Investigating Deep Learning Approaches for Hate Speech Detection in Social Media

Portuguese-BR tweets

Leila Weitzel
Universidade Federal Fluminense (UFF): dept. Computer Science
Rio das Ostras, Brazil
leila_weitzel@id.uff.br

Thalessa Hungerbühler Daroz
Universidade Federal Fluminense (UFF): dept. Computer Science
Rio das Ostras, Brazil,
thalessahd@id.uff.br

Luan Pereira Cunha
Universidade Federal Fluminense (UFF): dept. Computer Science
Rio das Ostras, Brazil
luanpereiracunha@id.uff.br

Rafael Von Helde
Universidade Federal Fluminense (UFF): dept Computer Science
Rio das Ostras, Brazil
rafaelvon@id.uff.br

Lucas Mendonça de Morais
Universidade Federal Fluminense (UFF): dept. Computer Science
rio das Ostras, Brazil
lucasmmc@id.uff.br

Abstract — **Social media constitutes a very open social network space, which lacks of barriers to access. It has become a way for some people to vent anger, fear, happiness, hate, hope, love and sadness without feeling self-conscious. Hate speech and offensive language published and diffused via online environments have the potential to cause harm and suffering to individuals and lead to social disorder beyond virtual space. Over the last decade, machine learning and natural language processing approaches have been developed to detect harmful user content on social media. This work aims to analyze and detect hate speech, on Portuguese language, during the 2018 Brazilian presidential electioneering period. In order to achieve this goal, we experiment different configurations of Deep Learning networks, hyperparameter and features. As a result, the paper presents best settings of parameters for practical applications of hate speech mining. We also presented a Portuguese hate lexicon and a labeled dataset. All resources are available on our GitHub.**

Keywords - Hate Speech; Machine Learning; Word Embeddings; Deep Learning.

I. INTRODUCTION

Social media constitute a very open social network space, which lacks barriers to access. Microblogging websites such as Twitter and Sina Weibo (a Chinese microblogging website) have rapidly established as an emerging global communication service due to its timeliness and convenience as a lightweight and easy way of communication and sharing of information about everything. Beyond merely displaying news and reports, the Twitter itself is also a large platform where different opinions and emotions are presented [1].

In our context of internet ubiquity and a massive use of online social media, the new paradigm of communication is oriented towards sociability and socialization, centered on the social use of technology. Its nature transcends the idea of communicative mediation between two poles, sender and receiver, expanding this perspective to multiple poles, individuals, communities and society, constituting new forms of interaction with the communication processes. On one hand, the progress of interactions in digital environments serve as a favorable mechanism to projection information and knowledge of the human being. On the other hand, it is also fertile ground for expansion of conflicting aspects of tangible reality and social relationships, such as hatred and all its manifestations [2]. The social media has become a way for some people to vent anger, happiness, hate, love and sadness without feeling self-conscious. Indeed, offensive content including insulting, hurtful or derogatory are pervasive in social media. Everyone has the right to say what they think and can express their opinions or emotions about practically everything very spontaneously and without "censorship". The social media also opens the door to manipulation of masses and defamation of specific individuals or groups of people [3].

In Brazil, during 2018 presidential election a lot of controversial and even unacceptable user comments begun to appear. The campaign period was marred by political violence and the spread of election-related disinformation and hate/offensive content on social media and messaging platforms. The conversational tone online has become noticeably more aggressive in recent years, as a space in which hostility and hate speech flourish, most online hate speech is directed toward politicians and minorities and focuses on their race, religion, and/or sexual orientation [4].

Based on the context described herein, this paper aims to classify hate speech, on Portuguese (PT-BR) language. Hate/offensive language identification is classification task in Natural Language Processing (NLP). It has been an active area of research in both academia and industry for the past two decades. Lately, many platforms of social media websites monitor user posts, thus, this leads to a pressing demand for methods to automatically identify suspicious posts. The vast majority of the hate content in circulation in Brazil is propagated by individuals (or groups) with extreme far-right sympathies. These kinds of comments have been most openly hostile and hateful towards the LGBT community, black/brown, indigenous and female community. These

contents pose a huge challenge to societies, as they can lead to the radicalization of debates threatening democracy and the health of the population.

