Comparação de modelos para classificação de discurso de ódio

Gabriel Vieira
Programa de Engenharia de Sistemas e Computação
COPPE - UFRJ
Rio de Janeiro, Brasil
gabrielv@cos.ufrj.br

Leticia Freire
Programa de Engenharia de Sistemas e Computação
COPPE - UFRJ
Rio de Janeiro, Brasil
lfreire@cos.ufrj.br

Rennan Gaio
Programa de Engenharia de Sistemas e Computação
COPPE - UFRJ
Rio de Janeiro, Brasil
rennangaio@cos.ufrj.br

***Abstract*—Discursos de ódio cada vez estão sendo comuns em redes sociais. Humanos conseguem facilmente distinguir quando isto ocorre, mas computadores possuem dificuldades em distinguir quando um texto está sendo ofensivo ou não. Utilizando dois trabalhos anteriores como base, comparamos quatro algoritmos de classificação, que não são específicos para processamento de linguagem natual, com a intenção de verificar como estes modelos lidam com datasets textuais.**

***Index Terms*—discurso, ódio, classificação, texto, frases, twitter, comentários**

I. Introdução

Nas redes sociais muitas pessoas são atacadas e perpetuam discursos de ódio, o que traz uma certa toxicidade àquele ambiente onde muitas pessoas se relacionam. Estes discursos são facilmente descritos como sendo um discurso de ódio ou não por humanos, entretanto computadores possuem uma dificuldade em classificar um texto como sendo deste tipo, aumentando ainda mais a dificuldade por ser em língua portuguesa. Atualmente, este tipo de atividade vêm sendo muito relevante pelas novas formas de legislação que obrigam as plataformas a combater a proliferação de conteúdo extremista [1] . Logo, criar uma forma automatizada de realizar este procedimento possui muito valor.

Para isso, os trabalhos de pesquisa de Rogers [1] e Paula [2] foram criados com a finalidade de classificar um texto como sendo um discurso de ódio com frases em português. Eles possuem uma base anotada de diversas frases coletadas do twitter e de comentários no G1 classificadas como sendo discurso de ódio ou não.

A fim de se construir um modelo mais interessante, foi utilizado o pré-processamento de word embedding do NILC [2] para extrair os atributos das palavras nas sentenças. Em conjunto com estes dados, verificou-se uma série de algoritmos de classificação da literatura de aprendizado de máquina para testar suas respectivas performances.

O objetivo deste trabalho é reproduzir alguns dos experimentos já feitos pelos artigos supracitados, assim como buscar uma aplicação de teste com algumas modificações nas bases de dados fornecidas. Em relação aos classificadores, foram utilizados: O XGBoost; o SVM (Support Vector Machine); o Multinomial Naive Bayes e o Random Forest. Foi também explorada a composição dos dois conjuntos de dados para avaliar seu resultado final.

II. Trabalhos relacionados

Na elaboração deste trabalho, utilizou-se como parâmetro as motivações e soluções dos estudos de Rogers [1] e Paula [2]. Para estabelecer uma padronização de ambas as abordagens, este capítulo tem como objetivo introduzir as técnicas e soluções apresentados por estes.

Primeiramente, na dissertação de Rogers, foram elaborados dois datasets pioneiros na tarefa que classificação de discursos de ódio para o vocabulário PT-BR (o OFFCOMBR-2 e o OFFCOMBR-3) [3]. Sua motivação se deu pelo caráter dinâmico da linguagem na web e pela maior especificidade com o português do Brasil. Além disto, foi elaborado o método de classificação nomeado de "Hate2Vec", com o objetivo de detectar o conteúdo ofensivo em textos. Para a construção dele foi utilizado um conjunto de classificadores que se alimentam dos embeddings de palavras e documentos. Posteriormente é utilizado um meta-classificador que combina a saída dos modelos e os embeddings dos termos criados para realizar a classificação. Como resultado, utilizando F1-score pelo autor, foi obtido valores que variam entre 0,90 e 0,97 de desempenho.

Em relação ao artigo de Paula, este também é focado na detecção de discursos de ódio, porém sua base de dados provêm da rede social Twitter e o estudo é de Portugal. Dentre suas contribuições, foi criada uma nova topologia

[1] https://www.reuters.com/article/us-eu-hatespeech/social-media-companies-accelerate-removals-of-online-hate-speech-eu-idUSKBN1F806X

[2] http://nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc

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

para esta classificação, criando diversas subclasses em um problema multiclasse. Desta forma foi possível distinguir discursos racistas de discursos homofóbicos por exemplo, que não seriam possíveis em um problema de classificação binária. Entretanto, por conta da dificuldade de concordância na anotação das classes específicas, utilizou-se uma simplificação dessas anotações apenas classificando como um discurso de ódio ou não por voto de maioria. Como ferramenta de classificação, utilizou-se a extração de atributos com Embedding das palavras e foi utilizado LSTM para a classificação. Como resultados, foi obtido com a métrica F1-score um desempenho de 0.72 na amostra de teste.

