Offensive Comments in the Brazilian Web: a dataset and baseline results

Rogers Prates de Pelle, Viviane P. Moreira

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

{rppelle,viviane}@inf.ufrgs.br

Abstract. *Brazilian Web users are among the most active in social networks and very keen on interacting with others. Offensive comments, known as* hate speech*, have been plaguing online media and originating a number of lawsuits against companies which publish Web content. Given the massive number of user generated text published on a daily basis, manually filtering offensive comments becomes infeasible. The identification of offensive comments can be treated as a supervised classification task. In order to obtain a model to classify comments, an annotated dataset containing positive and negative examples is necessary. The lack of such a dataset in Portuguese, limits the development of detection approaches for this language. In this paper, we describe how we created annotated datasets of offensive comments for Portuguese by collecting news comments on the Brazilian Web. In addition, we provide classification results achieved by standard classification algorithms on these datasets which can serve as baseline for future work on this topic.*

1. Introduction

Hate speech is defined as "any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion or other characteristic" [Nockleby, 2000]. Recently, companies like Facebook, Twitter, and YouTube have been facing legal action for allowing users to post texts that are considered offensive[1]. Since there is a massive number of posts every day, there is a need for automatically identifying and filtering such posts.

In Brazil, there have been prominent cases of racist texts posted in social networks targeting black celebrities[2][3]. Reports say that Brazilians spend the most time on social media and estimate that 96% of all Brazilian Internet users have at least one social network account [Webcertain, 2015]. Facebook alone has over 100 million Brazilian accounts[4]. The country ranks at third place in number of Facebook users and fifth in the number of Twitter users[5].

Hate speech is not limited to social networks. In a preliminary analysis, we took a sample of 145 news published on a single day (7-Jun-2016) by the biggest news site in the

[1] http://www.bbc.com/news/technology-36301772

[2] http://glo.bo/1Hl0HaB

[3] http://glo.bo/1UiV3i0

[4] http://www.internetworldstats.com/south.htm

[5] http://www.forbes.com/sites/victorlipman/2014/05/24/top-twitter-trends-what-countries-are-most-active-whos-most-popular

country[6] and identified that 90% had at least one hateful comment. In most cases, users start discussing the news and end up engaging in arguments using abusive language. Even though companies have mechanisms for preventing the publication of offensive texts, they still are not able to prevent cases such as the aforementioned ones.

In Brazil, the company Agência Nova/SB[7] carried out a study on intolerance in social networks. During three months, they monitored Facebook, Twitter, Instagram, and a number of blogs and websites. They recorded every time a subject such as racism and homophobia was mentioned, collecting a total of 542,781 mentions. Their findings, reported in [NoavaS/B, 2016], state that 84% of the comments on these subjects are negative.

Identifying offensive comments is not easy because authors tend to disguise offensive words by inserting asteriscs, spaces, or replacing characters by others with similar sounds. Thus, simply checking for the presence of terms that are in a precomputed list of offenses (*i.e.,* a blacklist) would miss many such comments. A better solution would be modeling this task as a text classification problem and generate models that would learn automatically how to identify offensive comments. This approach requires an annotated dataset with positive and negative examples (*i.e.,* offensive and non-offensive comments).

A number of recent works addressed the detection of hate speech on the Web focusing on different aspects, such as identifying racist tweets [Kwok and Wang, 2013, Silva et al., 2016] or blogs [Chau and Xu, 2007, Warner and Hirschberg, 2012], filtering pages with hate and violence [Liu and Forss, 2015], detecting hate groups [Ting et al., 2013], identifying flames [Razavi et al., 2010], and offensive news comments [Sood et al., 2012, Djuric et al., 2015]. To the best of our knowledge, none of such works have addressed the Brazilian Web or Portuguese texts.

The contributions of this work are: (i) an annotated dataset containing offensive (and non-offensive) comments collected from the Brazilian Web. We call it OFFCOMBR and make it available to the research community; (ii) baseline results of standard classification algorithms applied to the dataset, which may serve as reference for future works on this topic; and (iii) as a byproduct of the annotation process, we made available a system to help judges annotate sentences. We believe our contributions represent an initial step to enable the development of the identification of hate speech in Portuguese.

This paper is organized as follows. Section 2 presents the related literature on hate speech identification. Section 3 describes the process used to create OFFCOMBR and presents statistics on the dataset. Section 4 reports on experimental results on the datasets. Section 5 discusses our main findings; and Section 6 concludes the paper.

