# Contextual-Lexicon Approach for Abusive Language Detection

**Francielle Vargas*†, Fabiana Góes*, Isabelle Carvalho‡**
**Fabrício Benevenuto†, Thiago A. S. Pardo***
*Institute of Mathematical and Computer Sciences, University of São Paulo, Brazil
‡Ribeirão Preto Medical School, University of São Paulo, Brazil
†Computer Science Department, Federal University of Minas Gerais, Brazil
{francielleavargas,fabianagoes,isabelle.carvalho}@usp.br
fabricio@dcc.ufmg.br, taspardo@icmc.usp.br

## Abstract

Since a lexicon-based approach is more elegant scientifically, explaining the solution components and being easier to generalize to other applications, this paper provides a new approach for offensive language and hate speech detection on social media, which embodies a lexicon of implicit and explicit offensive and swearing expressions annotated with contextual information. Due to the severity of the social media abusive comments in Brazil, and the lack of research in Portuguese, Brazilian Portuguese is the language used to validate the models. Nevertheless, our method may be applied to any other language. The conducted experiments show the effectiveness of the proposed approach, outperforming the current baseline methods for the Portuguese language.

## 1 Introduction

In Brazil, hate speech is prohibited. Nevertheless, in government and civil society, the regulation of hate speech is not effective due to the difficulty to identify, quantify and classify abusive comments. Indeed, this is rather a difficult requirement to satisfy. According to Mesquita (2018), the *Safernet* non-governmental organization, which operates in cooperation with public organizations in Brazil, as well as companies, such as Google, Facebook, and Twitter, proposed a collection of data on actions that violate human rights. The data is very worrisome: during the 2018 year's election period, denunciations with xenophobia content had an increase of 2,369.5%; apology and public incitement to violence and crimes against life, 630.52%; neo-nazism, 548.4%; homophobia, 350.2%; racism, 218.2%; and religious intolerance, 145.13% [1]. Figure 1 shows the hate crimes evolution that occurred in the most populous Brazilian state [2]. The data was collected from São Paulo public security government. The pink line provides data on religious intolerance crimes, red on homophobia/transphobia, blue on race/ethnicity/color, green on region/origin, yellow on political intolerance, and light green on other crimes.

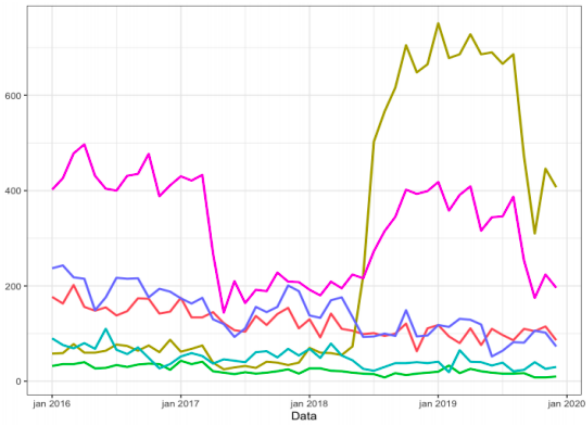


Figure 1: Hate crimes occurrence in São Paulo from 2016 to begining of 2020.

Indeed, it is generally agreed that the high incidence of hate crimes is boosted by the popularization of online social networks. In social media, people and organizations may use the language to defamation, oppression, and terrorism. The language used intentionally in order to disrespect, insult or attack the reader is denominated in literature by abusive language, unless otherwise stated such as offensive language (Çöltekin, 2020; Pitenis et al., 2020; Razavi et al., 2010), hate speech (Schmidt and Wiegand, 2017; Waseem and Hovy, 2016) and cyberbullying (Rosa et al., 2019).

According to Warner and Hirschberg (2012), hate speech is a particular form of abusive lan-

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

[2] https://www.ssp.sp.gov.br/



*Proceedings of Recent Advances in Natural Language Processing*, pages 1438–1447
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_161

guage considering stereotypes to express an ideology of hate. In the same settings, Nockleby (2000) defines hate speech as "any communication that disparages a person or a group based on some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic".

To the best of our knowledge, no previous methods exist in order to embody an offensive lexicon annotated with contextual information to automatically classify abusive language on social media. Therefore, the main contribution of this paper is providing a new method for abusive comment detection on social media. Moreover, as already mentioned, due to the severity of the hate crimes in Brazil, and the lack of research in this language, Brazilian Portuguese is the language used to evaluate the proposed method, which showed high performance, outperforming the current baseline methods for Portuguese. Despite the proposed approach has been orchestrated over Brazilian Portuguese comments, the method in this paper may be applied to any other language. Finally, this paper also presents the evaluation of algorithms used for feature selection.

The remainder of the paper is structured as follows. In Section 2, we briefly introduce the most relevant related work. Section 3 presents a overview of the data. Sections 4 and 5 describe the proposed method and the performed experiments. In Section 6, we report the evaluation results. In Section 7, we make some final remarks.

