Text Mining for Cyberbullying Detection: a Brazilian Portuguese Evaluation

Carolina Eberhart[1]**, Luciano Ignaczak**[1]**, Márcio Garcia Martins**[1]

[1] Universidade do Vale do Rio do Sinos (UNISINOS) – São Leopoldo, RS – Brasil

`carolieberhart@gmail.com, lignaczak@unisinos.br, marciog@unisinos.br`

Abstract. *Bullying and cyberbullying are words commonly seen in today's news. Although the scientific community has evaluated text mining techniques for cyberbullying detection, few studies have targeted Brazilian Portuguese datasets. Our study aims to assess the text mining application to detect cyberbullying messages written in Brazilian Portuguese. We gathered posts and comments from Reddit communities and extracted several text features. We then processed these features using Naïve Bayes and SVM classifiers to uncover cyberbullying activity. The outcomes of this experiment may not be used solo for cyberbullying detection; however, they can aid moderators in prioritizing content reviews and acting faster on real cyberbullying cases.*

Resumo. *Bullying e cyberbullying são assuntos abordados com frequência pela mídia. Embora a comunidade científica venha avaliando técnicas de mineração de texto para detecção de cyberbullying, poucos estudos utilizam datasets em português. Este estudo tem como objetivo avaliar a aplicação de mineração de texto para detectar mensagens em português associadas com cyberbullying. O estudo coletou posts e comentários de comunidades do site Reddit e extraiu diversas features, que foram usadas para treinar classificadores para descoberta de cyberbullying. Apesar dos resultados não demonstrarem que mineração de texto possa automatizar completamente a detecção de cyberbullying, as técnicas podem auxiliar moderadores na priorização da análise de mensagens.*

1. Introduction

Bullying and cyberbullying are not recent phenomenon. [Smith et al. 1999] defined bullying as ”a subcategory of aggressive behavior; but a particularly vicious kind of aggressive behavior, since it is directed, often repeatedly, towards a particular victim who is unable to defend himself or herself effectively”. [Hinduja and Patchin 2014] expanded this definition by stating that cyberbullying is an extension of bullying that includes harassment through electronic devices. [McCarthy 2018] gives a notion of the relevance of this problem when he mentions that, in 2018, 37% of Indian parents reported that their children experienced cyberbullying. Also, over 20% of the parents in the United States, Brazil, South Africa, and Canada reported the same issue. Still, cyberbullies target people of all ages and backgrounds.

The necessity of automated detection of cyberbullying messages on the web is unquestionable. Text mining provides a way of automating the message filtering process, and machine learning can offer support to the process. According to [Taeho 2019], text mining is the process of extracting knowledge from textual data. The usage of text mining to inhibit cyberbullying activities also shows in the scientific community worldwide. [Zhao and Mao 2016] study focuses on using text analysis to identify if English messages on Twitter are cyberbullying, even if those do not contain insults. [Nandhini and Sheeba 2015a]'s work focuses on analyzing English posts from Formspring and Myspace, categorizing those into bullying types such as harassment or racism. [Urtiga and Castro 2018] study uses data from Brazilian Portuguese Twitter messages to determine message topics that usually indicate bullying or cyberbullying activity. Studies utilized different text mining tasks and classification models to achieve their goals. However, a significant number of them incorporate an insulting words dictionary combined with another technique that assists in getting the context of the messages.

Our study aims to evaluate a text mining proposal for cyberbullying detection in messages written in Brazilian Portuguese. To achieve this, we collected posts and comments from specific Reddit communities and extracted information such as the text sentiment, key phrases, and the presence of insults. We then ran these features through two text classification algorithms: Naïve Bayes and SVM. Then, we also assessed how text mining could assist in finding the most toxic Reddit communities and users by ranking them based on our results. We want to contribute to the future state of automatic moderation of Brazilian Portuguese language online communities specifically. Although there is academic work on this subject, we only found a small number of studies focused on Brazilian Portuguese datasets.

