MIN_PT: An European Portuguese Lexicon for Minorities Related Terms

Paula Fortuna[1], **Vanessa Cortez**[2], **Miguel Sozinho Ramalho**[3], **Laura Pérez-Mayos**[1]
[1]NLP Group, Pompeu Fabra University,
[2]University of Minho,[3]University of Porto,
paula.fortuna@upf.edu|vanessalc1@gmail.com|m.ramalho@fe.up.pt|laura.perezm@upf.edu

Abstract

Hate speech-related lexicons have been proved to be useful for many tasks such as data collection and classification. However, existing Portuguese lexicons do not distinguish between European and Brazilian Portuguese, and do not include neutral terms that are potentially useful to detect a broader spectrum of content referring to minorities. In this work, we present MIN_PT, a new European Portuguese Lexicon for Minorities-Related Terms specifically designed to tackle the limitations of existing resources. We describe the data collection and annotation process, discuss the limitation and ethical concerns, and prove the utility of the resource by applying it to a use case for the Portuguese 2021 presidential elections.

1 Introduction

Dictionaries and lexicons are commonly used in the field of hate speech automatic detection (Fortuna and Nunes, 2018), with applications ranging from data collection (Silva et al., 2016) to feature extraction (Dadvar et al., 2013) and classification (Tulkens et al., 2016b) by applying some matching function with dictionary terms. However, even though such resources have been proved to be useful in numerous applications, lexical knowledge for hate speech classification has received little attention in literature (Bassignana et al., 2018). This work takes up this demand and focuses on presenting a new European Portuguese Lexicon for Minorities-Related Terms. The need for annotating a new resource derives from two different issues: lack of explicit European Portuguese lexicon, and the need for neutral terms.

Lack of European Portuguese lexicon The existent resources, e.g Hurtlex (Bassignana et al., 2018) or Hatebase[1], do not always distinguish European from Brazilian Portuguese. Both languages are similar and such simplification may serve the purpose of some applications. However, when addressing a nuanced and social phenomenon such as hate speech, the ethnographic differences between Portugal and Brazil require a more fine-grained annotation (e.g words such as "bicha" –fag– or "fufa" –dyke– refer to male and female homosexual individuals only in Portugal and not in Brazil).

Need for neutral terms The annotation of neutral terms in hate speech-related lexicons is not common, specially for low represented languages such as European Portuguese. This limits the application of such resources as those terms open new research venues. First, neutral terms can impact data collection stages as it is possible to identify a broader spectrum of online content referring to minorities. Second, it is possible to use neutral terms for bias detection and control if such terms are present equally in all the classes in a dataset. To overcome this limitation, we collect both offensive and non-offensive minorities' terms.

In what follows, Section 2 provides some background on existing annotated lexicons and their limitations. Section 3 describes the data collection and annotation process, and Section 4 presents the new lexicon. Section 5 presents a use case of the lexicon for the Portuguese 2021 presidential elections. Section 6 addresses some limitations and ethical concerns, and Section 7 summarizes the implications of our work for the automatic hate speech detection field.

2 Related Work

Lexicons can be analyzed in terms of how the data is generated and annotated. While some works have been manually annotated by humans, and others rely on automatic procedures where data is compiled by computational methods, other works

[1]https://hatebase.org/

Proceedings of the Fifth Workshop on Online Abuse and Harms, pages 76–80

conjugate both methods by manually curating the automatically compiled data.

Hatebase is one of the widely used lexicons in the field. It corresponds to a broad multilingual vocabulary manually annotated in terms of different categories (e.g nationality, gender) with data across 95 languages and 175 countries. However, the containing words and phrases have been compiled by non-trained crowdsourced internet volunteers, and therefore the quality of the annotation can not be guaranteed. Moreover, the lexicon does not differentiate Portuguese and Brazilian content. Several works have been using Hatebase terms as keywords for content search in social media platforms, e.g (Davidson et al., 2017; Founta et al., 2018; Radfar et al., 2020). One of these works has contributed particularly to enrich the lexicon English content (Davidson et al., 2017). The authors expand the initial term list with n-grams from the extracted messages when searching with the keywords and finally manually remove irrelevant terms.