Societies need to develop adequate response mechanisms in order to find a balance between freedom of expression on one side and the ability to life without oppressive remarks without imposing rigid censorship regimes. Impacts of abusive language are detrimental, ranging from short-term emotional reactions (anger, fear, self-blame, etc.) to long-term psychological effects (low self-esteem, depression, etc.), causing mental and physical health issues (sleep disorder, headache, eating disorder, etc.) [5].

It must be stressed that the research focuses on Supervised Learning tasks (with binary classification). It is not the intention of the paper to delve into issues of linguistic theory or other Deep Learning approaches other than those discussed here.

II. Related work

Hate speech has several definitions. According to [6], it is defined as the usage of language to insult, it is referred as expressions that prompt to harm, discrimination, hostility and violence. Offensive language is related to pejoratives or profanities, obscenities, swear words and curses. Although numerous definitions of hate speech have emerged in recent years, our study focused on the expression of hate or degrading attitudes instead of offensive language such as swear or curse words.

Hate speech and offensive language mining are subareas of sentiment (emotion) mining. Sentiment Analysis is carried out in three ways, Machine Learning (ML), Deep Learning (DL), Lexicon-Based and the Hybrid approach. ML and DL approaches use a set of features, usually some function of the vocabulary frequency, which are learned from annotated corpora or labelled examples. The Lexicon-Based approach uses a lexicon to provide the polarity, or a semantic orientation, for each word or phrase in the text. The hybrid uses both paradigm, ML-DL and lexicon-based approaches [7].

To the best of our knowledge, a few papers address the classification of texts in Portuguese. This section shows some of these works.

Mondal, Silva and Benevenuto [8] provide a kind systematic large-scale measurement and analysis study of hate speech in online social media. The main goal was to understand the abundance of hate speech in online social media, the most common hate expressions, the effect of anonymity on hate speech and the most hated groups across regions.

Nascimento, Carvalho, Cunha, Viana and Guedes [9] proposed an approach for hate speech detection in Portuguese PT-BR. In order to achieve this goal, first they generate a dataset from offensive 55chan[1] boards that was used as a baseline; second, tweets were collected and afterwards, the LIWC-2015 Brazilian Portuguese Lexicon was used to filter both collected data. Three classifiers were trained to hate speech binary classification: Multinomial Naïve Bayes (MNB), Random Forest (RF) and Support Vector Machine (SVM).

Silva, Serapião and Paraboni [10] proposed the implementation of a Convolutional Neural Network (CNN) with pre-trained (Wang2vec and GloVe) and trainable word embeddings for detecting hate speech. They used LIWC-2015 to manual annotation and they used Logistic Regression (LR) as a baseline.

III. Material and Methods

In this work we used a dataset of tweets in Portuguese, collected through Twitter's Application Program Interface - API and Tweepy Python library during 2018-2019. Regarding the origin of the data the names of users or mentions have been omitted for privacy purposes.

A. Data collection and anotation

Various steps are being followed to retrieve data. We used hashtag and a set of key words to extract data. The dataset is saved in CSV file format, a total of 80k tweets were collected.

We collect the dataset by using default keywords. Examples of hate/offensive key words used were: ***bolsobosta*** (high hostility term), this is a junction of the first two syllables of the president's name with a word "dung". The term ***bolsominion*** (pejorative term), it is used to label individuals politically aligned with the ideals of a far-right candidate. This is a combination of the first two syllables of the president's name and the term Minion that comes from the cartoon, Despicable Me. The term ***petralha*** is a derogatory term for ardent supporters of the Brazilian Left-wing Workers' Party. Party loyalists are ***petistas*** ("P.T.-ists"), hence, "***petralha***" is formed by merging ***petista*** with metralha, a Portuguese slang word for machine gun ("metralhadora"); superficially, he was calling them "machine-gun leftists" it also references "Os Irmãos Metralha" (Machine Gun Brothers), the Portuguese translation of The Beagle Boys cartoon. The term, ***coxinha***, is a derogatory expression used by the Brazilian left for the uptight, politically conservative, and socially reactionary group in Brazil (right-wing). We also include anti-government and pro-government news accounts to extract additional tweets.

