A Hierarchically-Labeled Portuguese Hate Speech Dataset

Paula Fortuna[1,3], **João Rocha da Silva**[1,2],
Juan Soler-Company[3], **Leo Wanner**[3,4], **Sérgio Nunes**[1,2]

[1]INESC TEC, [2]FEUP, University of Porto
Rua Dr. Roberto Frias, s/n 4200-465 Porto, Portugal
[3]NLP Group, ETIC, Pompeu Fabra University, Barcelona, Spain
[4]Catalan Institute for Research and Advanced Studies (ICREA), Barcelona, Spain
paula.fortuna@fe.up.pt, joaorosilva@gmail.com
juan.soler@upf.edu, leo.wanner@upf.edu, sergio.nunes@fe.up.pt

Abstract

Over the past years, the amount of online offensive speech has been growing steadily. To successfully cope with it, machine learning is applied. However, ML-based techniques require sufficiently large annotated datasets. In the last years, different datasets were published, mainly for English. In this paper, we present a new dataset for Portuguese, which has not been in focus so far. The dataset is composed of 5,668 tweets. For its annotation, we defined two different schemes used by annotators with different levels of expertise. First, non-experts annotated the tweets with binary labels ('hate' vs. 'no-hate'). Then, expert annotators classified the tweets following a fine-grained hierarchical multiple label scheme with 81 hate speech categories in total. The inter-annotator agreement varied from category to category, which reflects the insight that some types of hate speech are more subtle than others and that their detection depends on personal perception. The hierarchical annotation scheme is the main contribution of the presented work, as it facilitates the identification of different types of hate speech and their intersections. To demonstrate the usefulness of our dataset, we carried a baseline classification experiment with pre-trained word embeddings and LSTM on the binary classified data, with a state-of-the-art outcome.

1 Introduction

The Internet is the source of an immense variety of knowledge repositories (Wikipedia, Wordnet, etc.) and applications (YouTube, Reddit, Twitter, etc.) that everybody can access and take advantage of; it is also **the** communication forum of our time and the most important instrument to ensure freedom of speech. It allows us to freely state and disseminate our view on any private or public matter to vast audiences. But unfortunately it also opens the door to manipulation of masses and defamation of specific individuals or groups of people. One of these observed negative phenomena is the propagation of hate speech. Hate speech leads to a negative self-image and social exclusion of the targeted individuals, groups or populations, and incites violence against them. A clear example of the extreme harm that can be caused by hate speech is the 1994 Rwandan genocide; see Schabas (2000) for a detailed analysis. The detection of online hate speech is thus a pressing problem that calls for solutions. Over the last decade, a considerable number of supervised machine learning-based works tackled the problem. Most of them focused on English (Waseem and Hovy, 2016; Davidson et al., 2017; Nobata et al., 2016; Jigsaw, 2018), see also the overview by Schmidt and Wiegand (2017). As a result, also many more annotated datasets, which are the precondition for the use of supervised machine learning, are available for English (e.g., Waseem and Hovy (2016); Davidson et al. (2017); Nobata et al. (2016); Jigsaw (2018)) than for other languages. However, hate speech is not a phenomenon that is observed only in English discourse; it is notorious in online media in other languages as well; cf., e.g., Spanish (Fersini et al., 2018), Italian (Poletto et al., 2017; Sanguinetti et al., 2018), or German (Ross et al., 2016).

In this work, we aim to contribute to the field of hate speech detection. Our contribution is twofold: (i) diversification of the research on hate speech by provision of a new dataset of hate speech in another language than English, namely Portuguese; (ii) introduction of a novel fine-grained hate speech typology that improves on the common state-of-the-art used typologies, which tend to disregard the existence of subtypes of hate speech and either consider hate speech recognition as a binary classification task, or take into account only a few classes, such as 'racism'

Proceedings of the Third Workshop on Abusive Language Online, pages 94–104
Florence, Italy, August 1, 2019.

and 'sexism' (Waseem and Hovy, 2016) – despite the fact that such broad distinctions unduly overgeneralize. For instance, by classifying discrimination against both black people and refugees simply as 'racism', we ignore that in this case, different characteristics with a different motivation are targeted (also reflected in a different language style). In particular, we compile and annotate a new dataset composed of 5,668 tweets in Portuguese, which is one of the most commonly-used languages online (Fox, 2013). Two types of annotations are carried out. For the first, non-expert annotators classify the messages in a binary fashion ('hate' vs. 'no-hate'). For the second, we build a multilabel hate speech hierarchical annotation schema with 81 hate categories in total[1]. To demonstrate the usefulness of our dataset, we carried a baseline classification experiment with pretrained word embeddings and LSTM on the binary classified data, with a state-of-the-art outcome.

