UTILIZANDO BERTIMBAU PARA A CLASSIFICAÇÃO DE DISCURSO DE ÓDIO EM PORTUGUÊS

FÉLIX LEONEL VASCONCELOS DA SILVA[1]; LARISSA ASTROGILDO DE FREITAS[2]

[1]*Universidade Federal de Pelotas – flvdsilva@inf.ufpel.edu.br*
[2]*Universidade Federal de Pelotas – larissa@inf.ufpel.edu.br*

1. INTRODUÇÃO

Não existe uma única definição para discurso de ódio, a ONU (Organização das Nações Unidas) o descreve como qualquer tipo de comunicação na fala, escrita ou comportamento que ataque ou use linguagem pejorativa ou discriminatória com referência a uma pessoa ou grupo com base em quem ela é, em outras palavras com base em sua religião, etnia, nacionalidade, raça, cor, descendência, gênero ou outro fator de identidade, o discurso de ódio pode ser veiculado sob qualquer forma de expressão, incluindo: imagens, desenhos, memes, objetos, gestos e símbolos e pode ser disseminado *offline* e *online* (GUTERRES, 2021).

O Aprendizado Profundo (do inglês, *Deep Learning*) é uma sub-área do Aprendizagem de Máquina, que emprega algoritmos para processar dados e imitar o processamento feito pelo cérebro humano (DATA SCIENCE ACADEMY, 2021).

Transformer é uma abordagem do Aprendizado Profundo introduzido em 2017 que usa o mecanismo de *self-attention*. BERT (do inglês, *Bidirectional Encoder Representations from Transformers*) é uma metodologia de treinamento dos *Transformers*, mas também é o nome dos modelos pré-treinados por esta metodologia, ele atingiu o estado da arte quando foi aplicado a diferentes tarefas de NLP (do inglês, *Natural Language Processing*) (DATA SCIENCE ACADEMY, 2021).

Neste trabalho nós aplicamos o BERT na tarefa de classificação de discurso de ódio, utilizaremos o conjunto de dados criado por FORTUNA *et al.* (2019) e o modelo BERTimbau (SOUZA *et al.*, 2020).

2. METODOLOGIA

Como está representado na Figura 1, inicialmente, os *tweets* são pré-processados, depois são aplicadas algumas técnicas de *data augmentation*, nós usamos o modelo BERTimbau (SOUZA *et al.*, 2020) e o ajustamos (*fine tuning*) para a tarefa de classificação de discurso de ódio, no final os resultados obtidos são analisados.

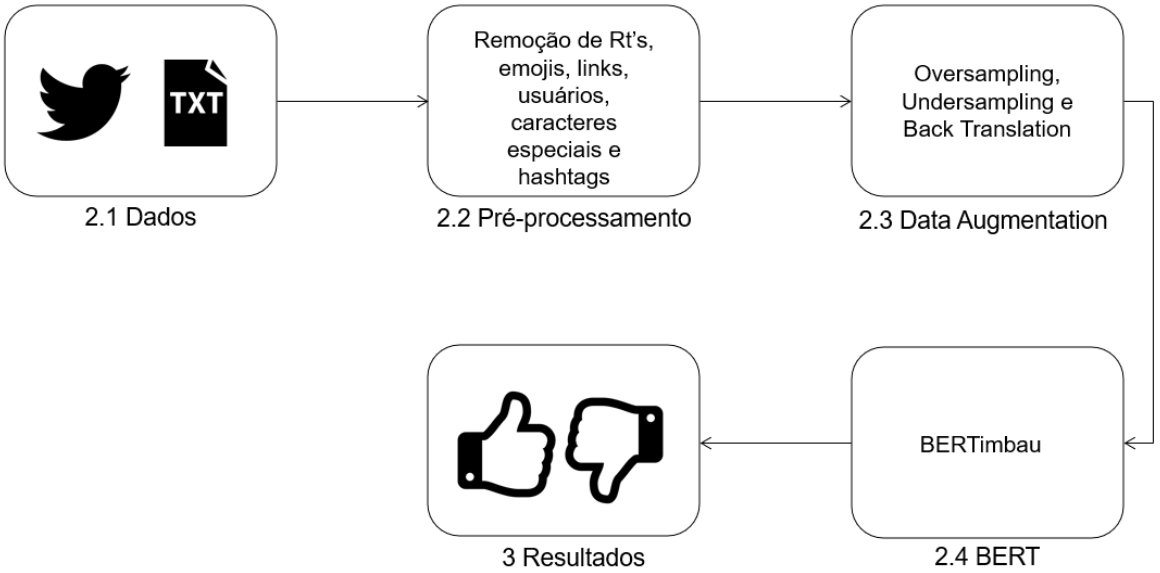

Figura 1. Fluxo da abordagem proposta. Fonte: Própria.

2.1 Dados

Nos experimentos realizados neste trabalho, foi utilizado o conjunto de dados criado por FORTUNA *et al.* (2019). Os dados foram coletados da rede social Twitter. Para isso, a autora utilizou a API de busca de perfil do Twitter para pesquisar palavras-chave e hashtag. Cada um dos 5668 *tweets* foram anotados por três anotadores (1786 sentenças ofensivas e 3882 sentenças não ofensivas). Um exemplo de discurso de ódio seria: “bom dia sapatao da minha vida” e um exemplo de não discurso de ódio seria: “E o sono… rs. Cheiros. Tb adorei!”.

2.2 Pré-processamento

Pré-processamento é um passo importante no NLP. Ele determina a qualidade final dos dados que vão ser analisados (JURAFSKY; MARTIN, 2009). Neste trabalho foram utilizados pré-processamentos conhecidos na literatura aplicados em tweets, como: remover os caracteres especiais, usuário, links, emojis, RT’s e hashtags.