This paper is organized as follows. In Section 2, we discuss recently published scientific papers that concern using text mining tasks in cyberbullying identification. Section 3 describes our research methodology, including the data source, chosen features, and the classifiers. In Section 4, we discuss the results obtained from the execution of our experiment and point out its contributions. Finally, Section 4 presents the conclusions and proposals for future work.

2. Related work

This section was based on using the search string "cyberbullying AND text mining" to find papers published on Association for Computing Machinery (ACM), Institute of Electrical and Electronics Engineers (IEEE Xplore), and Science Direct bases. The search string returned 72 papers. Then, we filtered the list by eliminating any papers published before 2015 or that had a Google H5-Index value lower than 15. The remaining ones were manually evaluated by their abstract contents, leaving a total of seven papers. Meanwhile, the second string "bullying AND mineração de dados" was used in Google Scholar to find related works specifically in Brazilian Portuguese. The strings resulted in 81 papers, which were filtered using the same published year and Google H5-Index criteria. Still, as Google H5-Index does not cover some of the smaller Brazilian journals, we also filtered the results by eliminating any papers with a Qualis "C" evaluation. Then, after reading the remaining results' abstracts, only one paper was deemed similar to our work. Therefore, this section presents the eight related works.

Every paper selected conveys a variation of cyberbullying identification techniques, providing a text-processing methodology combined with classification or clusterization tasks. These papers' general focus is building a classification or clusterization model for either binary message classification, non-binary message classification, or cyberbullying risk discovery.

The related papers under the binary message classification category focused on analyzing web content and evaluating whether it is cyberbullying or not. They all track the presence of insults in the messages; however, they do not consider this as the only significant feature to be analyzed. [Zhao et al. 2016] proposed evaluating the semantics and linguistic relationships of the words in the messages, as these also hold great importance in detecting cyberbullying. [Zhao and Mao 2016] also proposed using semantic and linguistic evaluation; however, their work focuses on creating a model to identify cyberbullying in messages that do not contain insults. [Singh et al. 2016] had a distinct approach to the problem. These authors' methodology was based on text and social features, such as the number of connections and their degree of centrality in the social network.

In the second category of papers - non-binary message classification - cyberbullying messages were classified into different cyberbullying categories (e.g., flaming, harassment, racism). To achieve this, [Nandhini and Sheeba 2015a] used the presence of insulting words, word frequencies, and part-of-speech as features. In another paper, [Nandhini and Sheeba 2015b] used fuzzy rules and a genetic algorithm to create their model, which was also based on part-of-speech and word frequency features.

The papers categorized as cyberbullying risk discovery aimed to identify words and expressions that indicate the presence or the risk of future cyberbullying activity occurrences and the topics associated with these cyberbullying messages. [Song and Song 2020] used features such as words related to cyberbullying methods and causes and the overall message sentiment. The final classification presents the level of cyberbullying risk associated with each message. Meanwhile, [Song et al. 2020] classified the messages into predetermined risk categories, using features such as term and document frequency, degree of diffusion, and degree of visibility. Finally, the study of [Urtiga and Castro 2018] was based on messages that already portray the cyberbullying theme but are not necessarily cyberbullying. They perform text clustering utilizing word frequencies as a feature.

Like most of the aforementioned papers, this work evaluated the automatic detection of cyberbullying activity. As in seven of the analyzed papers, we use text classification algorithms to detect cyberbullying. Furthermore, like [Singh et al. 2016], we utilize multiple classification algorithms and compare their results. Also, our feature selection focuses on sentiment analysis - as in [Song and Song 2020] - and the detection of insulting words - as in [Zhao et al. 2016, Zhao and Mao 2016, Nandhini and Sheeba 2015a]. The main distinction is that we search for the insults in the whole documents and also on their key phrases. We created this division to determine if the insults are a crucial part of the document or not. Finally, our dataset for this study was in Brazilian Portuguese.