The remainder of the paper is structured as follows: Section 2 reviews the literature. Section 3 describes our crawling procedure. In Section 4, we present the two annotation schemas we work with: the binary and the hierarchical schema. Section 5 discusses a baseline hate speech experiment that we carried out to validate our new dataset. Section 6 presents some ethical considerations of this work. In Section 7, finally, the conclusions of our work are presented.

2 Related Work

2.1 Hate Speech Concepts

Fortuna and Nunes (2018) analyze and compare several aggression-related concepts. As a result of their analysis, they present the following definition of *hate speech*:

> "Hate speech is 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 can occur with different linguistic styles, even in subtle forms or when humour is used."

We adopted this definition in our work. Our work has also been inspired by the taxonomy provided by Salminen et al. (2018), which includes 29 hate categories characterized in terms of hateful language, target, and sub-target types. To create their taxonomy, Salminen et al. followed an iterative and qualitative procedure called "open coding" (Glaser and Strauss, 2017).

There are obvious similarities between Salminen et al.'s approach and ours. However, there are also some significant differences. The first difference concerns the underlying definition of hate. While they use the very generic definition "hateful comments toward a specific group or target", the definition we adopt is more specific (cf. above). This leads to differences in the taxonomy. For instance, they introduce 'hate against media' and 'hate against religion', which is hate against abstract entities and not considered by us. Additionally, they merge in the same hate speech taxonomy the targets of hate and the type of discourse. In our case, we focus on the targets of hate speech only.

2.2 Dataset Annotation

Several hate speech datasets are publicly available, e.g., for English (Waseem and Hovy, 2016; Davidson et al., 2017; Nobata et al., 2016; Jigsaw, 2018), Spanish (Fersini et al., 2018), Italian (Poletto et al., 2017; Sanguinetti et al., 2018), German (Ross et al., 2016), Hindi (Kumar et al., 2018), and Portuguese (de Pelle and Moreira, 2017). In this section, we analyze the data collection strategy, the annotation method and the dataset properties of three representative hate speech datasets: the Hate speech, Racism and Sexism dataset by Waseem and Hovy (2016), the Offensive Language Dataset by Davidson et al. (2017), and the Portuguese News Comments dataset by de Pelle and Moreira (2017). We have chosen the first two because they are the most widely used datasets for English hate speech automatic classification. They show how Twitter can be used to retrieve information and how to conduct the manual classification relying on both expert and non-expert annotators. The third is another annotated and published dataset for Portuguese, which is rather different from ours.

Hate speech, Racism and Sexism Dataset. This dataset[2] (Waseem and Hovy, 2016) contains 16,914 tweets in English, which were classified by two annotators using the classes "Racism", "Sexism" and "Neither". Regarding the tweet collection, an initial manual search was conducted on Twitter to collect common slurs and terms related to religious, sexual, gender, and ethnic minorities.

[1] https://github.com/paulafortuna/Portuguese-Hate-Speech-Dataset

[2] https://github.com/ZeerakW/hatespeech

The authors identified frequently occurring terms in tweets that contain hate speech and used those terms to retrieve more messages. The messages were then annotated by the main researcher, together with a gender studies student; in total, 3,383 tweets as sexist, 1,972 as racist, and 11,559 as neither sexist nor racist. The inter-annotator agreement had a Cohen's Kappa of 0.84. The authors of the study concluded that the use of n-grams provides good results in the task of automatic hate speech detection, and adding demographic information leads to little improvement in the performance of the classification model.

Offensive Language Dataset. Davidson et al. (2017) annotated a dataset[3] with 14,510 tweets in English, using the classes "Hate", "Offensive" and "Neither". Regarding the collection of the messages, they started with an English hate speech lexicon compiled by Hatebase.org, searching for tweets that contained terms from this lexicon. The outcome was a collection of tweets written by 33,458 Twitter users. The collected tweets were completed by further follow-up tweets of these users, which resulted in a corpus of 85.4 million tweets. Finally, from this corpus, a random sample of 25,000 tweets containing terms from the lexicon has been extracted and manually annotated by CrowdFlower workers. Three or more workers from CrowdFlower annotated each message. The majority voting was used to assign a label to each tweet. Tweets that did not have a majority class were discarded. This resulted in a sample of 24,802 labeled tweets. The inter-annotator agreement score provided by CrowdFlower was 92%. However, a total percentage of only 5% of tweets were labeled as hate speech by the majority of the workers.