## 2 Related Work

Several efforts have been made to provide automated detection approaches for hate speech and offensive languages on social media (Gao and Huang, 2017; Davidson et al., 2017; Warner and Hirschberg, 2012). The basic state of the art framework consists of creating lists of words that contain sets of known hate keywords. Furthermore, corpora are manually annotated in order to construct training datasets labeled with hate speech and non-hate speech. At last, automated methods of learning, such as traditional machine learning or neural-based machine learning, are used to automatically detect hate speech in social media texts. However, most hate speech resources and models are proposed for English (Zampieri et al., 2019; Basile et al., 2019; Davidson et al., 2017; Njagi et al., 2015; Ting et al., 2013).

For Portuguese, Fortuna et al. (2019) adopted the definition of hate speech proposed by Fortuna and Nunes (2018), and proposed a new dataset composed of 5,668 tweets, as well as automated methods using a hierarchy of hate to identify social groups of discrimination. The authors have obtained 78% f1-score using a neural network (LSTM). Additionally, de Pelle and Moreira (2017) provide a new dataset composed of 1,250 comments collected from G1 Brazilian online newspaper and annotated with offensive and non-offensive tags. In addition, the authors present classification results achieved by classical machine learning algorithms (SMV and NB), reporting results over 81% f1-score.

## 3 Data Overview

### 3.1 HateBR Corpus

HateBR was proposed by Vargas et al. (2021), and consists of the first large-scale dataset for hate speech and offensive language detection for the Portuguese language. HateBR corpus annotation presents 89% of human inter-annotator agreement. The corpus is composed of 7,000 Instagram comments annotated with three different layers: (i) binary classes (offensive and non-offensive); (ii) offense-levels (highly, moderately, and slightly offensive); and (iii) nine hate group targets (xenophobia, racism, homophobia, sexism, religious intolerance, partyism [3], apology to dictatorship, antisemitism, and fatphobia).

The authors report that the comments were collected from six public Instagram accounts of the Brazilian political domain. Moreover, they selected three liberal-party accounts followed by three conservative-party accounts, being four women and two men. Due to the degree of complexity of the offensive language and hate speech detection tasks, mainly because it involves a highly politicized domain, the authors decided to enroll annotators at higher levels of education (Ph.D.), which are from different political orientations and colors in order to minimize bias.

Tables 1, 2, 3 show the HateBR dataset statistics.

[3] According to the professor at Harvard University, "partyism" is a form of hostility and prejudice that operates across political lines (Sunstein, 2016). Moreover, Westwood et al. (2018) demonstrated that partyism influences behaviors and non-political judgments.

Table 1: Binary class: offensive x non-offensive.

| Binary Class | Total |
|---|---|
| Non-Offensive | 3,500 |
| Offensive | 3,500 |
| Total | 7,000 |

Table 2: Offense levels.

| Offense-levels Classes | Total |
|---|---|
| Slightly Offensive | 1,281 |
| Moderately Offensive | 1,440 |
| Highly Offensive | 779 |
| Total | 3,500 |

Table 3: Hate group targets.

| Hate Groups | Total |
|---|---|
| Partyism | 496 |
| Sexism | 97 |
| Religious Intolerance | 47 |
| Apology to Dictatorship | 32 |
| Fat Phobia | 27 |
| Homophobia | 17 |
| Racism | 8 |
| Antisemitism | 2 |
| Xenophobia | 1 |
| Total | 727 |

### 3.2 MOL - Multilingual Offensive Lexicon

MOL (Multilingual Offensive Lexicon)[4] consists of a multilingual offensive lexicon, composed of 1,000 explicit and implicit offensive and swearing expressions of offense and swearing, which were annotated with a binary class: context-dependent and context-independent offensive. For example, the term *''vadia"* ("slut") consists of a context-independent offensive term. On 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".

The MOL was extracted from HateBR corpus (Vargas et al., 2021), and each term or expression was annotated by three different annotators obtaining a high human agreement score (73% Kappa). Furthermore, as already mentioned, implicit content also was extracted using "clue terms or expressions". For example, the expression *voltar para a jaula* ("go back to the cage") consists of a "clue expression" to identify the implicit offensive term *ladrão* ("thief"). Finally, terms that showed explicit potential to indicate hate speech targets were also annotated, for instance, *vadia* ("slut") and *judeus dos infernos* ("jews from hell"). Note that the occurrence of these cases may indicate sexist and antisemitism comments.

[4] https://github.com/francielleavargas/MOL

## 4 The Proposed Approach

We present a new approach to detect abusive comments on the web and social media. Our method embodies an offensive lexicon, which provides contextual information on hate speech and offenses. We show in detail our approach in Sections 4.1, 4.2, and 4.3.

### 4.1 Tasks

In this paper, we assume that abusive language detection may be divided into two main tasks: (i) offensive language detection, (ii) hate speech detection. Considering this premise, we train two different classifiers. The first classifier automatically identifies offensive comments. On the other hand, the second classifier automatically identifies comments that present hate speech content. Note that a hate speech comment is always an offensive comment, however, an offensive comment may present or not hate speech content. Figure 2 shows each of these different tasks in detail.