Thus, a lexicon named HALEX (HAte LExicon) with 512 terms of hate words (and terms) were created. We decide to remove profanity (swear words) terms from lexicon mostly because Brazilians have the habit of expressing themselves using swear words in both cases, either to express happiness or to express hatred, hence, we attempt to avoid the bias. The examples of hate terms are shown in Table I. It is important to note that some terms only make sense in the Portuguese language; the translation into English may seem meaningless.

[1] The 55chan was a Brazilian image board founded in October 2012 by hitmonkey and closed in January 2021. More information is avaible at: https://wikinet.pro/wiki/55chan_(2007%E2%80%9310)

TABLE I. EXAMPLE OF TERMS

terms	class
Brazil without commies	***hate***
bourgeois scoundrel	***hate***
president is a drunk and vagabond	***hate***
is more fake than	***hate***
out genocide	***hate***

B. Preprocessing approach

One of the main difficulties in processing a tweet, besides being an unstructured sentence, is that most of sentence may not use the correct grammar. Tweets contain many typographical errors, individual instances of typos, phonetic substitution and others that cause lexical deviation. This can be even worse if our goal is to identify hate speech [6].

Some individuals have the habit of masking hateful comments by inserting asterisks, spaces, or replacing characters with similar-sounding ones. For instance, the letter "a" is replaced by the character @. Other issues are related to the use of acronym slang and fashion words, acronyms, internet shorthand [6].

Apart from the plain text, tweets can have others elements such as hashtags; hyperlinks (typically a bitly URL, i.e., a URL shortening service), emoticons and references to other users, as "mention" (@<user>). In order to tackle these issues, we normalize tweets with the following procedures: first we remove: mentions; emoji, URL, special characters ($, %, &, # etc.), digits and punctuation (full stops, commas etc.) and stopwords. Then we delete duplicates tweets and tweets starting with "RT" because they refer to a previous tweet. We normalize abbreviations of: Political Parties, Brazilian States and Portuguese internet shorthand (also called internet slang), e.g. "pq" replaced by "porquê" ("why"), "aki" - "aqui" ("here"); "hj" - "hoje" ("today") and etc. Repeated characters were also deleted ("odeiooooo" –hateeee replaced by "odeio". We manual split hashtags e.g. "lulatapreso" by "lula está preso" – lula is in jail.

After the processing steps described herein, we remove tweets with less than three tokens. In the context of Twitter, it is common to see shortened messages e.g.: "que bixa safada !!" (what fags dirty !!) that correspond to a hate sentence but can also correspond to a joke (in Portuguese-BR). This tweet does not have a semantic interpretation for other people. It is basically understandable by the author (or the receiver) himself. It is common that the human annotator assigns it as hate speech, however, this supposition is based on beliefs of the annotator, and not on the real situation. Hence, afterward the complete processing step, we obtained 33,771 tweets.

A python routine was developed to perform the tweets automatic classification using the HALEX as a benchmark. Hence, we gathered 5,575 hate and 28,196 non-hate tweets. To assess the accuracy of the automatic annotation, we applied a manual inspection by taking a significant sample (using Z-score table) from the classified dataset. Four researchers were designated to perform this task. All of them are native Portuguese speakers, aged between 21-60, and everyone had at least a bachelor's degree qualification. We calculate the Cohen's Kappa metric to ensure that the inter annotator agreement were successful.

C. Imbalanced classifications

Imbalanced classification poses a challenge for predictive modeling. A severe imbalance is more challenging to model and require specialized techniques to tackle this issue [11].

Hence, we perform undersampling technique in order to attempt the balance the class frequencies. In this approach, we reduce the number of samples from the majority class to match the number of samples in the minority class. We build three datasets as follows (Table II). The approach aims to assess whether the processing has a significant impact on the evaluation of the classifiers' performance. These datasets were split into training, validation and test sets. In each epoch, the same training data is fed to the neural network architecture repeatedly, and the model continues to learn the features of the data. The test set is a separate set of data used to test the model after completing the training.