Portuguese News Comments Dataset. de Pelle and Moreira (2017) collected a dataset[4] with 1,250 random comments from the Globo news site on politics and sports news. Each comment was annotated by three annotators, who were asked to indicate whether it contained 'racism', 'sexism', 'homophobia', 'xenophobia', 'religious intolerance', or 'cursing'. 'Cursing' was the most frequent label, while the other labels had few instances in the corpus. Regarding the annotator agreement, the value was 0.71.

In comparison to this work, the dataset that we have compiled provides more data and is not restricted to specific topics. Additionally, our annotation focuses only on hate speech, instead of general offensive content. We also use and provide a complete labeling schema.

Compared to the previous two datasets, our second annotation schema is considerably more fine-grained. As we will see below, our annotation procedure with the fine-grained schema is similar to that of Waseem and Hovy (2016).

2.3 Classification methods

Different studies conclude that deep learning approaches outperform classical machine learning algorithms in the task of hate speech detection; see, e.g., Mehdad and Tetreault (2016); Park and Fung (2017); Del Vigna et al. (2017); Pitsilis et al. (2018); Founta et al. (2018); Gambäck and Sikdar (2017). For instance, Badjatiya et al. (2017) compare the use of different types of neural networks (CNN, LSTM) and deep learning libraries such as FastText with the use of classical machine learning techniques and experiment with different types of word embeddings. The setup that achieved the best performance consists of the combination of deep techniques with standard ML classifiers, and more precisely, of embeddings learned by an LSTM model, combined with gradient boosted decision trees. We will follow a similar methodology for classification.

3 Message Collection

Our overall approach to message collection is outlined in Figure 1. In what follows, we introduce in detail the individual steps.

Use of Keywords and Profiles. We used Twitter's search API for keywords and profiles because both can be complementary as message sources. With the first, we access a wider range of tweets from different profiles, but we restrict the search to specific words or expressions that indicate hate. With the second, we obtain more spontaneous discourse, but from a more restricted number of users:

- **Hate-related keywords:** We used Twitter's API search feature to look for keywords and hashtags related to hate speech, such as *fufas*, *sapatão* 'dyke' or *#LugarDeMulherENaCozinha* '#womensPlaceIsInTheKitchen'.

[3] https://github.com/t-davidson/hate-speech-and-offensive-language

[4] https://github.com/rogersdepelle/OffComBR

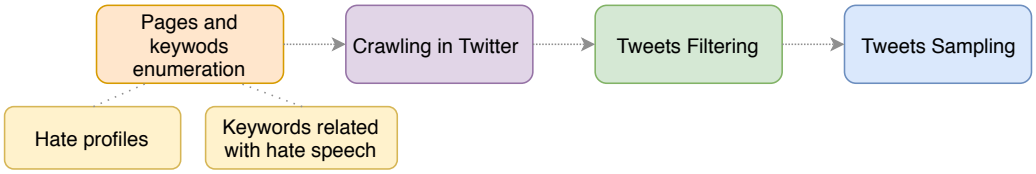

Figure 1: Method for message collection.

- **Hate-related profiles:** Using the profile search API, we query with words like *ódio* 'hate', *discurso de ódio* 'hate speech' and *ofensivo* 'offensive' in order to find accounts that post hateful messages. In Portuguese, there are social media users whose profile is built specifically for sharing hateful content against certain minorities. We collect the messages from those accounts with the expectation to find hate speech messages. This search also allowed us to find counter hate profiles. Those also use the same words in their description. It seemed adequate to keep these profiles because they reproduce hate speech messages from other users.

We looked at 29 specific profiles and used 19 keywords and ten hashtags in a total for 58 search instances.[5] The goal has been to be exhaustive and cover different types of discrimination, based on religion, gender, sexual orientation, ethnicity, and migration. We compiled this collection of search instances because there was no specific hate speech lexicon available for Portuguese, e.g., Hatebase contains generic hate (Hatebase, 2019).

Crawling. We used R to crawl content with respect to both keywords and profiles content on the 8th and 9th of March of 2017. A total of 42,930 messages were collected.

Tweet Filtering. We kept tweets categorized by Twitter as written in Portuguese. We eliminated repetitions and retweets from already collected messages to avoid duplication and removed HTML tags and messages with less than three words.

