Multilingual offensive lexicon annotated with contextual information

Anonymous ACL submission

Abstract

Online hate speech and offensive comments detection is not a trivial research problem since pragmatic (contextual) factors influence what is considered offensive. Moreover, offensive terms are hardly found in classical lexical resources such as wordnets, sentiment, and emotion lexicons. In this paper, we embrace the challenges and opportunities of the area and introduce the first multilingual offensive lexicon (MOL), which is composed of 1,000 explicit and implicit pejorative terms and expressions annotated with contextual information. The terms and expressions were manually extracted by a specialist from Instagram abusive comments originally written in Portuguese and manually translated by American English, Latin American Spanish, African French, and German native speakers. Each expression was annotated by three different annotators, producing high human inter-annotator agreement. Accordingly, this resource provides a new perspective to explore abusive language detection.

1 Introduction

According to linguistic studies, the pejorative connotation is used to express emotions, especially hate, anger, and frustration, as well as is heavily influenced by pragmatic (contextual) factors (Jurafsky, 1996; Rae, 2012; Anderson and Lepore, 2013; Frigerio and Tenchini, 2014; Bou; Kádár et al., 2019). 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; Kashyap, 2011; Culpeper et al., 2017). Moreover, several communities forbade hateful terms and expressions considering them as hate crimes [1] (Müller and Schwarz, 2020; Methven-Wasow, 2017; Brugger, 2007).

Hate speech and offensive language detection strategies consist of providing systems capable of recognizing and deleting offensive content without human moderation (Zampieri et al., 2019; Fortuna et al., 2019; Çöltekin, 2020; Pitenis et al., 2020; Fortuna and Nunes, 2018; Schmidt and Wiegand, 2017). According to Steimel et al. (2019), the differences between abusive comments on social media in different languages are not related to an effect of a topic. Furthermore, we would generally agree that in abusive comments in different languages there are terms that present only pejorative connotations, such as swear words, as well as there are terms that present both pejorative and non-pejorative connotations. As an example, observe the two abusive Instagram comments presented in Figure 1.

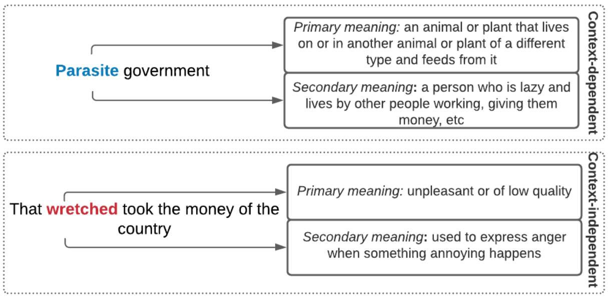
Parasite government
Primary meaning: an animal or plant that lives on or in another animal or plant of a different type and feeds from it
Secondary meaning: a person who is lazy and lives by other people working, giving them money, etc
Context-dependent
That wretched took the money of the country
Primary meaning: unpleasant or of low quality
Secondary meaning: used to express anger when something annoying happens
Context-independent

Figure 1: Examples of Instagram abusive comments with context-dependent and context-independent pejorative terms.

As shown in Figure 1, in the first abusive comment, the term "parasite"[2] presents a primary dictionary meaning without a pejorative context of use. However, the secondary dictionary meaning shows a pejorative context of use. On the other hand, in the second abusive comment, the term "wretched"[3] presents both primary and secondary

[1] https://www.fbi.gov/investigate/civil-rights/hate-crimes

[2] https://dictionary.cambridge.org/pt/dicionario/ingles/parasite

[3] https://dictionary.cambridge.org/pt/dicionario/ingles/wretched

dictionary meanings with the pejorative context of use, without any non-pejorative context of use.

Accordingly, in this paper, we explore a new perspective for hate speech and offensive language classification. Firstly, a linguist and hate speech skilled manually extracted 1,000 explicit and implicit pejorative terms and expressions from a 7,000 document-level Brazilian abusive comments dataset. Secondly, three different annotators classified the identified explicit and implicit terms and expressions according to two classes: context-dependent offensive and context-independent offensive. For classification of contextual information the annotators consult the dictionary meaning and assume that, whether a term or expression presents any non-pejorative context of use, it should be considered a context-dependent offensive term. Differently, whether a term or expression presents only pejorative meanings, or in other words, there is the only pejorative context of use, it should be considered a context-independent offensive term. Lastly, each term originally written in Portuguese was manually translated into American English, Latin American Spanish, African French, and German languages by native speakers. Therefore, this paper introduces the first multilingual offensive lexicon, manually annotated with contextual information and manually translated into four languages.