III. Dataset

A fim de realizar a classificação de padrões de discursos de ódio, levou-se em consideração os dois conjuntos de dados dos trabalhos que basearam o desenvolvimento deste. A utilização de um dataset já estabelecido favorece tanto em questões comparativas de resultados do desenvolvimento como também poupa um grande trabalho de crawling que seria necessário nas plataformas de redes sociais.

O primeiro dataset, produzido por Paula et. al, foi construído com base na busca pela API do twitter e em perfis específicos desta mesma plataforma. Para detectar possíveis mensagens que pudessem ter conotações de discurso de ódio, foram selecionadas algumas palavras chave e algumas hashtags específicas como "sapatão" ou "#LugarDeMulherENaCozinha". Após fazer esta seleção inicial de mensagens, foram escolhidas no máximo 200 mensagens por perfil com finalidade de gerar uma maior diversidade entre as possíveis sentenças. Ao final deste processo, este conjunto de dados totaliza 2668 mensagens de 1156 usuários distintos.

O segundo, presente na dissertação de mestrado de Rogers, foi o primeiro dataset na temática de discursos de ódio mais específico do português do Brasil (A pesquisa supracitada utilizava-se do português tanto de Portugal quanto do Brasil e outros paises que falam portugues, visto que a API do Twitter não distingue os países mas sim a língua). Como fonte de dados foram escolhidos os comentários das notícias do site "g1.globo.com", uma página web de noticias. Por ser um dos maiores sites de notícia do Brasil, observou-se uma grande quantidade de material, tendo em vista que cerca de 90% das postagens possuía pelo menos 1 comentário ofensivo. Como resultado final, esta base possui 10336 comentários de 115 notícias distintas.

Para a realização dos estudos deste trabalho, foi escolhido fazer uma composição entre os dois conjuntos de dados. Desta forma são aproveitados dois aspectos positivos para os resultados encontrados. Primeiramente, tem-se uma maior diversidade de plataformas as quais estes textos estão inseridos. Subsequentemente, foi possível fazer um melhor balanceamento dos dados, que originalmente possuíam-se muito mais exemplos sem discurso de ódio, sem prejudicar a quantidade de amostras deste problema. Foram selecionados então a mesma quantidade de dados das duas classes distintas, totalizando 2207 exemplos de cada uma delas.

IV. Experimentos e Resultados

Para este trabalho, foram utilizados quatro algoritmos de classificação a fim de identificar frases contendo discurso de ódio. Foram eles:

- Random Forest: utiliza árvores de predição para realizar a tarefa [3];
- Multinomial Naive Baye: representa cada documento como um vetor de palavras , indicando quantas vezes a palavra aparece no documento, calculando a probabilidade do documento pertencer às classes por meio de uma distribuição multinomial sobre o conjunto de palavras [4];
- SVM com Stochastic Gradient Descent: tenta encontrar um plano que melhor separe os dados das duas classes [5] e
- XGBoost: árvore de decisão que utiliza o conceito de boosting no seu processo de treinamento [6].

Cada um dos classificadores foi executado 10 vezes com os parâmetros apresentados na Tabela 1. A bilioteca em python sklearn [7] foi utilizado para os três primeiros classificadores; para o último, foi utilizada a biblioteca xgboost [4].

TABLE I
Parâmetros para os modelos

Random Forest	n_estimators = 10
Multinomial Naive Bayes	parâmetros padrão da biblioteca sklearn
SVM com SGD	parâmetros padrão da biblioteca sklearn
XGBoost	parâmetros padrão da biblioteca xgboost

Em cada uma das execuções, o dataset foi dividido em 30% para teste e 70% para treino. Para avaliar cada um dos classificadores, foram usadas 3 métricas de avaliação: acurácia, precisão e F1. A acurácia (1) mede a porcentagem de dados classificados corretamente, a partir de todos os dados que foram classificados [8]. Sendo assim, esta métrica calcula a quantidade de verdadeiros positivos e verdadeiros negativos.

$$\frac{VP + VN}{VP + FP + VN + FN} \tag{1}$$

sendo VP, verdadeiro positivo; VN, verdadeiro negativo; FP, falso positivo e FN, falso negativo.