Tweet Sampling. The procedure previously described resulted in 33,890 tweets. We noticed that the search instances returned several tweets from different magnitudes (e.g., some profiles had only around 30 messages while others had more than 3,000). We decided then to use a maximum of 200 tweets per search instance in order to keep a more diverse source of tweets.

Final Dataset. Our final dataset contains 5,668 tweets, containing content from 1,156 different users. The majority of the tweets (more than 95%) are from January, February, and March of 2017.

4 Annotation of Hate Speech

In what follows, we present the annotation procedures for binary hate speech and hierarchical hate speech annotation.

4.1 Binary annotation

Three annotators classified every message. 18 Portuguese native speakers (Information Science student volunteers) were given annotation guidelines to perform the task (cf. Appendix A.1). All of them received an equivalent number of messages. The annotation was binary and the annotators had to label each message as 'hate speech' or 'not hate speech'.

To check the agreement between the three classifications of every message, we used Fleiss's Kappa (Fleiss, 1971). We observed a low agreement with a value of K = 0.17. We think that this low value is the result of relying exclusively on non-expert annotators for classifying hate speech. For instance, in Waseem and Hovy (2016), the two annotators were the author of the study plus a gender studies student. On the other hand, the two other studies mentioned in Section 2 (de Pelle and Moreira, 2017; Davidson et al., 2017), are more generic in that they do not focus exclusively on hate speech (as we do), but rather consider offensive speech in general, which includes insults that are more explicit and easier to recognize, while hate speech is subtler and more difficult to identify.

For our final annotation, we applied the majority vote, which resulted in a dataset in which 31.5% of the messages are annotated as 'hate speech'.

[5] We use the term "search instance" to refer to profiles, keywords or hashtags used for the Twitter search.

4.2 Hierarchical annotation

When studying hate speech, it is possible to distinguish between different categories of it, like 'racism', 'sexism', or 'homophobia'. A more fine-grained view can be useful in hate speech classification because each category has a specific vocabulary and ways to be expressed, such that creating a language model for each category may be helpful to improve the automatic detection of hate speech (Warner and Hirschberg, 2012).

Another phenomenon we can observe when analyzing different categories of hate speech is their *intersectionality*. This concept appeared as an answer to the historical exclusion of black women from early women's rights movements often concerned with the struggles of white women alone. Intersectionality brings attention to the experiences of people who are subjected to multiple forms of discrimination within a society (e.g., being woman and black) (Collins, 2015). Waseem (2016) introduce a hate speech labeling scheme that follows an intersectional approach. In addition to 'racism', 'sexism', and 'neither', they use the label "both" arguing that the intersection of multiple oppression categories can differ from the forms of oppression it consists of (Crenshaw, 2018).

To better take into account different hate speech categories from an intersectional perspective, we approach the definition of the hate speech annotation schema in terms of a hierarchical structure of classes.

4.2.1 Hate speech and hierarchical classification

In hierarchical classification, there is a structure defining the hierarchy between the categories of the problem (Dumais and Chen, 2000). This is opposed to flat classification, where categories are treated in isolation. Several structures can be used to represent a hierarchy of classes. One of them is a *Rooted Directed Acyclic Graph* (rooted DAG), where each class corresponds to a node and can have more than one parent. Another property of this graph is that documents can be assigned to terminal categories and to non-terminal node categories alike (Hao et al., 2007). In the specific case of hate speech classification, we propose to use a rooted DAG in order to be able to cover hate speech subtypes and their intersections, as exemplified in Figure 2. The graph of classes has the following properties:

- The 'hate speech' class corresponds to the root of the graph.
- If hate speech can be divided into several types of hate, several nodes descend from the root node. This gives rise to the second level of classes (Table 1) according to the targets of the hate (e.g., 'racism', 'homophobia', and 'sexism').
- This second level of nodes can also be divided into subgroups of targets. For instance, racist messages can be targeted against black people, Chinese people, Latinos, etc.
- The division of classes can continue until we do not find more distinct groups, resulting in a terminal node.
- The lower nodes of the graph inherit the classes from the upper nodes, up to the root.
- The lower nodes of the graph can have one or more parents. In the second case, this gives rise to a class that intersects the parent classes.
- Instances are classified according to a multi-label approach and can belong to classes assigned to both terminal and/or non-terminal nodes.

Class	Definition
Sexism	Hate speech based on gender. Includes hate speech against woman.
Body	Hate speech based on body, such as fat, thin, tall or short people.
Origin	Hate speech based on the place of origin.
Homophobia	Hate speech based on sexual orientation.
Racism	Hate speech based on ethnicity.
Ideology	Hate speech based on a person's ideas, such as feminist or left wing ideology.
Religion	Hate speech based on religion.
Health	Hate speech based on health conditions, such as against disabled people.
Other-Lifestyle	Hate speech based on life habits, such as vegetarianism.

