Detection of Hate Speech on Twitter using a Brazilian Portuguese Dataset

Mirella Gomes Silva Nascimento
Departamento de Ciência da Computação
Universidade de Brasília
Brasilia, Brazil

***Abstract*—Social media have become a place for propagation of hate speech and toxic behavior. Thus, is important to develop tools that can identify this online content in order to analyze its social consequences. Recently, many researches have addressed the subject, however most of the publish content was done in English. With that in mind, this paper proposes a convolutional neural netwok using word embedding to classify hate speech with a Brazilian Portuguese dataset. The model tested archived 74% F-score.**

***Index Terms*—hate-speech, word2vec, CNN**

I. Introduction

The increase in the use of social networks and the arrival of new platforms enabled the creation of spaces for discussion and expression of opinions of the most diverse types and, consequently, the propagation of hate speech. Hate speech is characterized by the manifestation of thoughts, values and ideologies that aim to inferiorize, discredit and humiliate a person or a social group, due to characteristics such as race, sexual orientation, religious affiliation, place of origin or class [1]. In 2020, between the months of July and December, Twitter penalized 1,126,990 accounts for offensive behavior on the network[1]. Therefore, there is a need for digital platforms to identify and take action against these online behaviors. However, due to the large number of post, performing this filtering manually is not an effective method. Thus, automating the detection of hate speech is essential to create a safer online environment.

The automatic detection of hate speech on social networks is considered a classification task generally focused on supervised learning [2]. The task usually classifies content into non-hate or hateful speech, and in this case, the type of the content may be classified as well [3]. It can be divided into two categories: One based on a classical approach to feature engineering and one based on a deep learning paradigm [3]. Due to the characteristics of the content posted online, such as poorly written text, emoticons, hashtags, and the dependency on the context, both approaches come across numerous challenges when performing the task [4].

The goal of this paper is to determine if a tweet is classified as offensive or hateful based only on its text. It will be used the dataset generated by Leite et al. who proposed a new dataset with tweets posted in Brazilian Portuguese called ToLD-Br (Toxic Language Dataset for Brazilian Portuguese). The data was labeled as toxic or non-toxic, or in different types of toxicity.

This paper is organized according the following sections: Section 2 addresses related work regarding hate speech detection; Section 3 describes the method used on this specific task; Section 4 exposes the results obtained; Section 5 indicates the conclusion and future perspectives.

[1]https://time.com/6080324/twitter-hate-speech-penalties/

II. Related Work

There have been several research done into classifying media content as offensive or not, however, most of it was published in English, mostly due to the extended amount of annotated datasets available. Gamback and Sikdar [4] who used an architecture based on Convolution Neural Network to categorize tweets into four classes: racism, sexism, both and non-hate-speech. A feature embedding for all words was generated using word embedding, word2vec or random vectors, character n-grams or an association of both, that way four CNN models were trained. A max-pooling layer downsized the feature set and the softmax function determined the label for each tweet. The model with word2-vec without character n-gram achieved the best performance.

Focusing on the development of a Portuguese dataset, Fortuna et al [5] used an annotated dataset composed of 5,668 tweets that were categorized into 'hate' and 'no-hate'. After applying pre-processing techniques, feature extraction used pre-trained GloVe word embedding for Portuguese. For the binary classification, it was used LSTMs architecture obtained a state-of-the-art outcome.

De Pelle and Moreira [7] collected 10,336 comments and 1250 annotated from the most accessed news website in Brazil to create a new dataset of offensive comments in Brazilian Portuguese. Besides labeling the data as offensive or not, the data was also categorized on the type of the offense, such as racism, sexism, homophobia, xenophobia, religious intolerance, or cursing. For the feature selection, they used bag-of-words, and for the classification model Naive Bayes and SVM. The results showed that the SMV model performed better with the average F-measure equal to 0.80.