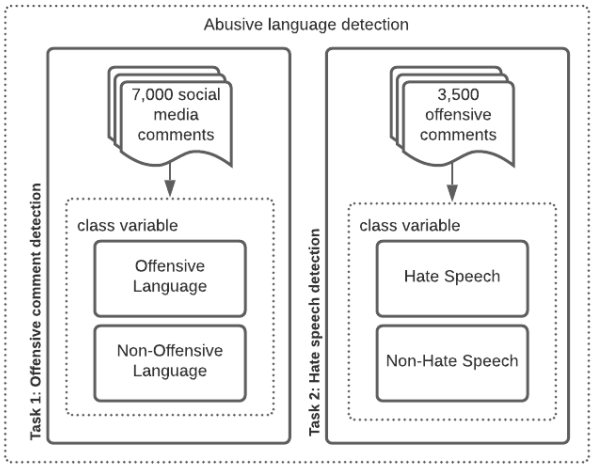


Figure 2: Our approach to use the HateBR dataset and automatically detect offensive comments, as well as offensive comments with hate speech content.

### 4.2 The Feature Set

Defining the most appropriate textual representation is a crucial task that directly influences the performance of the predictive model built by the classification algorithms. In this paper, we modeled hate speech and offensive language through

different representation paradigms and features. We describe each one in what follows.

#### 4.2.1 Lexical and Morphosyntactic Features

We selected lexical elements (each word into the document without stopwords), as well as part-of-speech-based features, using the Stanford Stanza POS tagger[5] for Portuguese.

#### 4.2.2 Lexicon-Based Features

We included features from three different lexicons: one sentiment lexicon (Sentilex-PT (Silva et al., 2012)), one emotion lexicon (WordNetAffect.BR (Pasqualotti, 2008)), and finally, one offensive contextual lexicon (MOL).

1. **Sentiment Lexicon**: we evaluated features based on sentiment (Silva et al., 2012) and emotion (Pasqualotti, 2008) lexicons, which present semantic polarity (e.g., positive, negative and neutral) and emotion types (e.g., anger, love, hate, disgust, suspicious and fear).

2. **Contextual Lexicon**: we evaluated features based on an offensive lexicon (MOL) annotated with contextual (context-dependent and context-independent) labels.

#### 4.2.3 Word Embedding Features

We also evaluated word embedding-based features. Different from other language models, BERT (Bidirectional Encoder Representations from Transformers) is usually used to pre-train deep bidirectional representations from unlabeled texts by jointly conditioning on both left and right contexts in all internal network layers (Devlin et al., 2019). In a similar setting, we also used fastText, the Facebook pre-trained models (Joulin et al., 2016).

#### 4.2.4 Feature Set Overview

We summarize in Table 4 the five feature representations used in this paper.

1. POS+S: we extracted the occurrence of part-of-speech tags for each comment. In addition, we extracted the occurrence of positive and negative words for each comment using the sentiment (Silva et al., 2012) and emotion (Pasqualotti, 2008) lexicons.

[5] https://stanfordnlp.github.io/stanza/pos.html

Table 4: Feature set representations.

| N. | Features | Description |
|---|---|---|
| 1 | POS+S | Bag-of-POS+Sentiment |
| 2 | BOW | Bag-Of-Words |
| 3 | MOL | Bag-Of-MOL |
| 4 | B+M | Bag-Of-Words embodying the MOL |
| 5 | BERT & fastText | Multilingual pre-trained models |

2. BOW: we created a bag-of-words representation or, in other words, we generated a text representation that describes the occurrence of dataset vocabulary for each comment. We simply calculate how many times each word of our dataset vocabulary (features) appears in each comment.

3. MOL: in this representation, a bag-of-words was generated using the terms or expressions extracted from the offensive lexicon (MOL), which were used as features. Therefore, for each comment, the occurrence of the MOL's terms was counted. Additionally, context labels (context-independent and context-dependent) have been considered in order to compute different weights to context-independent and context-dependent features. The frequency of the terms with context-independent labels were multiplied by 2, while the frequency of the terms with context-dependent features remained the same. Specifically for hate speech detection task, we also checked if a term presented any markers that identify hate speech content, and, if this condition was true, an additional weight was accounted. Therefore, in the MOL representation, the value of a term $x$ in the document (comment) $y$ for the offensive comment detection (task 1) is defined according to

$$MOL_{x,y} = freq_{x,y} * weightC_x \quad (1)$$

and for the hate speech detection (task 2) is given by

$$MOL_{x,y} = freq_{x,y} * weightC_x * weightH_x \quad (2)$$

where $freq$ is the frequency of the term in the document, $weightH = 2$ when the term is a marker that identifies hate speech and $weightH = 1$ otherwise, $weightC = 1$ for

context-dependent terms and $weightC = 2$ when the term is context-independent.