Table 1: Direct subtypes of the 'hate speech' type.

This annotation schema has several advantages compared to standard binary or disjoint flat classification. Firstly, it models in a better way the relationships between different subtypes of hate speech. Additionally, it preserves rare classes, while signaling them as part of more generic

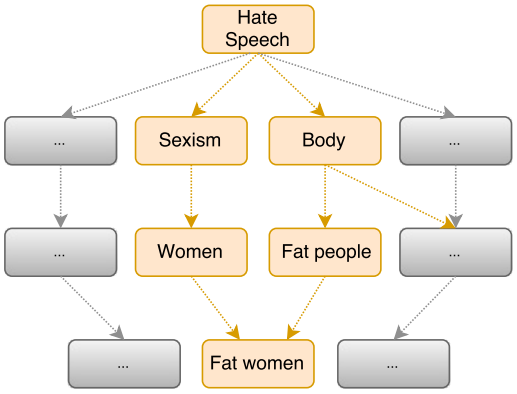

Figure 2: Part of the rooted directed acyclic graph used for hate speech classification.

classes. For instance, with this classification, we can use a message to build a model for predicting sexism even if the message was cataloged as 'hate against fat women'. Finally, with this approach, it is possible to study each subtype of hate speech individually or in relation to others, depending on the goal of the study.

In the next subsection, we outline the hierarchical annotation procedure conducted with the dataset described in Section 3, which complements the non-expert annotation.

4.2.2 Building the hierarchy of hate speech

Similarly to Salminen et al. (2018), we use for the annotation a data-driven approach based on an open coding methodology. This means that we iteratively protocol the different classes as they appear in the dataset while we read and classify the data. The classification hierarchy is then built by creating and reorganizing categories until all available data was analyzed. For this annotation, we applied an intersectional approach by enumerating all the possible groups cited in our dataset, no matter their frequency (e.g., 'feminist men' appears only once).

Based on all instances of the dataset, the hierarchy of classes was built by one researcher working in the area of automatic detection of hate speech, with training in social psychology. Then, the same researcher classified all the dataset messages using the hierarchical class structure.

4.2.3 Agreement between annotators

For verifying the validity of this annotation procedure, a second annotator classified 500 messages. Then, we used Cohens Kappa (Gamer et al., 2012) for checking the agreement between both. We observed K = 0.72. We also consider the agreement of the annotators by type of hate speech. We ranked the classes by the best agreement and removed the classes with only one instance for any of the annotators. We found diverse values in the different categories (Table 2), which points out that some specific types of hate speech can be more difficult to classify than others.

Classes	K	Annotator 1	Annotator 2
Lesbians	0.879	59	53
Health	0.856	3	4
Homofobia	0.823	69	61
Disabled people	0.799	2	3
Refugees	0.763	13	13
Migrants	0.751	15	14
Sexism	0.669	134	104
Trans women	0.662	6	9
Men	0.657	12	15
Women	0.642	109	75
Fat women	0.637	30	16
Body	0.637	32	17
Fat people	0.637	32	17
Ideology	0.609	14	15
Feminists	0.581	13	14
Hate speech	0.569	245	213
Racism	0.501	18	13
Religion	0.493	5	11
Black people	0.435	11	7
Origin	0.329	3	3
Islamists	0.329	2	10
Gays	0.300	4	9
Ugly women	0.276	24	4

Table 2: Annotator agreement by class, with the number of messages annotated by each annotator.

4.3 Hierarchical dataset

After the annotation phase, we obtain a multi-labeled dataset with 22% of hate speech instances. The resulting hierarchy, the node depth (ND) and class frequencies (Freq) are presented in Table 3. As expected, the classes corresponding to nodes with a higher depth tend to have a smaller frequency. Note that our schema also identifies categories that are less commonly mentioned in hate speech classification experiments, among them, e.g., 'fat people', 'fat women', 'ugly people', 'ugly women', 'men', 'feminists', 'people with left-wing ideology'. Some of them (such as, e.g., 'men') may look neutral at the first glance, but, in reality, they group messages whose vocabulary and language style reflect negative expectations towards the corresponding collective (in the case of men those expectations reflect toxic masculinity norms).

