DETECÇÃO DE DISCURSO DE ÓDIO EM PORTUGUÊS USANDO BERT

FÉLIX LEONEL V. DA SILVA[1]; LARISSA A. 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

O discurso de ódio é descrito como uma linguagem que ataca ou denigre um certo grupo baseado em sua raça, etnia, religião, sexo, idade, ou orientação sexual (NOBATA *et al.*, 2016). Com a internet, a disseminação de discurso de ódio se tornou muito ampla e passou a se disseminar principalmente em: redes sociais, blogs, vídeos e outros canais de comunicação. Com o anonimato e a sensação de impunidade as pessoas se sentem encorajadas a espalhar todo tipo de comentários ofensivos e discriminatórios. Em 2020 após o assassinato de George Floyd, foi iniciada uma campanha liderada por proeminentes grupos de direitos civis e organizações sem fins lucrativos, a chamada "Stop Hate for Profit"[1], pressionando marcas a suspenderem anúncios pagos ao Facebook[2] até que medidas de combate à desinformação e a disseminação de discurso de ódio fossem tomadas. Após a adesão de grandes companhias, o Facebook passou a tomar medidas contra o discurso de ódio.

Aprendizado Profundo (do inglês, *Deep Learning - DL*) é um tipo de aprendizado de máquina que treina computadores para realizar tarefas como seres humanos, o que inclui reconhecimento de fala, identificação de imagem e etc. O *Transformer* é uma abordagem de DL introduzido em 2017 que utiliza o mecanismo de self-attention. BERT (*Bidirectional Encoder Representations from Transformers*) é uma metodologia de pré-treinamento dos *Transformers* mas também é o nome dos modelos pré-treinados por essa metodologia. Ele tem atingido o estado da arte quando aplicado em diferentes tarefas de PLN (DATA SCIENCE ACADEMY, 2021). Neste trabalho aplicamos o BERT na tarefa de detecção de discurso de ódio em português.

2. METODOLOGIA

A abordagem proposta recebe como entrada um texto em português e o classifica como contendo ou não comentário ofensivo. O fluxo da abordagem proposta pode ser visualizado na Figura 1.

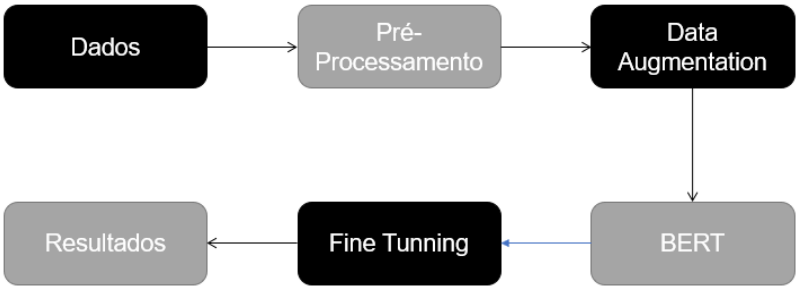

Figura 1. Fluxo da abordagem proposta. Fonte: Própria.

[1] https://www.stophateforprofit.org
[2] https://pt-br.facebook.com

2.1. Dados

Neste trabalho foram utilizados os seguintes conjuntos de dados: os conjuntos de dados criados por de Pelle e Moreira 2017 (OFFCOMBR-2 e OFFCOMBR-3) e o conjunto de dados criado por Fortuna *et al*. 2019.

Para criar OFFCOMBR-2 e OFFCOMBR-3 foram coletados comentários do site G1[3] das categorias política e esportes, dos quais, comentários compostos apenas por emojis foram descartados, no final do processo foram coletados 10336 comentários de 115 notícias. Foram escolhidos de forma aleatória 1250 dos 10336 comentários. Cada um dos 1250 comentários foram anotados por 3 anotadores, em relação a serem ou não ofensivos. O conjunto de dados OFFCOMBR-2 é composto por 1250 comentários (concordância de 3 anotadores, 419 sentenças ofensivas e 831 sentenças não ofensivas) e o OFFCOMBR-3 é composto por 1033 comentários (concordância de 2 anotadores, 202 sentenças ofensivas e 831 sentenças não ofensivas).