Tulkens et al. (2016a) presents another lexicon created to detect racist discourse in dutch social media. Starting with a list of words from the LIWC (Linguistic Inquiry and Word Count) (Tausczik and Pennebaker, 2010), the authors compile a set of terms by applying successive automatic expansions and manual annotation phases.

Hurtlex (Bassignana et al., 2018) is a multilingual lexicon automatically expanded and manually validated with 17 different dimensions such as: negative stereotypes ethnic slurs; professions and occupations; etc. In the set of the discussed lexicons, Hurtlex presents the more complete and complex taxonomy. However, for the purpose of our study, this taxonomy also misses neutral terms to refer to minorities, such as "mulheres" –women– or "muçulmanos" –muslim–, that can help to identify less explicit insults and also positive content. If we focus on the Hurtlex Portuguese subset of terms, we find again no distinction between Brazilian and Portuguese contexts.

Even though the discussed lexicons rely on automatic methods to compile an initial set of terms, they all require manual validation procedures to confirm relevant terms. In this procedure, annotators guarantee that terms match the taxonomy classification rules, which highlights the importance of human annotators to assure higher-quality resources. Accordingly, in this work we also rely on a manual enumeration and annotation of terms to create a new European Portuguese Lexicon for Minorities-Related Terms, containing both offensive and non-offensive terms. Our approach also aligns with the recommendation for synthetic data creation as the compiled data is generated, annotated and validated by experts in an attempt to mimic real behaviour (Vidgen and Derczynski, 2020).

3 Methodology

This section describes the data collection and annotation procedure followed to build MIN_PT, a European Portuguese Lexicon for Minorities Related Terms. We followed a qualitative approach with successive iterations and annotators' participation, as recommended in Vidgen and Derczynski (2020). Starting with and initial set of terms (Section 3.1), the annotators worked individually and collectively in successive iterations to create new annotation rules, remove undesired terms and expand the existent terms with new ones (Section 3.2). Then, two annotators discussed the lexicon terms to reach a consensus on a set of definitions and instructions, deciding which terms are kept and which terms must be eliminated (Section 3.3). The curated list of terms and their classification is available in a public GitHub repository[2].

3.1 Initial data source

For initial data seed, we rely on the Hatebase[3] for Portuguese hate terms; cf. Section 2. While it misses many terms, specially neutral, and mixes Brazilian and European Portuguese, it provides 319 terms and is a good starting point for our new lexicon.

3.2 Data Curation and Enrichment

Starting from the Hatebase for Portuguese hate terms, two annotators curated the list in three individual sessions and two collective sessions with the clear objective of achieving an exhaustive lexicon. The main discussions revolved around clarifying the meaning of diverse terms and deciding on ambiguous terms. The final annotation rules can be described as:

- Remove words that do not match vocabulary from Portugal, e.g. "sangue ruim" –mudblood–, "sapatão" –dyke–.

[2] https://github.com/paulafortuna/Portuguese-minority-terms

[3] https://hatebase.org/

- Enumerate all possible terms. An exhaustive list is achieved by manually: adding synonyms for the same term in case they exist; assuring all terms are present in singular, plural, masculine or feminine, in case such declensions apply; and adding all the known terms for all minority groups.

- Remove ambiguous terms that can have double meaning when the most common usage does not refer to minorities (e.g "preto´´, –black– may be used as an insult but is also a color, tea flavor, etc).

3.3 Data Annotation

After generating a curated list of terms, the annotators classified all terms into the following minorities-related categories: roma, LGBT, migrants, women, people based in religion, people based in ethnicity, and refugees. All the terms were further classified as being an insult (1) or not (0). It is important to notice that terms that can be used as both insults and in a neutral way were classified as not insults. This is the case for certain minority names that can also be used for name-calling (e.g "cigano" –gypsy–).

3.4 Annotators' Description

The two annotators of the MIN_PT lexicon are native Portuguese speakers –one for European and another for Brazilian Portuguese– living in Portugal and aware of the social context. Both identify as cis-gender women and correspond to two authors of the work with previous annotation experience.

4 Results

The MIN_PT European Portuguese Lexicon for Minorities Related Terms is composed of 155 carefully curated terms (cf. Section 3) related to 7 minority groups, as described in Table 1.