Class	ND	Parent nodes	Freq	Class	ND	Parent nodes	Freq
Hate speech	0	-	1228	Ageing	1	Hate speech	4
Sexism	1	Hate speech	672	Angolans	3	Africans	4
Women	2	Sexism	544	Nordestines	3	Rural people, Brazilians	4
Homophobia	1	Hate speech	322	Chinese	3	Asians	3
Homossexuals	2	Homophobia	288	Homeless	2	Other/Lifestyle	3
Lesbians	3	Homossexuals, Woman	248	Arabic	2	Origin	2
Body	1	Hate speech	164	Bissexuals	2	Homophobia	2
Fat people	2	Body	160	Blond women	2	Women, Body	2
Fat women	3	Women, Fat people	153	East europeans	2	Origin	2
Ugly people	2	Body	131	Jews	2	Religion	2
Ugly women	3	Women, Ugly people	130	Jornalists	2	Other/Lifestyle	2
Racism	1	Hate speech	94	Old people	2	Ageing	2
Ideology	1	Hate speech	92	Thin people	2	Body	2
Migrants	1	Hate speech	82	Thin women	3	Women, Thin people	2
Men	2	Sexism	70	Vegetarians	2	Other/Lifestyle	2
Refugees	2	Migrants	70	White people	2	Racism	2
Feminists	2	Ideology, Sexism	65	Young people	2	Ageing	2
Gays	3	Homossexuals	56	Agnostic	2	Ideology	1
Black people	2	Racism	52	Argentines	3	Latins	1
Religion	1	Hate speech	30	Autists	2	Health	1
Left wing ideology	2	Ideology	26	Brazilian women	3	Women, South Americans	1
Origin	1	Hate speech	26	Egyptians	3	Arabic	1
Trans women	3	Women, Transexuals	26	Football players women	2	Women, Other/Lifestyle	1
OtherLifestyle	1	Hate speech	20	Gamers	2	Other/Lifestyle	1
Islamists	2	Religion	17	Homeless women	3	Women, Homeless	1
Immigrants	2	Migrants	15	Indigenous	2	Racism	1
Transexuals	2	Sexism	14	Iranians	3	Arabic	1
Muslims	2	Religion	11	Japaneses	3	Asians	1
Black Women	3	Women, Black people	8	Men Feminists	3	Feminists, Men	1
Criminals	2	Other/Lifestyle	8	Mexicans	3	Latins	1
Latins	2	Racism, Origin	7	Muslim women	3	Muslims, Women	1
Health	1	Hate speech	6	Old women	3	Women, Old people	1
Rural people	2	Origin	6	Polyamorous	2	Other/Lifestyle	1
Travestis	3	Women	6	Poor people	2	Other/Lifestyle	1
Aborting women	3	Women	5	Russians	3	East europeans	1
Asians	2	Racism, Origin	5	Sertanejos	3	Rural people, Brazilians	1
Brazilians	3	South Americans	5	Street artists	2	Other/Lifestyle	1
Disabled people	2	Health	5	Ucranians	3	East europeans	1
South Americans	2	Origin	5	Venezuelans	3	Latins	1
Africans	2	Origin	4				

Table 3: Hate subclasses (Class) and respective parent categories (Parent nodes) sorted by frequency (Freq). Information of the node depth is also provided (ND).

5 Binary classification experiment

In order to obtain a first indicator of the usefulness of our dataset, we carry out a preliminary binary classification experiment.

5.1 Methodology

To perform the experiment, we use 10-fold cross-validation (Chollet, 2017), combined with holdout validation, in which one part of the data is used for cross-validation and parameter tuning with grid search and the other part of unseen data is then used for testing.

As already Badjatiya et al. (2017), we provide our source code [6]. We use Python 3.6, Keras (Chollet et al., 2015), Gensim (Řehůřek and Sojka, 2010) and Scikit-learn (Pedregosa et al., 2011) as main libraries. The following subsections describe how we implement each step performed by our system.

[6] https://github.com/paulafortuna/SemEval_2019_public

Text pre-processing As far as text pre-processing is concerned, we remove stop words using Gensim, and punctuation using the default string library and transform all tokens in the tweets to lower case.

Feature extraction: Regarding the features in our experiment, we use pre-trained Glove word embeddings with 300 dimensions for Portuguese (Hartmann et al., 2017). Methods provided by Keras are then used to map each token in the input to an embedding.