A precisão (2) avalia quantos dados foram corretamente classificados para uma classe (verdadeiros positivos) sobre todos os dados que foram classificados como sendo daquela classe [8].

$$\frac{VP}{VP + FP} \tag{2}$$

Já a métrica F1 (3) utiliza a precisão e o recall para avaliar o classificador, fazendo uma média entre eles. O recall informa a quantidade de dados que foram classificados corretamente pela classe verdadeira a partir de todos os dados que foram corretamente classificados[8].

$$\frac{2 * P * R}{P + R} \tag{3}$$

[4]https://xgboost.readthedocs.io/en/latest/python/index.html

sendo P, a precisão e R, o recall.

Na Fig. 1, podemos ver os resultados. Na acurácia, o Multinomial Naive Bayes obteve uma média melhor que os outros, mas em alguns casos, perdeu para o SGD. Já a precisão nos informa que o XGBoost consegue ser melhor para identificar os dados que são discurso de ódio.

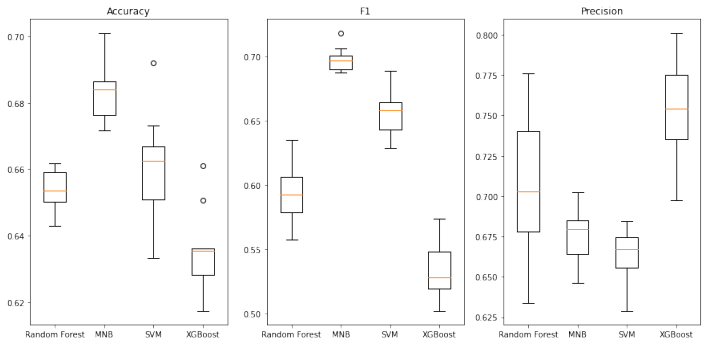

Fig. 1. Resultados dos classificadores utilizando o dataset unificado.

Comparando com [2], utilizando a métrica F1, o modelo Multinomial Naive Bayes alcançou resultados próximos. Em [2], o modelo utilizado alcançou 0.72 na métrica, enquanto que o Multinomial Naive Bayes alcançou média 0.697 e desvio padrão igual a 0.009. Já em [1], ainda utilizando a métrica F1 para comparação, nenhum dos modelos apresentados neste trabalho conseguiu alcançar resultados próximos aos apresentados em [1].

V. Conclusão

Neste trabalho, foram utilizados, como base, dois trabalhos anteriores que analisam discursos de ódio de fontes distintas. Em um deles, foi utilizado o Twitter para coletar os dados [2] e, no outro, comentários do portal de notícias G1 [1]. Para realizarmos os experimentos, construímos um único dataset, a partir da união dos dois datasets mencionados e os dados foram classificados utilizando quatro classificadores: Random Forest, Multinomial Naive Bayes, Stochastic Gradient Descent e XGBoost. Nos resultados, podemos notar que o Multinomial Naive Bayes foi o que melhor classificou todos os dados, observando a acurácia mas o XGBoost, a partir da precisão, melhor classificou os dados como sendo discurso de ódio.

References

[1] Rogers P. de Pelle and Viviane P. Moreira, “Offensive Comments in the Brazilian Web: a dataset and baseline results”, 6th Brazilian Workshop on Social Network Analysis and Mining (BraSNAM),2017.

[2] Fortuna, Paula, João Rocha da Silva, Juan Soler-Company, Leo Wanner and Sérgio Nunes, “A Hierarchically-Labeled Portuguese Hate Speech Dataset”, Proceedings of the 3rd Workshop on Abusive Language Online (ALW3), 2019.

[3] Breiman, Leo, “Random forests”, Machine learning, vol. 45, pp. 5–32, 2001.

[4] Juan, Alfons and Ney, Hermann, “Reversing and Smoothing the Multinomial Naive Bayes Text Classifier”, PRIS, pp. 200–212, 2002.

[5] Do, Thanh-Nghi, “Parallel multiclass stochastic gradient descent algorithms for classifying million images with very-high-dimensional signatures into thousands classes”, Vietnam Journal of Computer Science, vol. 1, pp. 107–115, 2014.

[6] Chen, Tianqi and Guestrin, Carlos, “Xgboost: A scalable tree boosting system”, Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794, 2016.

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

[8] Hossin, Mohammad and Sulaiman, MN, “A review on evaluation metrics for data classification evaluations”, International Journal of Data Mining & Knowledge Management Process, vol. 5, 2015.