TABLE II. DATASETS

database	preprocessed	Stopword removal	tweets
B0	***no***	***no***	***12,705***
B1	***Yes - full***	***yes***	***11,150***
B2	***Yes***	***no***	***10,575***

D. Model selection and configuration

The fundamental data structure in neural networks is the layer, which are combined into a network. Others important objects are loss function and optimizers. An Example of typical neural network are: Input layer; (ii) Output layer; (iii) Activation function: [Softmax, ReLU, etc.].; (iv) Loss function: [Cross Entropy, Mean Squared Error, etc.]; (v) Optimizer: [Adam, Nadam. RMSProp (Root Mean Squared Propagation) and SGD (Stochastic Gradient Descent)] [12].

We investigated three deep neural network: CNN Convolutional Neural Network, LSTM – Long Short Term Memory, Bi-LSTM - Bidirectional LSTM. Two optimizers were tested, Nadam and RMSProp. Nadam is Adam with Nesterov momentum. An in-depth study of these subjects is available in [12].

We build two baseline: Shallow Net and Multinomial Naive Bayes (MNB). MNB is widely used in supervised machine learning model. It provides an ability to classify data, that cannot be represented numerically [13]. Shallow neural networks consist of only 1 or 2 hidden layers. Understanding a shallow neural network gives us an insight into what exactly is going on inside a deep neural network [12]. Having constructed a baseline, the next step is to see what the baseline fails to capture. This will guide our choice of a more complex model. The baselines help us understand our data better.

To systematically study which model and dataset features lead to a better generalization in hate language-related models, we run about 36 experiments. As an example, the Figure 1

shows the architecture of one of neural network created, the CNN-B2.

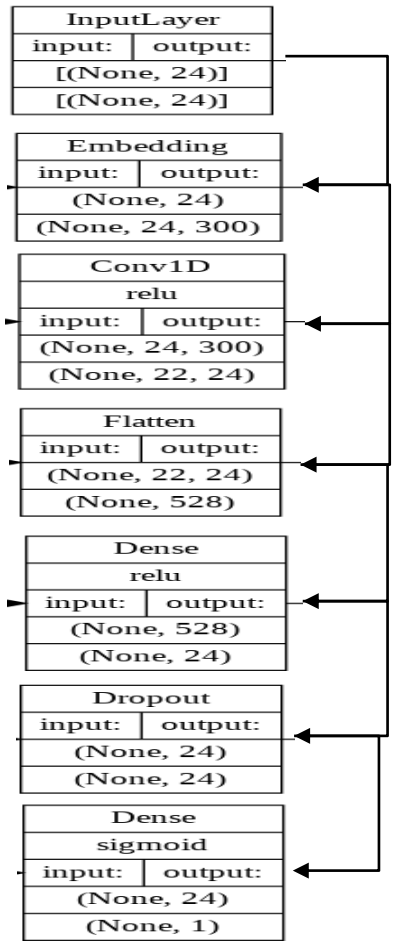

Figure 1. CNN-B2 architecture

E. hyperparameter tuning.

The hyperparameters tuning define how our model is actually structured. The process of setting the hyperparameters requires expertise and extensive Trial-Error. The Trial-Error is performed in a semi-automatic approach, where the hyperparameters are initialized with maximum and minimum values, and during training the performance was evaluated using a call-back (AUC and ROC-Curve). In order to find the optimal model architecture, we follow the steps as follow, we: define a model, define the range of possible values for all hyperparameters, define a method for sampling hyperparameter values, define evaluative criteria. The ranges tested were: batch size = [32, 64]; learn rate = [0.0001, 0.001]; optimizers: [Nadam, RMSProp, SGD]; Activation function: [ReLu (hidden layers), Sigmoid (output layer)].

To avoid overfitting, we use standard regularization techniques, e.g., early stopping (stops the training when a monitored quantity has stopped improving) and dropout. Dropout randomly select a neuron that will be ignored during training [14].

IV. RESULTS

The Table III shows the main outcomes achieved. One can be seen that there are not statistical differences between Bi-LSTM, CNN and LSTM except for LSTM-B0. This finding may show that the LSTM was affected by preprocessing step. As we expected, the ShallowNet achieved the worst performances. On the other hand, the MNB, surprisingly, showed a performance that can be considered acceptable (about ~88%).