In what follows, in Section 2, we present an overview of the HateBR dataset, which we used to extract the pejorative terms and expressions. Section 3 describes the construction of the proposed multilingual offensive lexicon (MOL). Section 4 presents an overview the MOL statistics, as well as in Section 5 final remarks are presented.

2 The HateBR Dataset

MOL was extracted from the HateBR dataset (OMITTED DUE TO DOUBLE-BLIND). The HateBR is the first large-scale annotated dataset for Brazilian abusive language detection, composed of 7,000 Brazilian Instagram comments with three different layers of annotation: (i) offensive binary class (offensive or non-offensive); (ii) offense-level classes (highly, moderately, and slightly offensive); and (iii) hate speech binary class (hate speech or non-hate speech). Furthermore, the authors report annotation of nine hate speech phenomena: xenophobia, racism, homophobia, sexism, religious intolerance, partyism, apology to the dictatorship, antisemitism, and fatphobia.

3 Multilingual Offensive Lexicon

3.1 Conceptualization

Post (2013) argues that the abusive language detection tasks present a conceptual difficulty of distinguishing hateful and offensive expressions from expressions that merely denote dislike or disagreement. Accordingly, in order to identify accurately the offensive vocabulary, we defined the following concepts, which has been used identify manually the terms:

1. *Pejorative word* consists of a pejorative term or expression that intends to undermine or disparage any of the following social aspects: moral, appearance, physical, psychological, sexual behavior and orientation, intellectual, economic, religious, and political aspects.

2. *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 interlocutor.

Table 1 shows examples of offensive and swear words and expressions.

Table 1: Pejorative and swear words.

Type	Term/Expression	Translation
Swear word	Vai Tomar no Cú	Go Fuck Yourself
Swear word	Filho da Puta	Son of a bitch
Swear word	Desgraçado	Wretched
Pejorative word	Hipócrita	Hypocritical
Pejorative word	Parasita	Parasite
Pejorative word	Mentiroso	Liar

3.2 Terms Identification

In our approach, terms and expressions were manually identified by a linguist, which is hate speech skilled. This process was performed in two steps. Firstly, for each one of 3.500 offensive comments from the HateBR corpus (see Section 2), the specialist extracted explicit and implicit terms and expressions that presented any pejorative context of use. Secondly, the specialist classified the identified terms according to the conceptualization adopted of offensive words and swear words (see definition in Section 3.1.

Table 2 shows examples of explicit and implicit terms and expressions identified by the specialist for each abusive Instagram comment from HateBR. We should point out that the underline terms indicate clues to identify implicit pejorative, and the boldface terms indicate the explicit pejorative terms.

Table 2: Explicit and Implicit Terms Identification.

N.	Instagram Comments	Explicit Terms	Implicit Terms
1	Tem que jogar esse **lixo** de volta para a cadeia	lixo (crap)	“de volta para a cadeia” (criminal)
2	Esse é o pastor que mais gosta do dinheiro alheio. Um **crápula**.	crápula (crook)	“gosta do dinheiro alheio” (thief)
4	Esse **hipócrita** precisa ir ver o sol nascer quadrado	hipócrita (hypocritical)	“sol nascer quadrado” (criminal)

As shown in Table 2, for each Instagram comment from the HateBR corpus, the specialist identified pejorative explicit and implicit terms or expressions. In the first example, *Tem que jogar esse lixo de vola para a cadeia*(“You have to throw this trash back to jail”), the specialist identified the explicit pejorative term “crap”, and the clue “back to jail”, which refers to the implicit pejorative term “criminal”. In the same settings, in the second example, *Esse é o pastor que mais gosta do dinheiro alheio. Um crápula* (“This is the pastor who most likes other people’s money. A crook.”), the specialist identified the explicit pejorative term “crook”, and the clue “like other people’s money”, which refers to the implicit pejorative term “thief”. At last, *Esse hipócrita precisa ir ver o sol nascer quadrado* (“This hypocrite needs to go see the sunrise square”), the specialist identified the explicit pejorative term “hypocritical”, and the clue “none are any good”, which refers to the implicit pejorative term “immoral”.