4. B+M: we generated a bag-of-words representation, which embodies context label information from the offensive lexicon (MOL). In other words, we firstly generated a bag-of-words from all comments into the dataset. Then, we performed the match with terms into MOL, and then we assigned a weight for terms or expressions labeled with context-dependent (weaker weight) and context-independent (stronger weight). The contextual labels are provided by MOL. Therefore, in B+M representation, the value of a term $x$ into the document (comment) $y$ is defined according to

$$B + M_{x,y} = freq_{x,y} * weightC_x \quad (3)$$

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

5. In a different setting, the feature extraction for the BERT and fastText followed state of the art text classification with a maximum size of 500. For the fastText classifier, we set the maximum size equal to 64 and the maximum number of features equal to 10,000. We used the standard processor model and evaluated the n-gram range for unigram, bigram, and trigram.

### 4.3 The Learning Methods

In general, previous works on hate speech detection use neural networks or traditional machine learning techniques on specific communities (Davidson et al., 2017; Founta et al., 2019; Del Vigna et al., 2017; Njagi et al., 2015; Djuric et al., 2015). In order to evaluate the performance of neural networks and traditional machine learning techniques, we used the following learning methods: Support Vector Machine (SVM) (Scholkopf and Smola, 2001) with linear kernel; Multinomial Naive Bayes (NB) (McCallum et al., 1998; Eyheramendy et al., 2003); Multilayer Perceptron (MLP) (Haykin, 2009) with one hidden layer (with 100 neurons), and ReLU activation function; Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) with two hidden layers (with 200 and 50 neurons, respectively) and a softmax output unit for the binary classification. ReLU was used as the activation function, as well as number of epochs equal to 10, and a random batch size of 100 documents. Moreover, we also used pre-trained models of word embeddings, such as BERT (Devlin et al., 2019) and fastText (Joulin et al., 2016).

## 5 Experiments

We carried out a wide variety of experiments. We describe the entire process in Sections 5.1, 5.2, 5.3 and 5.4.

### 5.1 Data Preparing

We accomplished an approach for data preparation, as shown in Figure 3

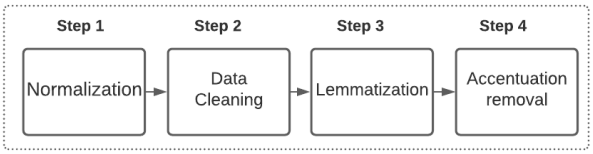


Figure 3: Data preparation.

Firstly, we normalized our dataset using the normalization tool for Brazilian Portuguese proposed by Bertaglia and Nunes (2016). The normalization process consists of identifying noise, which is very common in User-Generated Content (UGC), such as orthographic errors, often phonetically-motivated, abbreviations and expressions often used informally by web users, proper names and acronyms wrongly or not at all capitalized, agglutinated words that should be split, and wrong use of sentence delimiters; and suggesting possible substitutions.

Moving forward, in the second step, we remove emoticons, special characters, accounts, hyperlinks, and websites. In step 3, we lemmatize our dataset using Spacy[6]. Finally, in step 4, accentuation is removed.

### 5.2 Feature Selection

Feature selection (FS) allows the removal of irrelevant and redundant features. In this paper, in order to select the best feature set, we applied the following FS algorithms: (i) Correlation-based Feature Selection (CFS) (Hall, 1998), which selects characteristics that are highly correlated with

[6]https://spacy.io/models/pt

the class and not correlated with each other using Pearson coefficient[7] as criteria, and (ii) Information Gain Analysis (InfoGain) (Witten et al., 2016), which quantifies and chooses the characteristics that have the maximum information gain concerning the class. We apply both FS techniques on the NB, SVM, MLP and LSTM models. For BERT and fastText features, we do not apply FS techniques. Finally, we evaluated the performance of the FS techniques for each feature representation. More specifically, we measure the potential of the algorithms to help in the gain and loss of accuracy, precision, recall, and f1-score. Results are shown in Table 7.

### 5.3 Class Balancing

The most common class balancing methods are oversampling (Chawla et al., 2002) and undersampling (Witten et al., 2016). In undersampling, the number of examples of each class is maintained based on the number of examples from the minority class. Differently, in oversampling, the approach involves the construction of examples for the minority class, although these examples may not add any new information to the model. In our experiments, we adopted the undersampling on the unbalanced classes of hate speech, specially due to the fact that this approach makes overfitting unlikely. Note that, in our dataset (the HateBR), there are 727 labeled hate speech samples versus 2,227 labeled non-hate speech samples. As a result of the undersampling approach, we obtained 727 labeled samples for hate speech and 727 samples for non-hate speech.

### 5.4 Evaluation

Our models were trained and tested using 10-fold cross-validation (Stone, 1974). We have computed the classical machine learning evaluation measures of Precision, Recall and F1-Score. We present these evaluative measures for each class involved, as well as simple arithmetic means. The results are shown in Table 5. Moreover, we evaluated BERT and fastText pre-trained models, and show the obtained results in Table 6.