TABLE III. : RESULT FOR HATE SPEECH DETECTION

Network	Dataset	Accuracy	Loss function
Bi-LSTM	***B0***	***0,94***	***0,20***
Bi-LSTM	***B1***	***0,94***	***0,17***
Bi-LSTM	***B2***	***0,96***	***0,14***
CNN	***B0***	***0,94***	***0,17***
CNN	***B1***	***0,95***	***0,16***
CNN	***B2***	***0,96***	***0,15***
LSTM	***B0***	***0,69***	***0,63***
LSTM	***B1***	***0,91***	***0,20***
LSTM	***B2***	***0,95***	***0,15***
MNB	***B0***	***0,84***	***n/a***
MNB	***B1***	***0,87***	***n/a***
MNB	***B2***	***0,91***	***n/a***
ShallowNet	***B0***	***0,60***	***98,94***
ShallowNet	***B1***	***0,62***	***32,95***
ShallowNet	***B2***	***0,55***	***44,0***

The figure 2 illustrates the ROC curve and AUC measure. Observing that they are practically the same values as the accuracy, this may validate the results.

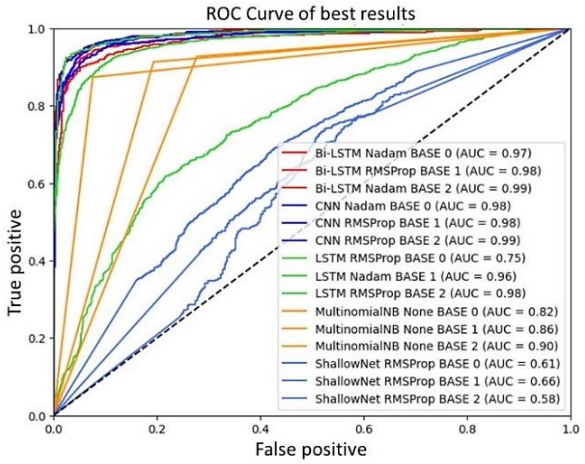

Figure 2. ROC and AUC values of the best results

The Figure 3 shows the Confusion Matrix of the CCN-B2 and the Figure 4 shows the Confusion Matrix of Bi-LSTM-B2, both in the test step. The zero label is not-hate label and label one is the hate label. These models achieved the best performance. The AUC measure was about 99% for both networks. The CNN-B2 network with Nadam optimizer, and Bi-LSTM network with RMSProp optimizer. However, we cannot ignore the performance that other networks reached (~ 98%.) such as, LSTM-B2 (RMSProp), CNN-B1 (RMSProp), CNN-B1 (RMSProp), CN-B0 (RMSProp), Bi-LSTM-B1 (RMSProp).