2. Related Work

The high volumes of hate speech on the Web has motivated the research community to come up with approaches to identify such offenses. Recent work have addressed different platforms such as web pages, social networks, blogs, and tweets.

Supervised classifiers have been used in a number of approaches for hate speech detection in different platforms: Xiang et al. [2012] applied them to a dataset of tweets looking for profane language; Kwok and Wang [2013] also worked with tweets but to

[6] www.g1.globo.com
[7] http://www.novasb.com.br

detect racist posts; Razavi et al. [2010] used them for flame detection on newsgroup messages; Warner and Hirschberg [2012] tackled anti-semitic comments from Yahoo! groups; and, while Sood et al. [2012] focused on detecting profanity, Djuric et al. [2015] addressed the broad category of hate speech in news comments.

Works that created annotated datasets. Supervised learning depends on the existence of annotated datasets containing instances with and without hate speech. A few studies report having created annotated datasets using human judges. Sood et al. [2012] used crowdsourcing to annotate 6,500 news comments. A total of 221 annotators identified 9,4% as containing profanity with at least 66% consensus. Warner and Hirschberg [2012] had three judges label 1,000 paragraphs extracted from web pages that might contain anti-semitic contents. Djuric et al. [2015] reports working with over 800K comments extracted from the Yahoo! Finance website, but the annotation process is not clear. Nobata et al. [2016] assembled a corpus with over 2 million comments from Yahoo! Finance and News. The annotations were carried out by trained employees. Chen et al. [2012] asked judges to assess whether 249 YouTube users were being offensive in their posts. Unfortunately, we found that *none* of the aforementioned datasets are readily available.

Available Datasets. Despite the research on hate speech detection dating back to 20 years, only in 2016 did the first datasets become publicly available. Ross et al. [2016][8] has 470 tweets in German classified as to whether they contain hate speech against refugees. Wulczyn et al. [2016][9] contains 115,737 Wikipedia discussion comments in English. The annotations indicate whether the comment has a personal attack. Waseem [2016][10] has 6,909 tweets in English annotated for hate speech through a crowdsourcing effort. We also found a dataset with about 4K English tweets classified as offensive and non-offensive on the Kaggle website[11].

None of the datasets available are in Portuguese. Thus, the dataset we describe in the next Section is, to the best of our knowledge, the first initiative in this direction.

3. Dataset Creation

In this section, we detail how the dataset was collected, the annotation process, and the statistics of our datasets.

3.1. Data collection

The source of the data was the news site g1.globo.com. This is the most accessed news site in Brazil[12] and, as a result, it has many comments. Although the comments on this site go through moderation, we found a considerable number of offensive contents. About 90% of the news we analyzed had at least one offensive comment. After a preliminary analysis, we noticed that the news categories with the most offensive comments are politics and sports. Thus, our data collection was limited to those sections.

We implemented a webscraper which sends requests to the web site on the sections of interest. The HTML pages of the news are then downloaded and parsed to extract the

