitos em escolas municipais de São Paulo*. Ph.D. thesis, Programa de Pós-Graduação em Educação, Pontifícia Universidade Católica de São Paulo, São Paulo, Brazil.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 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*, pages 4675–4684, Hong Kong, China.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deep learning for user comment moderation. In *Proceedings of the 1st Workshop on Abusive Language Online*, pages 25–35, British Columbia, Canada.

Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in Greek. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5113–5119, Marseille, France.

Robert Post. 2009. Hate speech. In *Extreme Speech and Democracy*, pages 123–138. Oxford Scholarship Online.

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

Anna Schmidt and Michael Wiegand. 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*, pages 1–10, Valencia, Spain.

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

Kenneth Steimel, Daniel Dakota, Yue Chen, and Sandra Kübler. 2019. Investigating multilingual abusive language detection: A cautionary tale. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 1151–1160, Varna, Bulgaria.

Cass R. Sunstein. 2016. Partyism. *University of Chicago Legal Forum*, 2016(2).

Zeerak Waseem and Dirk Hovy. 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*, pages 88–93, California, USA.

Sean J. Westwood, Shanto Iyengar, Stefaan Walgrave, Rafael Leonisio, Luis Miller, and Oliver Strijbis. 2018. The tie that divides: Cross-national evidence of the primacy of partyism. *European Journal of Political Research*, 57:333–354.

W.J. Wilson. 1999. *The Bridge over the Racial Divide: Rising Inequality and Coalition Politics*. University of California Press.

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

Savvas Zannettou, Joel Finkelstein, Barry Bradlyn, and Jeremy Blackburn. 2020. A quantitative approach to understanding online antisemitism. In *Proceedings of the 14th International AAAI Conference on Web and Social Media*, pages 786–797, USA.