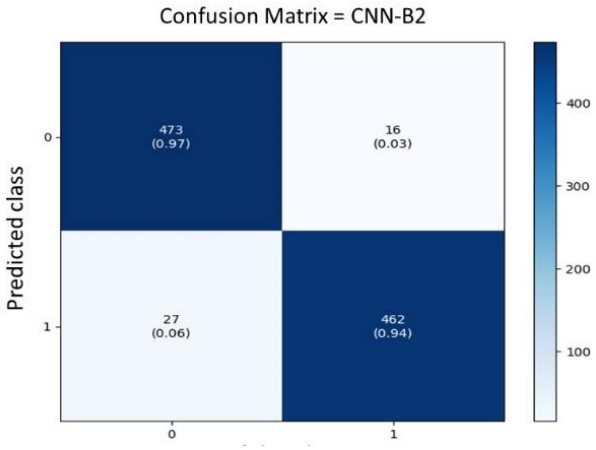

Figure 3. Confusion Matrix – CNN-B2

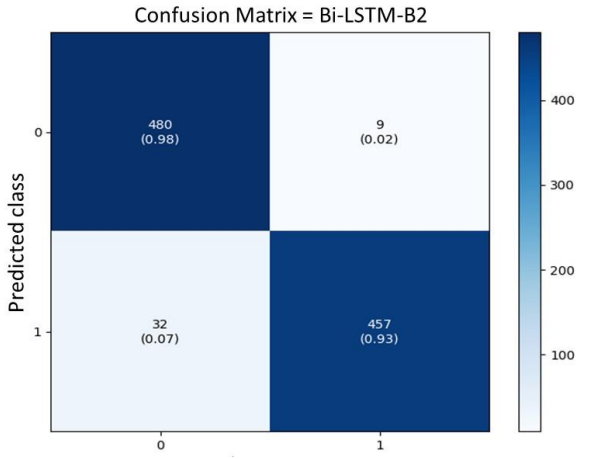

Figure 4. Confusion Matriz – Bi-LSTM-B2

V. CONCLUSIONS

The Internet became the platform for debates and expressions of personal opinions on various subjects. Social media have assumed an important role as a tool for interaction and communication between people. Users can express their opinions anonymously. To understand this phenomenon, it is indispensable to detect and assess what characterizes hate speech and how harmful it can be to society. In this paper we present a comprehensive evaluation of Portuguese-BR hate speech identification. The experiments demonstrate that the DL models tested can be used to detect hate speech.

We summarize our contribution in this paper: The performance acquired was basically the same, which means that intense processing was not necessary. We highlighted that these evidences are related to this dataset. We release an annotated hate speech dataset at sentence-level in Portuguese-BR language. We build a hate speech lexicon with 512 terms. The dataset, notebooks and dictionaries are available at https://github.com/LuanPCunha/TCC.

We believe that the research shed light on hate speech identification. Our future work we will extend this study to investigate different strategies. For example, we would like to extend the dataset using oversampling technique. There are several methods available to oversample. The most common data augmentation technique is Synthetic Minority named Oversampling Technique or SMOTE for short. Most automatic hate speech detection approaches tackle the problem as a binary classification thus the we aim to detect from multi-label perspective.

REFERENCES

[1] J. L. Alves, L. Weitzel, P. Quaresma, C. E. Cardoso, and L. Cunha, "Brazilian Presidential Elections in the Era of Misinformation: A Machine Learning Approach to Analyse Fake News," *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications.* pp. 72-84.

[2] R. C. M. Maia, and T. A. S. Rezende, "Respect and Disrespect in Deliberation Across the Networked Media Environment: Examining Multiple Paths of Political Talk," vol. 21, no. 2 %J J. Comp.-Med. Commun., pp. 121–139, 2016.

[3] B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1-167, 2012.

[4] M. Pérez-Escolar, and J. M. Noguera-Vivo, *Hate Speech and Polarization in Participatory Society*, 1 ed., London: Routledge, 2021.

[5] "Hate Speech on Social Media: Global Comparisons," *Council on Foreign Relations*.

[6] H. Watanabe, M. Bouazizi, and T. Ohtsuki, "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection," *IEEE Access*, vol. 6, pp. 13825-13835, 2018.

[7] B. Pang, and L. Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval*, vol. 2, no. 1-2, pp. 1-135, 2008.

[8] M. Mondal, L. A. j. Silva, and F. c. Benevenuto, "A measurement study of hate speech in social media." pp. 85-94.

[9] G. Nascimento, F. Carvalho, A. M. d. Cunha, C. R. Viana, and G. P. Guedes, "Hate speech detection using Brazilian imageboards." pp. 325-328.

[10] S. C. Silva, A. B. Serapião, and I. Paraboni, "Hate-speech detection in Portuguese using CNN and psycho-linguistic dictionary," *J. Inf. Data Manage.*, vol. 5, pp. 1-12, 2019.

[11] M. Kuhn, and K. Johnson, *Applied predictive modeling*, New York: Springer, 2013.
[12] F. Chollet, *Deep learning with Python*, Shelter Island, New York: Manning Publications Co, 2018.
[13] A. C. Müller, and S. Guido, *Introduction to machine learning with Python: a guide for data scientists*, First edition ed., Sebastopol, CA: O'Reilly Media, Inc, 2016.
[14] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929-1958, 2014, 2014.