We also present the evaluation of the methods with feature selection (FS). We measure the gain and loss of precision, recall, and f1-score for each selected algorithm (CFS and InfoGain) in both tasks: offensive language detection and hate speech detection, as well as for each representations: POS+S, BOW, MOL and B+M. Table 7 shows the results. We should point out that T1 is the sum of each representation, and T2 is the sum for each FS algorithm.

Finally, Table 8 shows the comparison of the results with the current baseline methods for Portuguese.

## 6 Results

As shown in Table 5, the B+M proposed method in this paper obtained better results of precision, recall, and f1-score in both tasks - offensive language and hate speech detection. The worse results were obtained using the POS+S approach, which combines part-of-speech and sentiment lexicon features. We should point out the considerable impact of an offensive lexicon for abusive language detection, when compared to the impact of a sentiment lexicon. Our results showed that sentiment lexicon approach present weak performance for abusive language detection on the web and social media.

Moving forward, the conducted experiments also show that the traditional machine learning techniques presented better performance than neural-based classifiers for offensive language tasks. Nevertheless, for the hate speech detection task, the neural-based classifier overcame the traditional machine learning methods.

In general, BERT and fastText, as shown in Table 6, presented a high performance for both tasks (offensive language and hate speech detection), even though our approach (B+M) has overcome the fastText (trigrams) in 2% (f1-score) for hate speech detection, as well as presented better precision performance, and the same recall and f1-score performances for offensive language detection.

Considering the feature selection (FS) performance, as shown in Table 7, the InfoGain algorithm produced better results for precision, recall, and f1-score than CFS algorithm for offensive language detection (task 1). On other hand, for hate speech detection (task 2), CFS algorithm obtained better performance than InfoGain in recall and f1-score. Moreover, for offensive language detection (task 1), InfoGain applied on BOW and B+M representations obtained performance gain, and POS+S and MOL presented loss of perfor-

[7]Pearson's correlation coefficient is a linear correlation coefficient that returns a value between -1 and +1. A -1 means there is a strong negative relationship, and +1 means there is a strong positive relationship.

Table 5: NB, SVM, MLP and LSTM Evaluation.

| Tasks | Features set | Class | Precision | | | | Recall | | | | F1-Score | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | NB | SVM | MLP | LSTM | NB | SVM | MLP | LSTM | NB | SVM | MLP | LSTM |
| Task 1: Offensive Language Detection | POS+S | 0 | 0.50 | 0.51 | 0.47 | 0.49 | 0.41 | 0.39 | 0.51 | 0.37 | 0.45 | 0.44 | 0.49 | 0.42 |
| | | 1 | 0.50 | 0.51 | 0.54 | 0.49 | 0.50 | 0.64 | 0.51 | 0.62 | 0.59 | 0.57 | 0.52 | 0.55 |
| | | Avg | 0.50 | 0.51 | 0.51 | 0.49 | 0.50 | 0.51 | 0.51 | 0.49 | 0.50 | 0.50 | 0.51 | 0.49 |
| | BOW | 0 | 0.85 | 0.82 | 0.92 | 0.83 | 0.86 | 0.96 | 0.81 | 0.89 | 0.86 | 0.88 | 0.81 | 0.86 |
| | | 1 | 0.86 | 0.95 | 0.79 | 0.88 | 0.85 | 0.79 | 0.90 | 0.81 | 0.85 | 0.86 | 0.90 | 0.85 |
| | | Avg | 0.85 | 0.88 | 0.86 | 0.85 | 0.85 | 0.87 | 0.86 | 0.85 | 0.85 | 0.87 | 0.84 | 0.85 |
| | MOL | 0 | 0.74 | 0.78 | 0.94 | 0.79 | 0.97 | 0.96 | 0.77 | 0.94 | 0.84 | 0.86 | 0.85 | 0.86 |
| | | 1 | 0.95 | 0.94 | 0.72 | 0.93 | 0.66 | 0.73 | 0.93 | 0.75 | 0.78 | 0.82 | 0.81 | 0.83 |
| | | Avg | 0.85 | 0.86 | 0.83 | 0.86 | 0.81 | 0.84 | 0.85 | 0.84 | 0.81 | 0.84 | 0.81 | 0.84 |
| | B+M | 0 | 0.84 | 0.84 | 0.91 | 0.86 | 0.93 | 0.94 | 0.83 | 0.85 | 0.88 | 0.88 | 0.87 | 0.85 |
| | | 1 | 0.93 | 0.93 | 0.81 | 0.85 | 0.83 | 0.81 | 0.90 | 0.86 | 0.88 | 0.87 | 0.86 | 0.85 |
| | | Avg | **0.89** | 0.88 | 0.86 | 0.85 | **0.88** | **0.88** | 0.87 | 0.85 | **0.88** | 0.86 | 0.86 | 0.85 |
| Task 2: Hate Speech Detection | POS+S | 0 | 0.52 | 0.49 | 0.42 | 0.52 | 0.48 | 0.78 | 0.53 | 0.47 | 0.50 | 0.60 | 0.47 | 0.50 |
| | | 1 | 0.52 | 0.47 | 0.63 | 0.52 | 0.56 | 0.20 | 0.52 | 0.57 | 0.54 | 0.28 | 0.57 | 0.54 |
| | | Avg | 0.52 | 0.48 | 0.53 | 0.52 | 0.52 | 0.49 | 0.53 | 0.52 | 0.52 | 0.44 | 0.52 | 0.52 |
| | BOW | 0 | 0.62 | 0.84 | 0.43 | 0.85 | 0.82 | 0.42 | 0.82 | 0.37 | 0.70 | 0.55 | 0.57 | 0.54 |
| | | 1 | 0.73 | 0.61 | 0.91 | 0.61 | 0.49 | 0.92 | 0.61 | 0.93 | 0.59 | 0.73 | 0.73 | 0.73 |
| | | Avg | 0.68 | 0.72 | 0.67 | 0.73 | 0.66 | 0.67 | 0.72 | 0.66 | 0.65 | 0.64 | 0.65 | 0.64 |
| | MOL | 0 | 0.61 | 0.62 | 0.58 | 0.60 | 0.74 | 0.80 | 0.68 | 0.93 | 0.67 | 0.69 | 0.63 | 0.73 |
| | | 1 | 0.67 | 0.71 | 0.73 | 0.84 | 0.53 | 0.50 | 0.63 | 0.38 | 0.59 | 0.59 | 0.68 | 0.52 |
| | | Avg | 0.64 | 0.66 | 0.66 | 0.72 | 0.64 | 0.65 | 0.66 | 0.65 | 0.63 | 0.64 | 0.66 | 0.63 |
| | B+M | 0 | 0.79 | 0.77 | 0.93 | 0.71 | 0.78 | 0.93 | 0.79 | 0.89 | 0.78 | 0.84 | 0.86 | 0.79 |
| | | 1 | 0.78 | 0.92 | 0.76 | 0.85 | 0.79 | 0.72 | 0.92 | 0.64 | 0.79 | 0.80 | 0.83 | 0.73 |
| | | Avg | 0.78 | 0.84 | **0.85** | 0.78 | 0.78 | 0.83 | **0.86** | 0.77 | 0.78 | 0.82 | **0.85** | 0.76 |