Even though our new lexicon contains much less terms than Hurtlex (Bassignana et al., 2018), 155 vs 3902 terms, it is worth noticing that only 23% of the terms in MIN_PT are present in Hurtlex. Therefore, the new lexicon presented in this work will prove to be a valuable resource for hate-speech detection, either on its own or in combination with other resources.

Minority group	Total	Insults
LGBT	44	20
People based on ethnicity	44	30
Women	29	24
Migrants	22	0
No minority	9	9
Roma	8	4
Religious people	6	0
Refugees	2	0

Table 1: MIN_PT lexicon terms frequency per class.

5 Lexicon Application: The case of Portuguese 2021 Presidential Elections

The annotation of this lexicon was motivated by the will to conduct an analysis on the Portuguese 2021 presidential elections twittersphere, aiming at understanding whether and how candidates' speeches and replies would tackle minority topics. The analyzed data is a subset from the *Portuguese Presidential Elections, Jan 24th 2021* (Ramalho, 2021) and corresponds to 35,101 tweets from September 2nd, 2020 to November 22th, 2020.

For the six candidates using Twitter, we performed a keyword matching with the terms in the MIN_PT lexicon to compute the percentage of tweets (Figure 1) and their replies (Figure 2) referring to minorities. Marisa Matias (*mmatias_*), André Ventura (*AndreCVentura*) and Ana Gomes (*AnaMartinsGomes*) are the candidates tackling a higher percentage of minorities topics. However, the targeted minorities are distinct depending on the candidate. While Ana Gomes focused more uniformly on the different groups, Marisa Matias discussed more refugees and women issues and André Ventura focused on Roma and people based on ethnicity, i.e racism issues. Comparing both figures, it is also interesting to see that the candidates' audience does not exactly resonate with the candidate in terms of mentioned minority topics. Moreover, while none of the candidates mentions any of the explicit insults in our lexicon, they were present in the audience.

While our lexicon proved to be valuable for an initial topic analysis, a more in depth analysis should be performed to get further insights on how politicians are referring to minorities.

6 Limitations and Ethical Concerns

Lexicons are static resources that can not mimic the contextual and mutating nature of language, and certain terms may refer to minorities, be considered as insults or just be neutral words depending on the

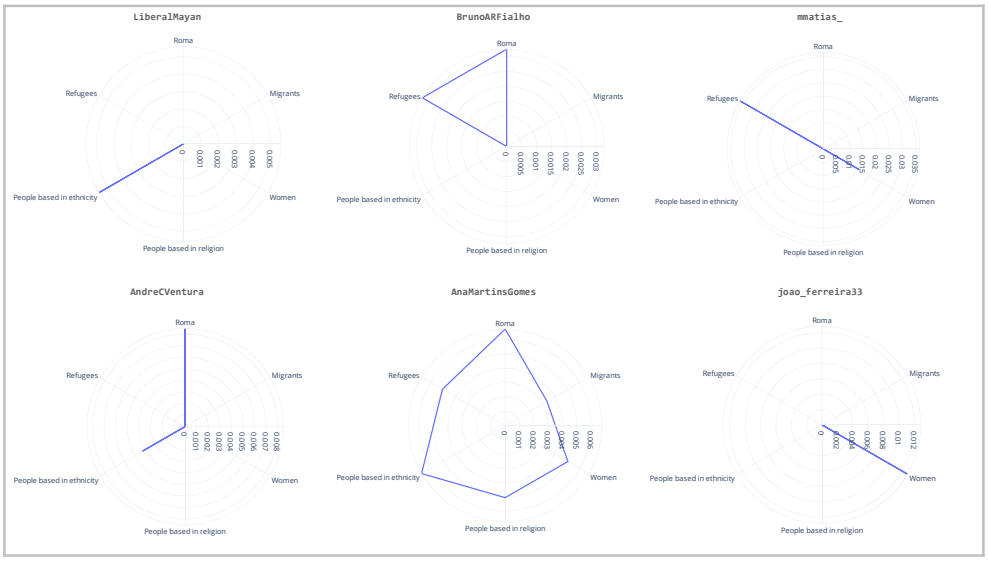

Figure 1: Relative frequencies of minority mentions in candidates' tweets.