Classification: For classification, we use a deep learning model, namely LSTMs, in an architecture as already proposed by Badjatiya et al. (2017). The architecture contains an embedding Layer with the weights from the word embeddings extraction procedure, an additional LSTM layer with 50 dimensions, and dropouts at the end of both layers. As loss function, we used binary cross-entropy and for optimization Adam, 10 epochs and 128 for batch size. With this model, we classify data into binary classes, and we save the last layer

before the classification to extract 50 dimensions as input to the xgBoost algorithm,[7] which is a gradient boosting implementation from the Python library (Chen and Guestrin, 2016).

For xgBoost, the default parameter setting has been used, except for 'eta' and 'gamma'. In this case, we conducted a grid search combining several values of both (eta: 0, 0.3, 1; and gamma: 0.1, 1, 10) in order to obtain the optimal eta and gamma settings. Figure 3 shows a graphical representation of our model.

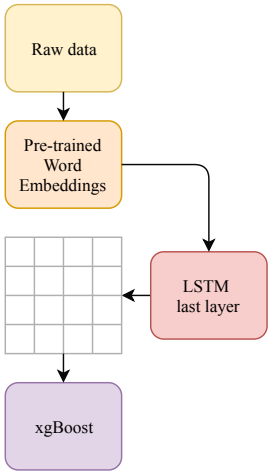

Figure 3: Classification method used as baseline for binary hate speech classification with the Portuguese dataset.

5.2 Results

In this section, we present the results of our classification experiment for classification of hate speech in Portuguese. Table 4 shows the baseline results of the LSTM-based model on our new dataset. We provide the cross validation and test set F1 scores and also the number of instances we used in each of these (N). The results show a state-of-the-art outcome. We can thus assume that even if annotated merely in terms of basic binary ('hate' vs. 'not hate speech') labels, our dataset already constitutes a valid hate speech resource.

6 Ethical considerations

Regarding the ethical aspects of this study, we took into consideration the privacy of the authors of the collected messages. However, we acknowledge the limitations of our sampling procedure when studying online hate speech. The data was anonymized by omitting the tweet id. As a consequence, it is possible to reach the original tweet and user only by a search for the exact text of the tweet. To also prevent this, we make our dataset available in GitHub only for research purposes under the condition that no such a search is performed. A disclaimer is attached, stating that any attempt to violate the privacy of Twitter users is against the established usage conditions, and that the authors of this paper cannot be made liable for this violation.

As far as the quality of the data collection is concerned, sampling bias may have been introduced. Firstly, because Twitter API was used and this provides only a subset of the all posted data in the platform. Secondly, we use a set of keywords and crawl profiles based on our decision criteria, as explained in Section 3. However, we do not aim to have a representative sample of online hate speech on Twitter. We consider that for building a dataset with examples of hate speech, our method is adequate, and that we could find diverse hate speech instances belonging to 80 different classes.

[7] We also experimented with higher dimensionality, but this did not improve the performance of the classifier.

	Hate speech dataset (PT)
CV f1-score	0.78
training data (N)	5099
test set f1-score	0.72
testing data (N)	567

Table 4: Results of Portuguese hate speech classification with the new dataset presented in this paper for binary classification. We provide the micro-averaged F1 scores and also the number of instances used in each of the datasets (N).

7 Conclusions and Future Work

In this work, we built a Portuguese dataset for research in hate speech detection.

To gather our data, we crawled Twitter for messages and manually annotated them using guidelines. Firstly, we developed a method for binary classification using the classification of three annotators per message as ground truth. With this dataset, we conducted a baseline classification experiment using pre-trained word embeddings and LSTM, achieving very competitive performance.

Furthermore, we provided a hate speech hierarchical labeling schema that integrates the complexity of hate speech subtypes and their intersections. This allowed us to find out that distinct types of hate speech present different agreement levels between annotators. Therefore, future guide-

lines for annotation may benefit from specifying the particularities of the different subtypes of hate speech.

As far as future work is concerned, in the context of the annotation procedure, the agreement between annotators can still be improved. We think that the subjectivity of the task makes the learning process challenging and more specific training is necessary for the annotators. Additionally, based on our experiment, we suggest that future data collection procedures should assure sampling of different subtypes of hate to improve the identification of less common subtypes.

Finally, in future explorations of this dataset, we will experiment with multilabel classification of hate speech to identify not only whether a message contains hate, but also the targeted groups.

Acknowledgments

This work was partially funded by the Google DNI project Stop PropagHate. Soler-Company and Wanner have been supported by the European Commission under the contract numbers H2020–7000024-RIA and H2020-786731-RIA. We would like to thank the anonymous reviewers for their insightful comments and to the annotators for their contribution to this work.