3. Research Methodology

Our evaluation for automatic identification of cyberbullying messages along with harmful communities and users explored the Reddit (https://www.reddit.com/) website using a group of Python scripts. We chose Python as a programming language because of its text mining and Reddit information extraction libraries. The following sections expose the details regarding this study's dataset source and language, feature selection methods, and text classification algorithms. The scripts created for this experiment, as well as the utilized data and obtained results, are available on our GitHub[1].

3.1. Data source and collection

Reddit is a very active website, and it hosts thousands[2] of communities in many languages, including Brazilian Portuguese, which is our specific target in this study. Since Reddit does not have a list of subreddits by language, we used the emportugues.org website (https://emportugues.org/), which contains a list of Portuguese language subreddits, to obtain the base of our data source. On April 11th, 2021, we collected the complete list of 1993 communities' data from the emportugues.org website. We then filtered the records which would not be valuable for our study: duplicated, non-Brazilian Portuguese, and less active - less than 1,000 members - communities. The final dataset obtained from this filtering process contained 133 communities.

Then, on April 12th, 2021, we collected the actual data used in our experiment - the community comments - by utilizing the PRAW[3] (Python Reddit API Wrapper) library to access the selected communities' top 10 topics from the past 12 months - April 13th, 2020 to April 12th, 2021. We then cleared out each comment's emojis before storing them in our database, along with the respective author and community name. Through this process, we were able to obtain a total of 30,634 comments.

3.2. Data preprocessing

To obtain an appropriate dataset for text mining, we performed a few preprocessing steps. We began by removing comments that were too small in length to provide valuable information. According to [Song et al. 2014], the problem of short text classification presents itself differently from the general text classification one. Short text tends to be sparse, containing only a few words, few features, and a low level of information that does not provide context. [Song et al. 2014] mentions 30 characters news titles as short text examples, so we used this number as the base for our experiment and eliminated comments with less than 30 characters.

We also observed the presence of blank, foreign language, and bot comments in our data. Since we are interested in identifying cyberbullying activity as well as toxic users and communities in Brazilian Portuguese, none of these comments would be valuable in our dataset. Therefore, we removed all of their instances from our data. Our final dataset contained a total of 19,272 comments, which we then manually labeled as cyberbullying or not cyberbullying.

3.3. Selected features

We selected 12 features for our experiment, some based on related work, some in the characteristics of cyberbullying, and others in our knowledge of the internet forum Reddit. We then categorized these features into three groups: document sentiment features, insult / explicit wording presence features, and URL presence features.

The first set of features - positive, neutral, and negative sentiment confidence of the overall document - is meant to identify the possibility of maliciousness in a comment.

[1] `https://github.com/ceberhart2611/Cyberbullying-article-2021`

[2] `https://www.redditinc.com/press`

[3] `https://praw.readthedocs.io/en/latest/`

We based this set of features on the fact that insults in a message may not always mean it is malicious. Therefore, the sentiment would assist in deciding the message tone and aid the classifier in a decision not exclusively based on certain words. We obtained these sentiment confidence features by utilizing the Microsoft Azure Text Analysis API sentiment analysis module[4]. This API module mines the text for positive, neutral, and negative sentiment clues and outputs a confidence score between zero and one for each sentiment. Thus, the sum of these positive, neutral, and negative sentiment confidences for a given text always equals one.

The second set of features - the number of insults and sexually explicit words in the document and its key phrases - adds the possibility of explicitly identifying aggression and sexual harassment. We based this set of features on the articles we found in our related work search - [Zhao et al. 2016]; [Singh et al. 2016]; [Nandhini and Sheeba 2015a]; [Choi et al. 2020]. However, in all of these articles, a single dictionary of insulting words is used. Thus, we implemented a different approach by assigning insults into four dictionaries according to their category, allowing a broader set of insults in our experiment and enabling the classifier to work with different aggression types separately.