[8] https://github.com/UCSM-DUE/IWG_hatespeech_public

[9] https://figshare.com/articles/Wikipedia_Detox_Data/4054689

[10] https://github.com/zeerakw/hatespeech

[11] https://www.kaggle.com/c/detecting-insults-in-social-commentary/data

[12] http://www.alexa.com/topsites/countries/BR

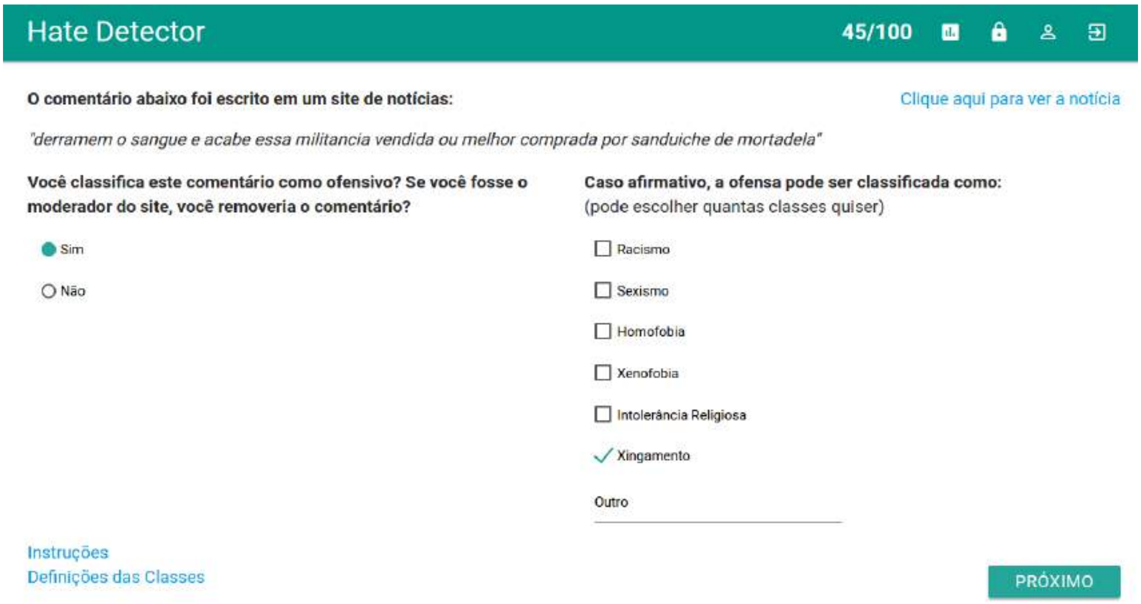
text of the news and the link where all comments for a given news can be obtained in the JSON format. All comments for a given news were extracted. Comments that were composed only of emojis or other non-alphabetical characters were discarded. To help preserve the anonymity of the authors, whenever a comment mentioned the full name of the author of another comment, we kept only the first name. At the end of this process, we obtained 10,336 comments posted for 115 news.

3.2. Annotation Process

Since our bottleneck for creating the dataset is the availability o human judges, we could not label all 10K instances. Thus, a sample of 1,250 comments was randomly selected. Following the standard procedure adopted for dataset annotation, each comment was annotated by three judges which were asked to whether it was offensive. In case of an affirmative answer, the annotator was also asked to categorize the offence as racism, sexism, homophobia, xenophobia, religious intolerance, or cursing.

We developed a tool to help the judges during the annotation process. The Web interface (depicted in Figure 1) showed the comment, the categories, and a link to the news for which the comment was written. A help screen with definitions of the categories and a definition of offensive text were also provided. We believe this tool could be useful to other annotation tasks, so we made it available at `http://inf.ufrgs.br/~rppelle/hatedetector/`.

Two datasets were generated from the annotations. The first, called OffComBr-2, has all 1,250 instances and the class assigned to each comment was the one picked by *at least two* of the judges. The second is a more strict dataset, called OffComBr-3. This dataset was composed solely of the comments for which *all three judges agreed* as to whether or not the comment was offensive. The datasets can be obtained at `http://inf.ufrgs.br/~rppelle/hatedetector/`.



O comentário abaixo foi escrito em um site de notícias:

"derramem o sangue e acabe essa militancia vendida ou melhor comprada por sanduiche de mortadela"

Você classifica este comentário como ofensivo? Se você fosse o moderador do site, você removeria o comentário?

● Sim

○ Não

Caso afirmativo, a ofensa pode ser classificada como:
(pode escolher quantas classes quiser)

□ Racismo

□ Sexismo

□ Homofobia

□ Xenofobia

□ Intolerância Religiosa

✓ Xingamento

Outro

Figure 1. Interface of the annotation tool

3.3. Statistics for the datasets

To measure the level of agreement among the judges, we calculated the Fleiss Kappa Fleiss [1971] measure which quantifies degree of agreement over that which would be expected by chance. For OFFCOMBR-2, the value was 0.71, which is considered substantial. This value is within the range of agreement found in other works that also carried out annotations (0.63 for Warner and Hirschberg [2012], 0.73 for Chen et al. [2012], and 0.84 for Nobata et al. [2016]). Since OFFCOMBR-3 only contains instances for which the class has been agreed by all three judges, it made no sense to calculate Kappa.

In OFFCOMBR-2, 419 (out of 1,250) comments were considered offensive by at least two judges, representing 32,5% of the total (we noticed that no comment was found offensive by only one judge)[13]. For OFFCOMBR-3, there are 202 offensive comments (out of 1,033), amounting to 19,5% of the cases. As with other datasets for hate speech detection, both versions of OFFCOMBR are unbalanced with negative examples outnumbering positive ones.