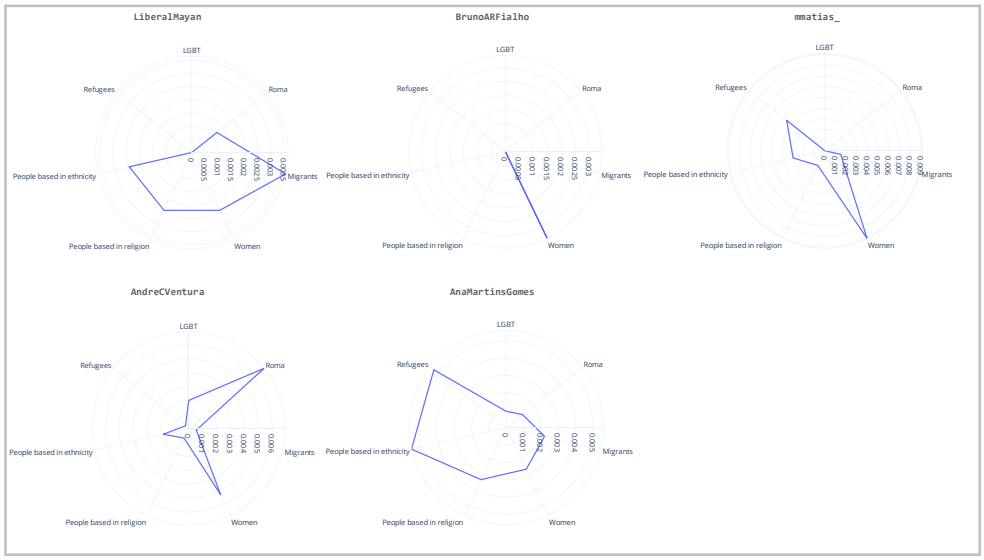

Figure 2: Relative frequencies of minority mentions in replies to candidates.

context in which they are used. Our annotation was done with the objective of analyzing politicians discourses and interactions on Twitter, and we explicitly removed ambiguous terms from the lexicon. Therefore, future users must be warned that the terms should be validated when used with other datasets and contexts.

Finally, even though the presence of the terms in our lexicon may imply hate speech against minorities, it should not be used for direct hate speech classification with keyword matching. Depending on the context and the data author, such terms may have a neutral and even positive meaning.

7 Conclusions

We presented MIN_PT, a new European Portuguese Lexicon for Minorities-Related Terms. We discussed existing annotated lexicons, grounding the need for a new lexicon. Following a qualitative approach, we produced a high-quality lexicon containing also neutral words and specific for European Portuguese. We also presented a use case of the lexicon on the analysis of Portuguese politicians' tweets. Future iterations of this work would benefit from the contribution of more annotators to increase the diversity of the available vocabulary.

Acknowledgements

Paula Fortuna is supported by the research grant SFRH/BD/143623/2019, provided by the Portuguese national funding agency for science, research and technology, Fundação para a Ciência e a Tecnologia (FCT), within the scope of Operational Program *Human Capital* (POCH), supported by the European Social Fund and by national funds from MCTES.

References

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018*, volume 2253 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving cyberbullying detection with user context. In *Advances in Information Retrieval*, pages 693–696, Berlin, Heidelberg. Springer Berlin Heidelberg.

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 Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*, pages 512–515. AAAI Press.

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

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*, pages 491–500. AAAI Press.

Bahar Radfar, Karthik Shivaram, and Aron Culotta. 2020. Characterizing variation in toxic language by social context. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020*, pages 959–963. AAAI Press.

Miguel Sozinho Ramalho. 2021. High-level approaches to detect malicious political activity on twitter.

Leandro Araújo Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. In *Proceedings of the Tenth International Conference on Web and Social Media, Cologne, Germany, May 17-20, 2016*, pages 687–690. AAAI Press.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. 2016a. The automated detection of racist discourse in dutch social media. *Computational Linguistics in the Netherlands Journal*, 6(1):3–20.

Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. 2016b. A dictionary-based approach to racism detection in Dutch social media. In *Proceedings of the LREC 2016 Workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS)*. European Language Resources Association (ELRA).

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.