Table 6: BERT and fastText Evaluation.

| Models | Class | Task 1: Offensive Language Detection | | | Task 2: Hate Speech Detection | | |
|---|---|---|---|---|---|---|---|
| | | Precision | Recall | F1-Score | Precision | Recall | F1-Score |
| BERT | 0 | 0.85 | 0.86 | 0.86 | 0.76 | 0.65 | 0.70 |
| | 1 | 0.85 | 0.85 | 0.85 | 0.64 | 0.75 | 0.69 |
| | Avg | 0.86 | 0.86 | 0.86 | 0.70 | 0.70 | 0.70 |
| fastText (unigram) | 0 | 0.88 | 0.88 | 0.88 | 0.78 | 0.76 | 0.77 |
| | 1 | 0.87 | 0.87 | 0.87 | 0.76 | 0.79 | 0.77 |
| | Avg | **0.88** | **0.88** | **0.88** | 0.77 | 0.79 | 0.77 |
| fastText (bigrams) | 0 | 0.83 | 0.87 | 0.85 | 0.77 | 0.84 | 0.80 |
| | 1 | 0.87 | 0.84 | 0.85 | 0.80 | 0.72 | 0.76 |
| | Avg | 0.85 | 0.85 | 0.85 | 0.78 | 0.78 | 0.78 |
| fastText (trigrams) | 0 | 0.83 | 0.91 | 0.87 | 0.77 | 0.97 | 0.86 |
| | 1 | 0.90 | 0.81 | 0.85 | 0.96 | 0.70 | 0.81 |
| | Avg | 0.86 | 0.86 | 0.86 | **0.86** | **0.84** | **0.83** |

mance. For hate speech detection (task 2), InfoGain applied on B+M representation presented performance gain. Differently from this, POS+S, BOW, and MOL had a loss of performance using InfoGain. Differently, CFS algorithm applied to BOW and B+M obtained performance gain, and when applied to POS+S and MOL representations, presented loss of performance.

Table 7: Feature selection performance.