TABLE I
DISTRIBUTION OF TOXIC AND NON TOXIC TWEETS IN THE DATASET

Label	Number of tweets
Non toxic	11745
Toxic	9255

III. PROPOSED METHOD

A. Pre-processing

This section describe the preparation of the dataset and the pre-processing used in the tweets. The dataset provided by Leite et all classified the texts in the following categories: LGBT+phobia, obscene, insult, racism, misogyny, and/or xenophobia. Since this paper proposes a binary classification, it was considered that if the text was classified as toxic at least once, then it would be labeled that way.Therefore, the number of tweets for each class is seen below:

The dataset was split into training (80%), validation (10%) and testing (10%) sets. A simple pre-processing step was performed on each text. This includes: applying lowercase, removing stopwords, URLs and user id. Hashtags and rts were treated by removing only their symbols.

B. Feature Embedding

In order for the computer to process words, they must be in a numerical format. For this reason, a word embedding technique is applied to the dataset. In this work, it was used the Word2Vec algorithm, which applies an unsupervised learning technique to create a vector representation of each word in the text based on its context [8].There are two Word2Vec models, Continuous Bags of Words(CBOW) and Skip-gram(SG) model. This paper uses the Skip-gram model to generate the vector. The SG model predicts the context of words based on a word of interest [8].

C. CNN Model

Despite being mostly used for computer vision tasks, convolutional neural networks can be applied to NLP as well. The main idea of this architecture is to combine convolutional layers with pooling layers [9]. According to Yoav Goldberg [9], in the convolutional layer the widows of k words (filters) will be transformed into a d-dimensional vector that registers important properties of the words seen. Then, a pooling layer will combine this vectors into a d-dimension vector. Last, the vector will be fed to a fully connected layer that predicts the text classification.

As mentioned earlier, the feature embedding layer was generated with Word2Vec. The output was added to a drop-out layer with rate of 0.2. The drop-out layer is used to prevent overfitting in neural networks [8]. Three 1D convolutional layers with a window size of 2, 3 and 4 are added to the model. Each output is reduced by a MaxPooling layer and added to another drop-out layer, but with a rate of 0.5. For the calculation of the class distribution probability, a fully connected layer with 'sigmoid' activation function is applied. To train the model, a binary cross entropy function and the accuracy optimizer was used.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 50, 100)           2765400

 dropout (Dropout)           (None, 50, 100)           0

 conv1d (Conv1D)             (None, 49, 128)           25728

 max_pooling1d (MaxPooling1D  (None, 24, 128)          0
 )

 conv1d_1 (Conv1D)           (None, 22, 128)           49280

 max_pooling1d_1 (MaxPooling  (None, 11, 128)          0
 1D)

 conv1d_2 (Conv1D)           (None, 8, 128)            65664

 max_pooling1d_2 (MaxPooling  (None, 4, 128)           0
 1D)