References

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760. International World Wide Web Conferences Steering Committee.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

François Chollet et al. 2015. Keras. https://keras.io, accessed last time in February 2019.

Francois Chollet. 2017. *Deep learning with python*. Manning Publications Co.

Patricia Hill Collins. 2015. Intersectionality's definitional dilemmas. *Annual Review of Sociology*, 41:1–20.

Kimberle Crenshaw. 2018. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory, and antiracist politics [1989]. In *Feminist legal theory*, pages 57–80. Routledge.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of ICWSM*.

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 First Italian Conference on Cybersecurity*, pages 86–95.

Susan Dumais and Hao Chen. 2000. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 256–263. ACM.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018. Overview of the task on automatic misogyny identification at ibereval 2018.

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 (CSUR)*, 51(4):85.

Antigoni-Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2018. A unified deep learning architecture for abuse detection. *arXiv preprint arXiv:1802.00385*.

Zoe Fox. 2013. Top 10 most popular languages on twitter. Available in http://mashable.com/2013/12/17/twitter-popular-languages/, accessed last time in May 2017.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using Convolutional Neural Networks to Classify Hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90.

Matthias Gamer, Jim Lemon, Maintainer Matthias Gamer, A Robinson, and W Kendall's. 2012. Package ?irr? *Various coefficients of interrater reliability and agreement*.

Barney G Glaser and Anselm L Strauss. 2017. *Discovery of grounded theory: Strategies for qualitative research*. Routledge.

Pei-Yi Hao, Jung-Hsien Chiang, and Yi-Kun Tu. 2007. Hierarchically svm classification based on support vector clustering method and its application to document categorization. *Expert Systems with applications*, 33(3):627–635.

Nathan Hartmann, Erick Fonseca, Christopher Shulby, Marcos Treviso, Jéssica Silva, and Sandra Aluísio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In

Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology, pages 122–131, Uberlândia, Brazil. Sociedade Brasileira de Computação.

Hatebase. 2019. Hatebase. Available in https://www.hatebase.org/, accessed last time in February 2019.

Jigsaw. 2018. Toxic comment classification challenge identify and classify toxic online comments. Available in https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge, accessed last time in 23 May 2018.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking Aggression Identification in Social Media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbulling (TRAC)*, Santa Fe, USA.

Yashar Mehdad and Joel Tetreault. 2016. Do characters abuse more than words? In *Proceedings of the SIGdial 2016 Conference: The 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 299–303.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, pages 145–153. International World Wide Web Conferences Steering Committee.

Ji Ho Park and Pascale Fung. 2017. One-step and Two-step Classification for Abusive Language Detection on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rogers Prates de Pelle and Viviane P Moreira. 2017. Offensive comments in the brazilian web: a dataset and baseline results. In *6º Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2017)*, volume 6. SBC.

Georgios K Pitsilis, Heri Ramampiaro, and Helge Langseth. 2018. Detecting offensive language in tweets using deep learning. *arXiv preprint arXiv:1801.04433*.

Fabio Poletto, Marco Stranisci, Manuela Sanguinetti, Viviana Patti, and Cristina Bosco. 2017. Hate speech annotation: Analysis of an italian Twitter corpus. In *Ceur Workshop Proceedings*, volume 2006, pages 1–6. CEUR-WS.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Björn Ross, Michael Rist, Guillermo Carbonell, Ben Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, pages 6–9.

Joni Salminen, Hind Almerekhi, Milica Milenković, Soon-gyo Jung, Jisun An, Haewoon Kwak, and Bernard J Jansen. 2018. Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *Twelfth International AAAI Conference on Web and Social Media*.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An italian Twitter corpus of hate speech against immigrants. In *Proceedings of LREC*.

William A Schabas. 2000. Hate speech in rwanda: The road to genocide. *McGill Law Journal*, 46:141.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. *SocialNLP 2017*, page 1.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26. Association for Computational Linguistics.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the 1st Workshop on Natural Language Processing and Computational Social Science*, pages 138–142.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of NAACL-HLT*, pages 88–93.

.

A Appendices

A.1 Non-expert annotators guidelines translated to English

Analyse the tweets from the first set and evaluate if according* to your opinion, these tweets contain hate speech.

For every tweet, mark manually with 1 or 0 if you think the tweet contains or not hate, respectively, accordingly with Table 5.

Tweet	HS	A
Black people should go back to their land!!	1	A
Meat and black beans are delicious!	0	A
Muslim people are terrorists!	1	A

Table 5: Hate speech (HS) annotation examples with respective annotator (A) in English.