Regarding the categories (racism, sexism, homophobia, xenophobia, religious intolerance, or cursing), the results for each dataset are shown in Table 1. The most common category was by far cursing. The other categories had few comments. Religious intolerance was found in only one comment and it was not unanimous.

Table 1. Prevalence of each Category in the Annotations

# Judges	Xenophobia	Homophobia	Sexism	Racism	Cursing	Religious Intolerance
1	13 (1,0%)	35 (2.8%)	14 (1,1%)	19 (1.5%)	375 (30.0%)	1 (0.1%)
2	12 (1.0%)	14 (1,1%)	8 (0.6%)	18 (1.4%)	286 (22.9%)	1 (0.1%)
3	5 (0.5%)	9 (0.9%)	4 (0.4%)	1 (0.1%)	175 (16,9%)	0 (0.0%)

4. Implementing Classifiers to Identify Offensive Comments

In this section, we address the identification of offensive comments as a text classification problem. The task was then to take each comment perform a binary classification as to whether it is offensive (*i.e.,*, the classes were `yes` and `no`). Our goal here is to provide baseline results of standard classification algorithms and data preprocessing tasks on our datasets.

4.1. Experimental Setup

The instances in both datasets were submitted to a number of standard preprocessing tasks. These are identified below:

- *Case folding*. Two options were adopted: (i) converting the text of the comments to lowercase (`lower`) and (ii) leaving the comments in the case it was typed by the authors (`original`).

[13] Note that here we report on the percentage of comments that were found offensive, while in the Introduction we did a preliminary analysis of the percentage *news articles* with offensive comment. We found that, while 90% of the news articles had at least one offensive comment, out of all comments, 32,5% were considered offensive.

Table 2. Statistics of the two datasets

OffComBr-2		OffComBr-3	
File	**#features**	**file**	**#features**
original_1G	4,980	original_1G	4,348
original_1G_FS	241	original_1G_FS	119
original_1G+2G	17,374	original_1G+2G	15,085
original_1G+2G_FS	396	original_1G+2G_FS	164
original_1G+2G+3G	30,711	original_1G+2G+3G	26,600
original_1G+2G+3G_FS	448	original_1G+2G+3G_FS	182
lower_1G	4,123	lower_1G	3,647
lower_1G_FS	250	lower_1G_FS	122
lower_1G+2G	15,899	lower_1G+2G	12,385
lower_1G+2G_FS	426	lower_1G+2G_FS	103
lower_1G+2G+3G	29,126	lower_1G+2G+3G	25,303
lower_1G+2G+3G_FS	489	lower_1G+2G+3G_FS	196

- *Tokenization.* The text of the comments is tokenized and the tokens are used as features by the classification algorithms. This is known as bag-of-words approach. Three options of extracting n-gram (sequences of n tokens) features from the texts were tested: (i) using unigrams (`1G`), (ii) using unigrams and bigrams (`1G+2G`), (iii) using unigrams, bigrams and trigrams as features (`1G+2G+3G`). The idea is that by using longer n-grams, we could capture better the structure of the discourse.
- *Feature selection.* we compared the use of all n-grams as features against using only the features selected by Information Gain. Information Gain measures how correlated each feature is with respect to the class we wish to predict. We wanted to keep only features that had a positive correlation with the class, thus the threshold of zero was applied to the *InfoGainAttributeEval* method in Weka.

Combining all possibilities for case folding, tokenization, and feature selection across both datasets yielded 24 files to be processed by the classification algorithms. Statistics for these files are shown in Table 2.

We tested two algorithms that are widely used for text classification: Naive Bayes and SVM (called SMO in Weka's implementation). A 10-fold-cross-validation approach was followed and the results we report on the next section are the averages of the ten executions of each algorithm over the 24 data files.

4.2. Results

Our analysis of the results is based on the macro weighted average F-measure (*i.e.*,, the weight is given by the class size). For each class, the F-measure is the harmonic mean between recall and precision. In order to test whether the different performances were statistically significant, we used paired T-tests with the F-measures of the executions with the standard threshold of statistical significance of α = 0.05.