| Measures | FS | Features set | Task 1: Offensive Language Detection | | | | | | Task 2: Hate Speech Detection | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Learning Methods | | | | T1 | T2 | Learning Methods | | | | T1 | T2 |
| | | | NB | SVM | MLP | LSTM | | | NB | SVM | MLP | LSTM | | |
| Precision | CFS | POS+S | -0.25 | -0.01 | -0.02 | -0.01 | -0.29 | -0.49 | -0.04 | 0.02 | 0.03 | -0.04 | -0.03 | 1.04 |
| | | BOW | -0.01 | -0.02 | 0.00 | -0.02 | -0.05 | | 0.10 | 0.20 | 0.17 | 0.16 | 0.63 | |
| | | MOL | -0.04 | -0.02 | -0.04 | -0.04 | -0.14 | | 0.00 | 0.05 | 0.01 | 0.10 | 0.16 | |
| | | B+M | -0.03 | -0.03 | 0.00 | 0.05 | -0.01 | | 0.06 | 0.05 | 0.08 | 0.09 | 0.28 | |
| | Info Gain | POS+S | -0.25 | 0.01 | -0.02 | -0.01 | -0.27 | -0.01 | -0.02 | 0.01 | 0.04 | -0.01 | 0.02 | 1.14 |
| | | BOW | 0.00 | 0.02 | 0.05 | 0.07 | 0.14 | | 0.09 | 0.22 | 0.16 | 0.17 | 0.64 | |
| | | MOL | -0.02 | 0.00 | 0.00 | 0.02 | 0.00 | | -0.01 | 0.06 | 0.02 | 0.11 | 0.18 | |
| | | B+M | 0.00 | 0.01 | 0.04 | 0.07 | 0.12 | | 0.07 | 0.06 | 0.10 | 0.07 | 0.30 | |
| Recall | CFS | POS+S | 0.00 | 0.00 | 0.02 | 0.01 | 0.03 | -0.37 | -0.04 | 0.01 | 0.02 | -0.04 | -0.05 | 0.80 |
| | | BOW | -0.04 | -0.05 | 0.02 | -0.02 | -0.09 | | 0.08 | 0.08 | 0.21 | 0.10 | 0.47 | |
| | | MOL | -0.12 | -0.05 | -0.01 | -0.10 | -0.28 | | 0.00 | 0.01 | 0.07 | 0.02 | 0.10 | |
| | | B+M | -0.05 | -0.05 | 0.02 | 0.05 | -0.03 | | 0.06 | 0.04 | 0.10 | 0.08 | 0.28 | |
| | Info Gain | POS+S | 0.00 | 0.00 | -0.02 | -0.01 | -0.03 | 0.16 | -0.02 | 0.01 | 0.04 | -0.01 | 0.02 | 0.36 |
| | | BOW | 0.00 | 0.02 | 0.05 | 0.07 | 0.14 | | -0.02 | -0,01 | 0.04 | -0.02 | -0.01 | |
| | | MOL | -0.04 | -0.01 | 0.01 | -0.04 | -0.08 | | -0.02 | 0.02 | 0.08 | 0.04 | 0.12 | |
| | | B+M | 0.00 | 0.01 | 0.04 | 0.08 | 0.13 | | 0.07 | 0.05 | 0.11 | 0.00 | 0.23 | |
| F1-Score | CFS | POS+S | -0.17 | -0.03 | 0.04 | 0.02 | -0.14 | -0.74 | -0.04 | -0.06 | 0.03 | -0.04 | -0.11 | **0.48** |
| | | BOW | -0.04 | -0.05 | -0.07 | -0.02 | -0.13 | | 0.07 | 0.12 | 0.15 | 0.07 | 0.41 | |
| | | MOL | -0.14 | -0.06 | -0.02 | -0.12 | -0.39 | | -0.01 | -0.01 | -0.02 | -0.03 | -0.07 | |
| | | B+M | -0,06 | -0.05 | -0.02 | 0.05 | -0.08 | | 0.06 | 0.04 | 0.08 | 0.07 | 0.25 | |
| | Info Gain | POS+S | -0.17 | 0.11 | -0.01 | 0.03 | -0.04 | **0.12** | -0.07 | -0.02 | 0.03 | -0.01 | -0.07 | 0.03 |
| | | BOW | 0.00 | 0.02 | -0.01 | 0.07 | 0.14 | | 0.07 | -0.02 | -0.01 | -0.04 | -0.09 | |
| | | MOL | -0.05 | -0.01 | 0.05 | -0.04 | -0.11 | | -0.01 | 0.00 | -0.01 | 0.02 | -0.01 | |
| | | B+M | 0.00 | 0.01 | 0.04 | 0.08 | 0.13 | | 0.06 | 0.05 | 0.10 | -0.02 | 0.20 | |

### 6.1 Comparing Results

Table 8 shows a comparison of results between our new proposed method and baseline methods for Portuguese. Although a direct comparison is unfair (as the authors use different datasets), it offers an idea of the general performance of the methods.

de Pelle and Moreira (2017) report a f1-score of 81% using SVM and NB algorithms. For the same algorithms, our approach presented 88% of f1-score, improving the performance. In the same settings, Fortuna et al. (2019) report a f1-score of 78% using the LSTM algorithm. In our experiments, we obtained an f1-score of 86%, also using the LSTM algorithm, consequently, our approach presented better performance.