We created three different dictionaries of insulting words and one sexually explicit language dictionary containing adjectives and nouns only. The first insulting words dictionary contains only swear words, the second general insults, and the third context-related insults. General insults are offensive but socially accepted as they do not present inappropriate terminology. ”Shit” and ”dumbfuck” are examples of swear words, while ”idiot” and ”stupid” are general insults. Meanwhile, context-related insulting words are the ones that not only do not work solo as insults but also carry generally unoffensive meanings. For instance, saying ”I have pigs on my farm” is not insulting; however, saying ”Jack is such a pig” carries an offensive tone. These dictionaries aim to find the general cases of aggression tied to cyberbullying activity while still allowing other features to provide context cues. For example, even though swear words are generally unacceptable, they still can be used without offensive intent.

The sexually explicit words dictionary has the specific goal of assisting in revealing instances of sexual harassment. Its contents include jargon for body parts, sexual positions, and sexual acts. However, even though anatomic words such as ”penis” or ”vagina” may also indicate sexual harassment, these were not included in the dictionary. Reddit hosts many communities where healthy discussions about sexuality and sexual relationships, and these terms may be used in those. The final set, including the three insulting words dictionaries and the sexually explicit terms, has 1,191 unique records. The complete dictionaries[5] and sexually explicit words[6] used are available on GitHub.

In general, these features are build by finding the number of matches between each dictionary and a text record, resulting in two features per dictionary - one for the

[4] https://docs.microsoft.com/en-us/azure/cognitive-services/text-analytics/how-tos/text-analytics-how-to-sentiment-analysis?tabs=version-3-1

[5] https://github.com/ceberhart2611/Cyberbullying-article-2021/blob/main/insult-dictionaries.csv

[6] https://github.com/ceberhart2611/Cyberbullying-article-2021/blob/main/sex-slang-dictionary.csv

comment's whole text and then one for this comment's key phrases. We extracted the key phrases of each document by utilizing the Microsoft Azure Text Analysis API's key phrases module[7], which outputs the main points of a given unstructured text document. The goal of using key phrases instead of only the complete document to obtain features is to check if being aggressive or offensive is a core part of a comment.

Finally, the third group is a single feature: the presence of URLs in a comment. This feature indicates if a URL is present in a comment by trying to find ”HTTP://”, ”HTTPS://”, or ”www.” within its text. While some websites generate direct URLs to their content using content IDs, Reddit creates links utilizing the name of its users and communities or the title of the posts. Thus, a user posting a URL for a community or a user profile with an insulting name, or a post with an insulting title, may get unintended matches in sexually explicit and insulting words features. The general goal of tracking the URL presence is to give a chance for our classifiers to weigh these specific cases differently. This feature is also a contribution of this article as we did not find it in our related work corpus.

3.4. Data partitioning and classifiers

As our experiment utilizes supervised classifiers, we split our data into two different training and testing sets. 90% training / 10% testing and 80% training / 20% testing. Still, we identified that our testing results might not be satisfactory as, even though we had a high percentage of training cases, our dataset is highly imbalanced. In a total of 19,272 records, only 17.6% of them are labeled as cyberbullying. To mitigate this data imbalance issue, we used the Synthetic Minority Oversampling Technique - also known as SMOTE - as proposed by [Chawla et al. 2002]. The application of this oversampling method resulted in balanced training datasets, which had 50% of cyberbullying records and 50% of non-cyberbullying records. The specific SMOTE code used in this experiment is part of version 0.8.0 of the imblearn[8] Python library.

To acquire a fair evaluation of the efficiency of our features, we also used two distinct classification algorithms: Naïve Bayes and SVM. The choice of Naïve Bayes and SVM algorithms was associated with their significant presence in the related work and their different approaches. While Naïve Bayes does not explore relationships between the features and assumes their independence, SVM explores the possible relationships between them. The goal of processing our feature set in different classifiers was to verify how their different approaches perform on our dataset. As discussing classifiers implementation and optimization was not part of our scope, we used the Gaussian Naïve Bayes and the linear SVM implementations available on SciKit-learn.