Classification results are shown in Figure 2. The columns reflect the average F-measure across all executions for a given parameter (*i.e.,* choices for preprocessing and

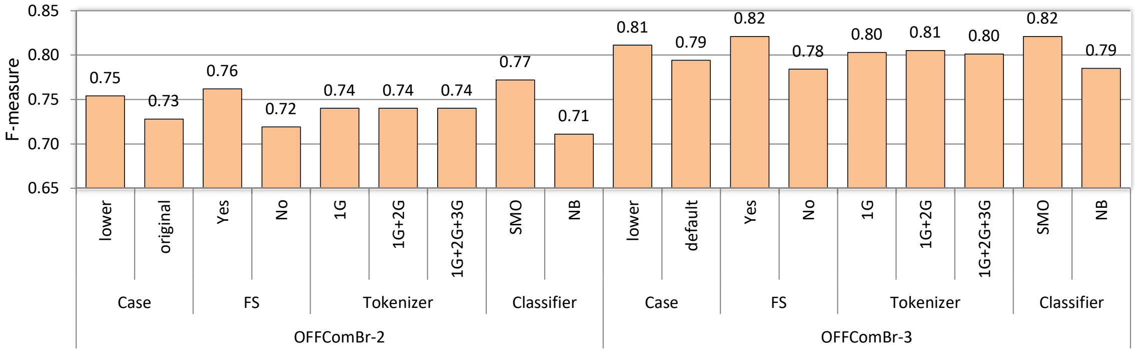


Figure 2. Classification results for different data preparations

classification algorithm). For example, the first column shows the average of the weighted F-measure in all executions in which the text of the comments was converted to lowercase considering both classification algorithms. Absolute values for F-measure ranged from .69 to .85. The configuration with the best result, in both OFFCOMBR-2 and OFFCOMBR-3, was achieved by SMO over the files in which the comments converted to lowercase and when feature selection was applied. The only difference was regarding the tokenization options – `1G+2G` was the best in OFFCOMBR-2 and `1G` was the best in OFFCOMBR-3, however, the differences were very small.

Regarding the results for each of the preprocessing tasks we found that:

- *Case folding*. Converting the text of the comments to lowercase brought significant improvements, with the F-measure for `lower` being statistically superior to the results for `original`. This suggests that the loss of information incurred by case folding is compensated by a gain of generality in the classification model.
- *Tokenization*. We found no statistical difference among the three tokenization options. Considering bigrams increases features by a factor of three, and considering trigrams increases the number further by a factor of two. Given the similar results that the three alternatives yield and the additional cost in processing this many more features, using unigrams only is preferable.
- *Feature selection*. Feature selection had a positive impact on the classifiers. Not only did the F-measure get significantly better but also processing time was drastically reduced by the use of much fewer features. The number of features was reduced to a maximum of 489 in OFFCOMBR-2 and 196 in OFFCOMBR-3. This represents a dramatic reduction as the selected features account for 0.8% to 4.8% of the total features.

With respect to the classifiers, SMO performed significantly better than Naive Bayes. The average F-measure for SMO was 0.80, while for Naive Bayes the average was 0.75. Besides the difference in the scores, an important difference between the two algorithms appears when we look at the misclassified instances. While SMO has more false negatives (*i.e.,* 188 offensive comments that were classified as non-offensive) and only a few false positives (*i.e.,* 15 non-offensive comments that were classified as offensive), Naive Bayes has fewer false negatives (154) and many more false positives (184).

Although these results cannot be directly compared to other works that used supervised classification algorithms for hate speech detection (since the datasets were different), we found that we have reached similar scores. Warner and Hirschberg [2012] and Sood et al. [2012] report achieving 0.63, while Chen et al. [2012] and Nobata et al. [2016] report scores of 0.8 and 0.83, respectively.

As expected, the classifiers performed better with OFFCOMBR-3 than with OFFCOMBR-2, since the difference between the yes and no classes are clearer in the former.

5. Discussion

In this Section we describe the main findings of this study and their implications.

Prevalence of cursing. Insults including vulgar language were the most frequent category of offensive comments. They were present in almost 70% of the comments found offensive. The targets of the insults were either people mentioned in the news (politicians or soccer players) or authors of other comments.

The importance of context. In many cases, we found that the same words present in offensive comments could also be used in non-offensive text. The distinction is the context in which they are used. For example, in all the words in the comments "*a minha mãe há anos esta no céu mas a sua tá na zona*" and "*voce nasceu assim ou bateu a cabeca quando era crianca*" are also very common in non-offensive comments. This means that one needs to know the context in which the words appear to be able to accurately classify them. In our experiments, we tested the use brigrams and trigrams as features hoping that they would provide more context and thus achieve better classification performance. Our results, however, did not improve with the use of longer n-grams. This suggests that other forms of providing context, such as considering sentence structure, should be investigated.