Table 8: Comparison of results.

| | **Dataset language** | **Algorithms** | **F1-score** |
|---|---|---|---|
| Our approach | Brazilian Portuguese | SVM and NB | **88%** |
| de Pelle and Moreira (2017) | Brazilian Portuguese | SVM and NB | 81% |
| Our approach | Brazilian Portuguese | LSTM | **86%** |
| Fortuna et al. (2019) | European and Brazilian Portuguese | LSTM | 78% |

## 7 Conclusions

In this work, we provide a new approach for the automatic detection of abusive comments on social media. Our approach embodies an offensive lexicon that provides contextual information. Due to the increase of abusive comments on social media in Brazil, as well as the lack of research in Portuguese, we decided to use an Brazilian annotated dataset to evaluate the models. The proposed approach obtains high performances: 88% f1-score for offensive comments detection, and 85% for comments with hate speech, which overcame the current baseline methods for Portuguese. We also evaluated the performance of feature selection (FS) methods, and conclude that InfoGain algorithm is the best algorithm for the offensive comment detection task, considering the obtained gains in recall and f1-score. For the hate speech detection task, CFS algorithm obtained better performance. Accordingly, based on the obtained results, we concluded that the proposed approach in this paper for automated detection of abusive comments is efficient and highly relevant, bearing in mind the current Brazilian social scenario, in which hateful comments are a very relevant social problem. Moreover, in the next year (2022), there will be presidential elections in Brazil, and this paper may provide a reliable automated approach for abusive comments detection in order to minimize political polarization, as well as hate crimes on social media.

## Acknowledgements

The authors are grateful to DCC-UFMG research project: WHATSAPP MONITOR - SIMP 18048 for supporting this work.

## References


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.

Thales Felipe Costa Bertaglia and Maria das Graças Volpe Nunes. 2016. Exploring word embeddings for unsupervised textual user-generated content normalization. In *Proceedings of the 2nd Workshop on Noisy User-generated Text*, pages 112–120, Osaka, Japan.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

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, Montreal, Canada.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the 1st Italian Conference on Cybersecurity*, pages 86–95, Venice, Italy.

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

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web*, page 29–30, Florence, Italy.

Susana Eyheramendy, David D Lewis, and David Madigan. 2003. On the naive bayes model for text categorization. In *Proceedings of the 9th International Workshop on Artificial Intelligence and Statistics*, pages 93–100, Florida, USA.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computer Survey*, 51(4).

Paula Fortuna, Joao Rocha da Silva, Leo Wanner, Sérgio Nunes, et al. 2019. A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the 3rd Workshop on Abusive Language Online*, pages 94–104, Florence, Italy.

Antigoni Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2019. A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM Conference on Web Science*, page 105–114, Massachusetts, USA.

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.

Mark A Hall. 1998. *Correlation-based feature subset selection for machine learning*. Ph.D. thesis, Thesis submitted in partial fulfillment of the requirements of the degree of Doctor of Philosophy at the University of Waikato.

S. Haykin. 2009. *Neural networks and learning machines*, 3 edition. Pearson Upper Saddle River.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

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

Andrew McCallum, Kamal Nigam, et al. 1998. A comparison of event models for naive bayes text classification. In *Proceedings of the AAAI-98 Workshop on Learning for Text Categorization*, pages 41–48, Madison, Wisconsin.

Lígia Mesquita. 2018. Denúncias de discurso de ódio online dispararam no 2º turno das eleições, diz ong. *BBC News Brasil*, 9(11):1–7.

Dennis Njagi, Z. Zuping, Damien Hanyurwimfura, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10:215–230.

John T. Nockleby. 2000. Hate speech. In *Encyclopedia of the American Constitution*, page 1277–1279, New York, USA. Macmillan.

P. R Pasqualotti. 2008. Reconhecimento de expressões de emoções na interação mediada por computador. Master's thesis, Dissertação de Mestrado em Ciência da Computação. Pontifícia Universidade Católica do Rio Grande do Sul - PUCRS, Porto Alegre, Brasil.

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 510–5019, São Paulo, Brasil.

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.

Amir H. Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive language detection using multi-level classification. In *Proceedings of the 23rd Canadian Conference on Advances in Artificial Intelligence*, page 16–27, Berlin, Heidelberg. Springer-Verlag.

Hugo Rosa, Nádia Pereira, Ricardo Ribeiro, Paula Costa Ferreira, João Paulo Carvalho, Sofia Oliveira, Luísa Coheur, Paula Paulino, AM Veiga Simão, and Isabel Trancoso. 2019. Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, 93:333–345.

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.

Bernhard Scholkopf and Alexander J Smola. 2001. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, Cambridge.

Mário J. Silva, Paula Carvalho, and Luís Sarmento. 2012. Building a sentiment lexicon for social judgement mining. In *Proceedings of the 10th International Conference on Computational Processing of the Portuguese Language*, pages 218–228, Berlin, Heidelberg.

Mervyn Stone. 1974. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133.

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

I-Hsien Ting, Hsing-Miao Chi, Jyun-Sing Wu, and Shyue-Liang Wang. 2013. An approach for hate groups detection in facebook. In *Proceedings of the 3rd International Workshop on Intelligent Data Analysis and Management*, pages 101–106.

Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Fabrício Benevenuto, and Thiago Pardo. 2021. Building an expert annotated corpus of brazilian instagram comments for hate speech and offensive language detection. *CoRR*, abs/2103.14972:01–11.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the 2nd Workshop on Language in Social Media*, pages 19–26, Montréal, Canada.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California.

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.

Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. 2016. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

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*, page 1415–1420, Minnesota, USA.