3.3 Hate Speech Targets Identification

The specialist has also identified a set of explicit and implicit terms and expressions that provide the considerable potential to indicate hate speech targets. For example, the term *vadia* (“slut”), and the expression *judeus dos infernos* (“jews from hell”), both cases provide considerable indicative to identify sexists and antisemitism comments. Table 3 shown other examples. In total, the specialist has been identified 150 (one hundred and fifty) hate speech targets, as shown in Section 4.

Table 3: Hate Speech Targets.

Term/Expression	Hate Speech Target
Bitch	Sexism
Dyke	Homophobia
Jews from hell	Antisemistim
Military intervention now!	Apology to dictatorship

3.4 Contextual Information Annotation

As already mentioned, we annotated context information for each identified term or expression. Three different annotators annotated them into a binary class: context-dependent offensive or context-independent offensive. Each annotator has checked whether the identified terms or expressions presented or not any pejorative context of use considering the dictionary meaning, as well as the personal world vision. For example, the expression “wretched” or the terms “hypocritical” and “slut” were labeled as context-independent offensive, because in general these terms are used only in pejorative contexts of use. Differently, the “crazy”, “hit”, and “illiterate” terms were labeled as context-dependent offensive terms because the annotators identified a possible non-pejorative context of use in which these terms may be employed. Figure 2 shown the annotation schema for contextual information classification.

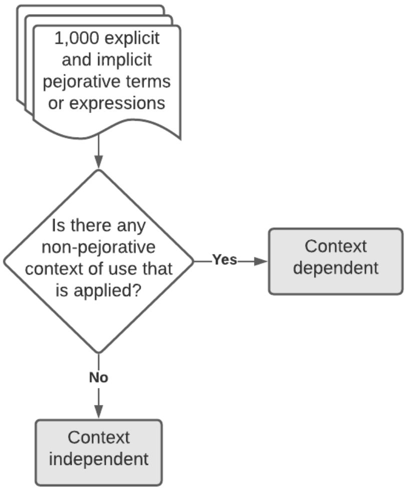
1,000 explicit and implicit pejorative terms or expressions
Is there any non-pejorative context of use that is applied?
Yes
Context dependent
No
Context independent

Figure 2: Annotation schema for contextual information classification.

As displayed by Figure 2, firstly each one of three annotators check if the implicit or explicit terms or expressions presented any non-pejorative context of use considering the dictionary meaning, as well as personal world vision. Subsequently, terms or expressions that had any non-pejorative context of use were classified as *context-dependent offensive*, and terms or expressions that had the only pejorative context of use, were classified as *context-independent offensive*.

We should point out that a couple of identified terms and expressions did not present formal dictionary definitions, such as expressions that are deeply culture rooted (e.g., “macumbeira”[4], “bolsonazi”[5], *moças que ficam na rodovia*[6](“girls who are on the highway”). These cases were signaled and the meanings were proposed by the specialist. In total, we identified 70 (seventy) Brazilian deeply culture rooted terms or expressions, in which the most of cases refer to political domains neologisms.

3.5 Translation

The entire translation process was guided by a linguist and carried out manually by American English, Latin American Spanish, African French, and German native speakers. Firstly, the linguist provides the 1.000 identified pejorative terms and expressions with definitions extracted from the Cambridge dictionary [7]. Table 4 shows examples of translated terms and expressions.

Table 4: Translated Terms or Expressions.

Portuguese (original)	English	Spanish	French	German
Cu	Anus	Culo	Cul	Arsch
Idiota	Idiotic	Idiota	Idiot	Idiot
Inútil	Useless	Inútil	Inutile	Nutzlos

As shown in Table 4, the terms originally written in Portuguese were translated for English, Spanish, French and German languages. The native speaker’s translators were supported by dictionary meaning for each term. The doubts and clarifications were discussed and carried out jointly with the specialist. We should only point out that the Brazilian deeply culture rooted terms were not translated. Lastly, each native speaker’s translator provides a set of deeply culture rooted terms that were incorporated into the proposed MOL and properly marked (e.g. “Brazilian deeply culture rooted”, or “Latin Spanish deeply culture rooted”, etc).