Language is constantly changing. The language on the Web is full of jargon, misspellings, and abbreviations. With the increase on the number of users, the evolution of the language becomes faster. New words are created and old words gain new meanings. This makes the use of static black-lists of offensive words inadequate. Also, if machine learning classifiers are to be used, they need to be capable of adapting and continue to learn new patterns.

Size of the comments. We observed a tendency that longer comments tended to contain fewer offenses. This suggests that longer comments require more careful elaboration and present fewer insults compared to shorter comments.

Freedom of speech. There is a fine balance between filtering offensive comments and interfering with people's freedom of speech. This calls for careful consideration in the implementation of automatic methods which should aim to minimize false positives.

Limitations. Given that the availability of human judges to annotate the instances is the bottleneck in the creation of datasets, OFFCOMBR is limited to 1,250 instances. We feel that many more comments are needed to allow for a better coverage of the offensive language identified on the Web. A larger dataset would also help classifiers distinguish between the categories. In addition, the low prevalence of comments with racism, sexism, homophobia, xenophobia, and religious intolerance makes OFFCOMBR not suitable for research that focuses on these categories. Furthermore, since we did not evaluate every comment for a given news article, we are unable to analyze whether articles with more comments tend to have more offensive comments.

6. Conclusion

Methods for hate speech detection that rely on supervised machine learning require annotated datasets. Although a number of studies have been conducted in this topic, there are only a handful of datasets available, most of them in English.

This paper addresses the lack of hate speech datasets in Portuguese by describing the creation of annotated datasets of offensive posts collected from news comments. The datasets, called OFFCOMBR-2 and OFFCOMBR-3, are freely available to the research community.

We have also run standard classification algorithms on the datasets to provide baseline results. The F-measure scores we obtained are within the range found in other work that addressed hate speech identification in English.

Out future work will include enlarging our datasets through a crowdsourcing effort using a larger number of volunteer annotators. In addition, we plan on adding an (anonymized) userId to each comment in the dataset. This will enable finding out the prevalence of offensive comments by user. This could be used as additional evidence by automatic methods for filtering offensive text.

Acknowledgements: This work has been partially supported by CNPq. R. P. de Pelle receives an MSc grant from CNPq. We thank the volunteers Cássio Alan Garcia, Diego de Vargas Feijó, Geisiane Martini, Lucas Pessutto, and Paula Burguêz for their effort in annotating our datasets.

References

Michael Chau and Jennifer Xu. Mining communities and their relationships in blogs: A study of online hate groups. *International Journal of Human-Computer Studies*, 65(1): 57–70, 2007.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. Detecting offensive language in social media to protect adolescent online safety. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 71–80, 2012.

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

J.L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

Irene Kwok and Yuzhou Wang. Locate the hate: Detecting tweets against blacks. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.

Shuhua Liu and Thomas Forss. Text classification models for web content filtering and online safety. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 961–968, 2015.

NoavaS/B. The Webcertain Global Search & Social Report, 2016. URL `http://www.comunicaquemuda.com.br/dossie/quando-intolerancia-chega-as-redes/`.

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

John T. Nockleby. Hate speech. In *Encyclopedia of the American Constitution (2nd ed.,edited by Leonard W. Levy, Kenneth L. Karst et al., New York: Macmillan, 2000)*, pages 1277–1279, 2000.

Amir H Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. Offensive language detection using multi-level classification. In *Advances in Artificial Intelligence*, pages 16–27. 2010.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 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*, volume 17 of *Bochumer Linguistische Arbeitsberichte*, pages 6–9, 2016.

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

Sara Owsley Sood, Judd Antin, and Elizabeth F Churchill. Using crowdsourcing to improve profanity detection. In *AAAI Spring Symposium: Wisdom of the Crowds*, volume 12, pages 69–74, 2012.

I Ting, Shyue-Liang Wang, Hsing-Miao Chi, Jyun-Sing Wu, et al. Content matters: A study of hate groups detection based on social networks analysis and web mining. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 1196–1201, 2013.

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

Zeerak Waseem. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, November 2016.

Webcertain. The Webcertain Global Search & Social Report, 2015. URL `http://internationaldigitalhub.com/en/publications/the-webcertain-global-search-and-social-report-2015`.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. *CoRR*, abs/1610.08914, 2016.

Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1980–1984, 2012.