O conjunto de dados criado por Fortuna *et al.* 2019 foi coletado da rede social Twitter[4]. Para isso a autora utilizou a API de busca de perfil do Twitter para pesquisar palavras-chave e hashtags como: sapatão, dyke ou #LugarDeMulherENaCozinha. Foram observados 29 perfis específicos, 19 palavras-chave e 10 hashtags. No final do processo foram coletados 42930 tweets. Foram escolhidos 200 tweets por cada instância de pesquisa, resultando em 5668 dos 42930 tweets. Cada um dos 5668 tweets foram anotados por três anotadores (1786 sentenças ofensivas e 3882 sentenças não ofensivas), os quais classificaram cada tweet como sendo ou não ofensivos.

2.2. Pré-processamento

O pré-processamento é um passo importante no PLN. Neste trabalho foram retirados os caracteres especiais (por exemplo: #, @, :, !, ?) e as sentenças foram convertidas para letra maiúscula dos três conjuntos de dados. Ainda, no conjunto de dados de Fortuna *et al.* 2019 foram retirados os *retweets* (RT's).

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 (over-sampling) que cria mais sentenças da classe minoritária e subamostragem (under-sampling) que retira sentenças da classe maioritária (LYASHENKO, 2021). Neste trabalho foi utilizado *over-sampling*.

2.4. BERT

Neste trabalho foi utilizado o BERT *for Sequence Classification* que é o modelo BERT[5] com uma camada *Feed Forward Network para* detectar discurso de ódio em português.

2.5. *Fine Tuning*

Fine Tuning significa fazer pequenos ajustes em um processo para obter a saída ou desempenho desejado. No caso do DL, envolve o uso de pesos de um algoritmo de DL anterior para programar outro processo de DL semelhante. Neste trabalho foi utilizado o otimizador Adam que é um algoritmo baseado em gradiente de primeira ordem de funções objetivas estocásticas, baseado em estimativas adaptativas de momentos de ordem inferior (KINGMAN e BA, 2014), com o tamanho do batch de 32, taxa de aprendizado de 2e-5 e 4 épocas para o treinamento.

---

[3] https://g1.globo.com
[4] https://twitter.com
[5] https://huggingface.co/transformers/model_doc/bert.html#bertforsequenceclassification

3. RESULTADOS E DISCUSSÃO

Nas Tabelas 1, 2 e 3 apresentamos os resultados obtidos. Para isso, utilizamos três configurações: (1) Original que é quando as sentenças são mantidas como são escritas nos conjuntos de dados; (2) *Data Augmentation* que quando usamos a técnica de *over-sampling*; (3) Sem C.E. (Caracteres Especiais), Sem RT's e L.M. (Letra Maiúscula) que é quando os caracteres especiais e *retweets* são retirados e as sentenças são convertidas para letra maiúsculas.Para avaliar os resultados obtidos utilizaremos as métricas Acurácia (Acc), Acurácia Balanceada (Bacc) e Medida-F (PEDREGOSA *et al.*, 2011). Acc é o número de acertos dividido pelo total de exemplos. Bacc é o cálculo de todos os acertos divididos por todos os acertos mais os erros. Medida-F é a média harmônica entre precisão e revocação.

Para o OFFCOMBR-3 como pode ser observado na Tabela 1, os melhores resultados para as métricas Acc, Bacc e Medida-F foram encontradas usando *Data Augmentation* com valores de 0,89, 0,83 e 0,88, respectivamente.

Tabela 1. Experimentos com OFFCOMBR-3.