 flatten (Flatten)           (None, 512)               0

 dropout_1 (Dropout)         (None, 512)               0

 dense (Dense)               (None, 2)                 1026

=================================================================
Total params: 2,907,098
Trainable params: 141,698
Non-trainable params: 2,765,400
_________________________________________________________________
```

Fig. 1. Architecture of CNN model.

D. Evaluation

The confusion matrix is utilized for the performance evaluations of the classification method and allows to visualize how to calculate each value. To understand the results of the model, metrics such as precision, recall and f1-score were analyzed.

TABLE II
CONFUSION MATRIX

	Positive prediction	Negative prediction
Positive values	2128	242
Negative Values	1200	630

Precision is represented as the ratio between corrected classified data and the total predicted data.

$$Precision = \frac{TP}{TP + FP}$$

Recall is defined by the proportion of false negatives predicted.

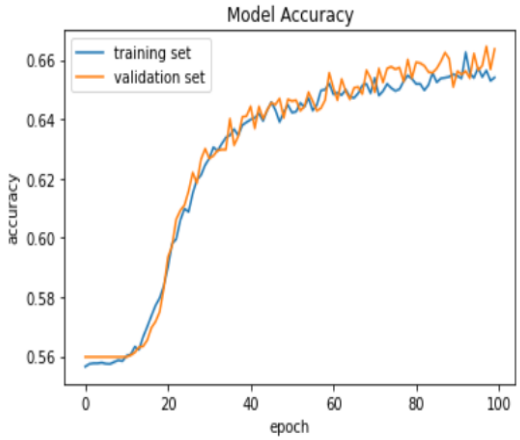
$$Recall = \frac{TF}{TN + FN}$$

F1 score combines precision and recall, determining the equilibrium between both metrics. It is the most used metrics to compare performance in the model and between different models

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

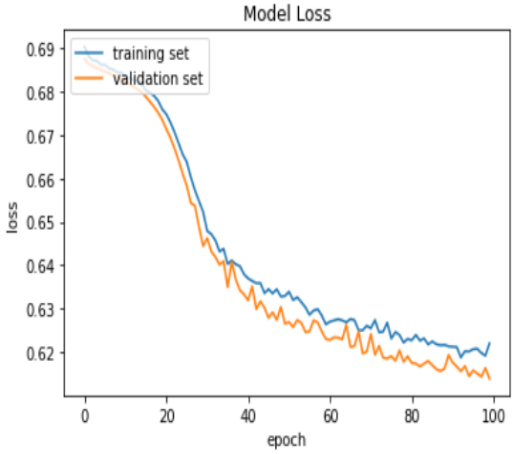
IV. Experimental Results

The classification report shows that the model obtained a 0.74% f1-score value for identifying non-toxic behavior. Also, by looking at both accuracy and loss graphs it can be perceived overfitting of the data.



0.5

Fig. 2. Training and validation accuracy



0.5

Fig. 3. Training and validation loss

V. Conclusion

This work proposes a deep-learning model to classify Brazilian Portuguese tweets in toxic or non-toxic. The classifier uses word2vec and a cnn for training and classification of the dataset available. The F1-score for the non-toxic behavior was 0.74, while for the toxic behavior was 0.5. The results indicate that, despite being a well research task, hate speech still has its challenges.

For future work it may be interesting to explorer with different neural networks such as LSTM and GRU for text classification, and also with different word embedding like BERT. Furthermore, the definition and the prediction of various types of hate speech may help understand and solve online problems.

References

[1] Trindade, L. V., Ribeiro, D. (2022). Discurso de ódio nas redes sociais. Brazil: Editora Jandaíra.

[2] A. Schmidt and M. Wiegand, A survey on hate speech detection using natural language processing, in: International Workshop on Natural Language Processing for Social Media, Association for Computational Linguistics, 2017, pp. 1–10.

[3] Zhang, Ziqi, and Lei Luo. "Hate speech detection: A solved problem? the challenging case of long tail on twitter." Semantic Web 10, no. 5 (2019): 925-945.

[4] Kovács, György, Pedro Alonso, and Rajkumar Saini. "Challenges of hate speech detection in social media." SN Computer Science 2, no. 2 (2021): 1-15.

[5] Gambäck, Björn, and Utpal Kumar Sikdar. "Using convolutional neural networks to classify hate-speech." In Proceedings of the first workshop on abusive language online, pp. 85-90. 2017.

[6] Fortuna, Paula, Joao Rocha da Silva, Leo Wanner, and Sérgio Nunes. "A hierarchically-labeled portuguese hate speech dataset." In Proceedings of the third workshop on abusive language online, pp. 94-104. 2019.

[7] de Pelle, Rogers Prates, and Viviane P. Moreira. "Offensive comments in the brazilian web: a dataset and baseline results." In Anais do VI Brazilian Workshop on Social Network Analysis and Mining. SBC, 2017.

[8] Hapke, Hannes, Cole Howard, and Hobson Lane. Natural Language Processing in Action: Understanding, analyzing, and generating text with Python. Simon and Schuster, 2019.

[9] Goldberg, Yoav. "A primer on neural network models for natural language processing." Journal of Artificial Intelligence Research 57 (2016): 345-420.