4. Results and discussion

In this section, we discuss the results obtained by applying our research methodology. Table 1 shows that, through our evaluation, we obtained accuracy as high as 81%, and in general, the variations of our performance metrics do not surpass three percentage points. This same table also shows that our highest values for accuracy and precision are

[7] https://docs.microsoft.com/en-us/azure/cognitive-services/text-analytics/how-tos/text-analytics-how-to-keyword-extraction

[8] https://imbalanced-learn.org/stable/

81% and 45%, which appear under both the Naïve Bayes 80/20 column and the SVM 80/20 column. Still, the highest recall - our primary metric - and F1-score are under the SVM 80/20 column, which portrays the results of processing our features using an SVM classifier with an 80% training, 20% testing data fold. Hence, we concluded that this approach, which we will call SVM8020 onwards for simplicity, is the best result of our experiment.

Tabela 1. Experiment results - performance metrics

	Naïve Bayes		SVM	
Metric / Distribution	**90/10**	**80/20**	**90/10**	**80/20**
Accuracy	0.80	**0.81**	0.79	**0.81**
Precision	0.43	**0.45**	0.42	**0.45**
Recall	0.53	0.54	0.53	**0.56**
F1-score	0.47	0.49	0.47	**0.50**

Since presenting the results related to the top 10 toxic users and Reddit communities for our four different approaches would result in an extensive set of tables, we opted to show only the ones related to SVM8020. This evaluation was the one that obtained the best recall, meaning that it captured most of the comments deemed cyberbullying correctly. We had, however, to perform some manipulation of the data presented in the testing records before we obtained our final ranking. Therefore, communities and users with less than 14 and five comments were eliminated from the ranking data to allow a fair evaluation of their behavior.

Table 2 shows the outcomes of this process, comparing user and community toxicity as obtained from the actual comment labels and the predicted ones. The results shared in the table reveal that SVM8020 pointed at least half of the most toxic users and communities correctly; however, it was utterly deficient in the ranking order. Given that SVM8020 achieved a recall value of 56%, we can state that this result is satisfactory. However, if the prediction can only find around half of the actual cyberbullying cases, it will not correctly order the top offending authors and communities.

In order to find the issues in our predictions, we had to explore the misclassified records. When looking into the false negatives, we were able to identify cases in which the insults were not in our dictionary or were in the verbal form, which we did not explore. Commonly, the cases identified as cyberbullying in our dataset involved using some insults combined with a generally negative sentiment. Hence, cyberbullying cases in which there were no insulting words or that contained some insults, but had a highly neutral sentiment, were also not caught. Furthermore, we observed cases in which insults were combined with other words - for instance, ”shitface” - which made them go unnoticed as we searched for exact matches. Finally, we found that internet slang and instances of sarcasm also impacted our predictions and generated false negatives.

Meanwhile, in the false positives group, we found many insulting comments unrelated to people or groups, but rather companies, products, or the comment author himself. We also found instances of insulting words being used to intensify the effect of praising something, for instance, saying, ”fucking amazing”. Finally, the last common false-positive cases are tied to phrases with a high percentage of negative sentiment con-

Tabela 2. Most toxic users and communities in percentage according to the actual comments labels and the SVM8020 label predictions

	Actual label		**Predicted label**	
Rank	**Toxic users**	**Toxic communities**	**Toxic users**	**Toxic communities**
1	user560*	comm23*	user951*	comm5
2	user1776	comm107*	user560*	comm40*
3	user1302*	comm40*	user2722	comm51
4	user2528*	comm84*	user1302*	comm107*
5	user411	comm43	user2528*	comm53
6	user951*	comm86	user433	comm23*
7	user1747	comm45	user1094	comm104
8	user388	comm35*	user84	comm41
9	user1626	comm26	user839	comm35*
1	user1962	comm13	user2207	comm84*

* Users or communities ranked in the top ten for toxicity both by the actual labels and the predicted labels

fidence that contain only context-related insults. In these cases, the sentiment is generally sadness, not maliciousness; however, our sentiment analysis does not distinguish types of negativity.