	Acc	Bacc	Medida-F
Original	0,88	0,79	0,87
Data Augmentation	**0,89**	**0,83**	**0,88**
Sem C.E. e L.M.	**0,83**	**0,69**	**0,82**

Para o OFFCOMBR-2 como pode ser observado na Tabela 2, os melhores resultados para as métricas Acc, Bacc e Medida-F foram encontradas usando as sentenças originais com valores de 0,89, 0,87 e 0,89, respectivamente.

Tabela 2. Experimentos com OFFCOMBR-2.

	Acc	Bacc	Medida-F
Original	**0,89**	**0,87**	**0,89**
Data Augmentation	0,87	0,84	0,86
Sem C.E. e L.M.	**0,75**	**0,77**	**0,76**

Para o conjunto de dados de Fortuna *et al.* 2019 como pode ser observado na Tabela 3, a melhor Acc foi encontrada usando as sentenças originais e *Data Augmentation* com valor de 0,86 e a melhor Bacc e Medida-F foram encontradas usando as sentenças originais com os valores de 0,83 e 0,86, respectivamente.

Tabela 3. Experimentos com Fortuna *et al.* 2019.

	Acc	Bacc	Medida-F
Original	**0,86**	**0,83**	**0,86**
Data Augmentation	**0,86**	0,82	0,85
Sem C.E. e Sem RT's e L.M.	**0,78**	**0,75**	**0,78**

Já, os piores resultados para os três conjuntos foram encontrados ao remover C.E e convertendo as sentenças para L.M.

4. CONCLUSÕES

Neste trabalho foi utilizado o BERT para detectar discurso de ódio nos conjuntos de dados OFFCOMBR-2, OFFCOMBR-3 e de Fortuna *et al.* 2019. Para isso, foram usados alguns pré-processamento e os melhores resultados tanto para Acc, Bacc e Medida-F foram obtidos com o OFFCOMBR-2 usando sentenças originais. Comparando nosso trabalho com outros, os resultados obtidos usando o BERT foram melhores. No trabalho de Pelle e Moreira 2017, os autores utilizaram *Support Vector Machines* (SVM) e *Naïve Bayes* (NB) nos conjuntos de dados OFFCOMBR-2 e OFFCOMBR-3. Os autores obtiveram Medida-F de 0,77 para o SVM e 0,71 para o NB (OFFCOMBR-2) e de 0,82 para o SVM e 0,79 para o NB (OFFCOMBR-3). Nosso trabalho obteve Medida-F de 0,89 usando BERT com as sentenças originais (OFFCOMBR-2) e 0,89 usando BERT com *Data Augmentation* (OFFCOMBR-3). Ainda, no trabalho de Fortuna *et al.* 2019, a autora obteve Medida-F de 0,78 usando *Long Short-Term Memory* (LSTM). Nosso trabalho obteve Medida-F de 0,86 usando BERT com as sentenças originais e *Data Augmentation*.

5. REFERÊNCIAS BIBLIOGRÁFICAS

de Pelle, R. and Moreira, V. (2017). Offensive comments in the brazilian web: a dataset and baseline results. In: 6th Brazilian Workshop on Social Network Analysis and Mining, Porto Alegre, RS, Brasil. SBC.

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.

Nobata, C., Tetreault, J. R., Thomas, A., Mehdad, Y., and Chang, Y. (2016). Abusive language detection in online user content. In: 25th International Conference on World Wide Web, pages 145–153, Montreal, Canadá. ACM.

Data Science Academy (2021). Deep learning book. Disponível em: https://www.deeplearningbook.com.br. Acesso em: 26 Julho 2021.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. In: 3rd International Conference for Learning Representations, San Diego, Estados Unidos. SJR.

Lyashenko, V. (2021). Data Augmentation in Python: Everything You Need to Know. Disponível em: https://neptune.ai/blog/data-augmentation-in-python. Acesso em: 27 Julho 2021.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. Journal of Machine Learning Research, p.2825–2830.