3.6 Annotation Evaluation

We evaluated the consistency of the annotation process using Cohen’s kappa (McHugh, 2012; Sim and Wright, 2005) coefficient. The obtained results are displayed in Table 5, where A, B, and C letters stand for the annotators, and agreement is measured for pairs of them. According to (Landis and Koch, 1977), values around 70% are considered a substantial agreement.

Table 5: Kappa score.

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

4 MOL Statistics

In this section, we present the Multilingual Offensive Lexicon overview, as shown in Tables 6, 7 and 8. We should only point that the MOL is an open-source recourse available in (OMITTED DUE TO DOUBLE-BLIND).

Table 6: Type of Terms and Expressions

Type	Total	%
Pejorative words	909	90,9
Swear words	91	9,10
Total	1,000	100

Table 7: Contextual Labels.

Offense-level Classes	Total	%
Context-Independent Offensive	613	61,30
Context-Dependent Offensive	387	38,70
Total	1,000	100

Table 8: Hate Speech Targets.

Groups	Total	%
Partyism	70	46,66
Sexism	34	22,66
Homophobia	16	10,66
Religious intolerance	9	6,00
Fatphobia	9	6,00
Apology to dictatorship	5	3,33
Racism	4	2,66
Antisemitism	3	2,00
Total	150	100

5 Final Remarks

This paper describes the first Multilingual Offensive Lexicon annotated with contextual information and manually translated by native speakers. The proposed resource consists of 1,000 explicit and implicit pejorative terms and expressions manually identified by a specialist and annotated according to the following classes: context-dependent offensive and context-independent offensive. The MOL is currently available in Portuguese, English, Spanish, French, and German.

[4] *Macumbeira* refers to an Afro-Brazilian cultural element that is pejoratively cited by some religious communities, and names any female practitioner of witchcraft.

[5] *Bolsonazi* is a Brazilian neologism formed by agglutination process of words “Bolsonaro”, which is the current Brazilian’s President, with the word “Nazism”.

[6] *Moças que ficam na rodovia* is a Brazilian pejorative expression to designate prostitute.

[7] https://dictionary.cambridge.org/pt/dicionario/portugues-ingles/

References

Luvell Anderson and Ernie Lepore. 2013. What did you call me? slurs as prohibited words. *Analytic Philosophy*, 54(3):350–363.

Winfried Brugger. 2007. Proibição ou proteção do discurso do ódio? Algumas observações sobre o direito alemão e o americano. *Direito Público*, 4(15):117–136.

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.

Jonathan Culpeper, Paul Iganski, and Abe Sweiry. 2017. Linguistic impoliteness and religiously aggravated hate crime in england and wales. *Journal of Language Aggression and Conflict*, 5(1):1 – 29.

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.

Aldo Frigerio and Maria Paola Tenchini. 2014. *On the Semantic Status of Connotation: The case of Slurs*, Lodz Studies in English and General Linguistics 2, pages 57–75.

Timothy Jay and Kristin Janschewitz. 2008. The pragmatics of swearing. *Journal of Politeness Research - language Behaviour Culture*, pages 267–288.

Daniel Jurafsky. 1996. Universal tendencies in the semantics of the diminutive. *Language*, pages 533–578.

Abhishek Kumar Kashyap. 2011. Book review: Derek bousfield, impoliteness in interaction. amsterdam/philadelphia: John benjamins, 2008. xiii + 281 pp. *Discourse & Society*, 22(1):111–112.

Dániel Z. Kádár, Vahid Parvaresh, and Puyu Ning. 2019. Morality, moral order, and language conflict and aggression: a position paper. *Journal of Language Aggression and Conflict*, 7(1):6 – 31.

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

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

Elyse MethvenWasow. 2017. *Dirty talk: A critical discourse analysis of offensive language crimes*. Ph.D. thesis, University of Technology Sydney, Doctor of Philosophy in Law, Sydney, Australia.

Karsten Müller and Carlo Schwarz. 2020. Fanning the flames of hate: social media and hate crime. *Journal of the European Economic Association*, pages 1–47.

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. 2013. Hate speech. In *Hare, Ivan; Weinstein, James (org). Extreme Speech and Democracy*, volume I, pages 123–138. Oxford-UK: Oxford University Press.

Langton Rae. 2012. Beyond belief: Pragmatics in hate speech and pornography1. *Speech and Harm: Controversies Over Free Speech*, pages 72–93.

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.

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.