Given all the performance metrics and the analysis we shared about our experiment, it is fair to say that the overall results obtained are satisfactory and may be applied to real-world scenarios of cyberbullying identification. Even though 56% of recall would not allow our automatic cyberbullying identification approach to work solo on Reddit, it could help the website moderation team to prioritize the investigation of specific comments and act faster when it comes to real cyberbullying cases. However, we would not recommend using the most toxic users and communities raking, as the ordering piece could misguide the exploration of the source of the most harmful content on the website.

Conclusion

In this article, we evaluated text mining for the identification of cyberbullying messages written in Brazilian Portuguese. Our evaluation comprised the presence of insults and the sentiment of text comments in Reddit communities. We verified the performance by running our dataset through two different text classifiers and then measuring it using accuracy, precision, recall, and F1-score metrics. The results of our experiment were overall satisfactory, yielding the possibility of its application in real-world scenarios. Therefore, we determined that text mining is helpful for the automatic moderation of Brazilian Portuguese online communities. In conclusion, this work sets a starting point for exploring machine learning-based online moderation in the Brazilian Portuguese language.

As the next step for future research, we suggest adding entity recognition and insulting verbs to our features to improve the recall percentage in the experiment. We also suggest considering moving the sexual harassment identification to another research more specifically tied to this subject and involving image processing techniques. Even though sexual harassment also pertains to the cyberbullying subject, it contains particularities that could benefit from joining text mining and image analysis.

Referências

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Choi, Y.-J., Jeon, B.-J., and Kim, H.-W. (2020). Identification of key cyberbullies: A text mining and social network analysis approach. *Telematics and Informatics*, page 101504.

Hinduja, S. and Patchin, J. W. (2014). *Bullying beyond the schoolyard: Preventing and responding to cyberbullying*. Corwin Press.

McCarthy, N. (2018). Where cyberbullying is most prevalent. Statista, 2018. Available at: <https://www.statista.com/chart/15926/the-share-of-parents-who-say-their-child-has-experienced-cyberbullying/>. Acessed in: November 24, 2020.

Nandhini, B. S. and Sheeba, J. (2015a). Cyberbullying detection and classification using information retrieval algorithm. In *Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology*, pages 1–5.

Nandhini, B. S. and Sheeba, J. (2015b). Online social network bullying detection using intelligence techniques. *Procedia Computer Science*, 45:485–492.

Singh, V. K., Huang, Q., and Atrey, P. K. (2016). Cyberbullying detection using probabilistic socio-textual information fusion. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 884–887. IEEE.

Smith, P. K., Catalano, R., Junger-Tas, J., Slee, P., Morita, Y., and Olweus, D. (1999). *The nature of school bullying: A cross-national perspective*. Psychology Press.

Song, G., Ye, Y., Du, X., Huang, X., and Bie, S. (2014). Short text classification: A survey. *Journal of multimedia*, 9(5):635.

Song, J., Han, Y., Kim, K., and Song, T. M. (2020). Social big data analysis of future signals for bullying in south korea: Application of general strain theory. *Telematics and Informatics*, 54:101472.

Song, T.-M. and Song, J. (2020). Prediction of risk factors of cyberbullying-related words in korea: Application of data mining using social big data. *Telematics and Informatics*.

Taeho, J. (2019). Text mining concepts, implementation, and big data challange,(p. 1). *Seoul, Korea: Hongik University*.

Urtiga, T. and Castro, T. (2018). Detecção de bullying escolar em redes sociais e suas implicações na educação de adolescentes. In *Brazilian Symposium on Computers in Education (SBIE)*, volume 29, page 1693.

Zhao, R. and Mao, K. (2016). Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder. *IEEE Transactions on Affective Computing*, 8(3):328–339.

Zhao, R., Zhou, A., and Mao, K. (2016). Automatic detection of cyberbullying on social networks based on bullying features. In *Proceedings of the 17th international conference on distributed computing and networking*, pages 1–6.