2.3 Data Augmentation

Data augmentation é uma técnica para gerar novos exemplos de dados de treinamento para balancear os conjuntos de dados, há alguns tipos como: sobreamostragem (do inglês, *oversampling*), subamostragem (do inglês, *undersampling*), tradução reversa (do inglês, *back translation)*, substituição de sinônimos, entre outros (BEDDIAR *et al.* 2021).

Neste trabalho foi utilizado subamostragem, tradução reversa e sobreamostragem. Subamostragem é uma técnica que retira sentenças da classe majoritária até igualar com a classe minoritária. Sobreamostragem é uma técnica que adiciona sentenças a classe minoritária até igualar com a majoritária (MOHAMMED *et al.* 2019). Tradução reversa é o processo de traduzir de uma

linguagem alvo de volta a sua linguagem de origem, nós utilizamos inglês como a linguagem alvo (BEDDIAR *et al.* 2021).

2.4 BERT

O BERT treina os modelos linguísticos com base no conjunto completo de palavras, em uma consulta ou frase conhecido como treinamento bidirecional, tornando os modelos linguísticos capazes de discernir o contexto das palavras com base nas palavras ao seu redor (DATA SCIENCE ACADEMY, 2021). Neste trabalho foi utilizado o modelo BERTimbau base cased (SOUZA *et al.* 2020), um modelo pré-treinado do BERT para português, nós utilizamos taxa de aprendizagem 2e-5, *batch* de 32 e 4 épocas.

3. RESULTADOS E DISCUSSÃO

Na Tabela 1 apresentamos os resultados obtidos. Os dados foram divididos em 80% para treinamento, 10% para validação e 10% para teste. Para atingir os resultados foram utilizadas cinco configurações: (1) Original; (2) Sem Caracteres Especiais; (3) Sobreamostragem; (4) Subamostragem; (5) Tradução Reversa.

Neste trabalho foram utilizadas três métricas para avaliar os resultados obtidos, Acurácia (Acc), Acurácia Balanceada (Bacc) e Medida-F. Acc é o número de acertos dividido pelo número total de exemplos. Bacc é o cálculo de todos os acertos dividido por todos os acertos mais os erros. Medida-F é a média harmônica entre Precisão e Revocação (MÜLLER; GUIDO, 2017).

Tabela 1. Resultados obtidos nos experimentos.

	Acc	Bacc	Medida-F
(1)	**0,86**	**0,83**	**0,86**
(2)	0,83	0,77	0,82
(3)	**0,86**	0,82	0,85
(4)	0,82	0,81	0,82
(5)	0,82	0,78	0,82

Na maioria dos casos, há uma melhor performance quando as sentenças são utilizadas na forma em que elas foram escritas pelos seus autores, este conjunto de dados não é muito afetado pelo desbalanceamento das classes, atingindo os melhores resultados com a configuração (1) com 0,86 de Acc 0,83 para Bacc e 0,86 para Medida-F. Na configuração (3) também atingiu 0,86 de Acc, tendo os piores desempenhos com as configurações (4) e (5), obtendo os mesmos resultados para Acc e Medida-F. A configuração (2) atingiu o pior resultado de Bacc.

4. CONCLUSÕES

Neste trabalho foi utilizado o BERT para identificar o discurso de ódio no conjunto de dados criado por FORTURNA *et al.* (2019). Para isso, foram usados alguns pré-processamentos e três técnicas de *data augmentation*. Os melhores resultados foram obtidos com a configuração (1), indicando que deixar os *tweets* como eles foram escritos pelos usuários é a melhor abordagem para esse conjunto de dados. Ainda, obtivemos melhores resultados quando comparados com o trabalho FORTUNA *et al.* (2019), o nosso trabalho obteve Medida-F de 0,86 utilizando BERTimbau (SOUZA *et al.*, 2020) e o trabalho de FORTUNA *et al.* (2019) obteve Medida-F de 0,78 utilizando LSTM (do inglês, *Long Short-Term Memory*)

5. REFERÊNCIAS BIBLIOGRÁFICAS

Fortuna, P., Rocha da Silva, J., Soler-Company, J., Wanner, L., and Nunes, S. (2019). A hierarchically-labeled Portuguese hate speech dataset. In: 3rd Workshop on Abusive Language Online, pages 94–104, Florence, Itália. ACL.

Data Science Academy (2021). Deep learning book. Acesso em: 20 de Julho de 2022. Disponível em: https://www.deeplearningbook.com.br.

Guterres, A. (2021). What is hate speech?. Disponível em:https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech. Acessado em: 08 de Agosto de 2022.

Jurafsky, D. e Martin, J. H. (2009). Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, 2nd Edition. Prentice Hall series in artificial intelligence. Prentice Hall, Pearson Education International.

Mohammed, R., Rawashdeh, J., and Abdullah, M. (2020). Machine learning with oversampling and undersampling techniques: Overview study and experimental results. 2020 11th International Conference on Information and Communication Systems(ICICS), pages 243–248

Beddiar.R. D, Jaham S Md, Oussalah M (2021). Data expansion using back translation and paraphrasing for hate speech detection,Online Social Networks and Media, Volume 24, 2021, 100153, ISSN 2468-6964, https://doi.org/10.1016/j.osnem.2021.100153.

Souza, Fábio & Nogueira, Rodrigo & Lotufo, Roberto. (2020). BERTimbau: Pretrained BERT Models for Brazilian Portuguese. 10.1007/978-3-030-61377-8_28.

Müller, A C., and Sarah Guido. Introduction to Machine Learning with Python: A Guide for Data Scientists.Sebasttopol CA, 2017.