O IMPACTO DO AUMENTO DE DADOS NA DETECÇÃO DE DISCURSO DE ÓDIO NA LÍNGUA PORTUGUESA.

FÉLIX LEONEL VASCONCELOS DA SILVA[1]; ARTUR CERRI[2]; LARISSA ASTROGILDO DE FRETAS[3]

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

1. INTRODUÇÃO

A história da mídia social on-line ou digital é evidentemente um desenvolvimento recente, indicando o uso crescente de computadores para conectar pessoas. As redes sociais, como o Facebook, são dedicadas a criar e refletir relacionamentos dentro de comunidades com interesses semelhantes (BANDGAR, 2014).

De acordo com MATHEW *et. al.* 2018, a prevalência do discurso de ódio on-line contribuiu para crimes de ódio horríveis no mundo real, como o genocídio em massa dos muçulmanos *Rohingya*, a violência comunitária em Colombo e o massacre na sinagoga de *Pittsburgh*. Consequentemente, é imperativo entender a difusão desse conteúdo de ódio em um ambiente on-line.

Detectar discursos de ódio é uma tarefa desafiadora. A divergência no entendimento sobre a rotulagem humana do discurso de ódio ressalta a dificuldade crescente dessa classificação para modelos de computador (FIRMINO, 2022).

O Processamento de Linguagem Natural (PLN) é um campo da Inteligência Artificial e da Linguística que se concentra em permitir que os computadores entendam expressões ou palavras em idiomas humanos, conhecidos como idiomas naturais (KHURANA *et. al.*, 2022).

O discurso de ódio é definido como um discurso hostil e mal-intencionado motivado por preconceito, direcionado a indivíduos ou grupos com base em características inatas, reais ou percebidas. Ele expressa atitudes discriminatórias, intimidadoras, desaprovadas, antagônicas e prejudiciais em relação a atributos como gênero, raça, religião, etnia, cor, nacionalidade, deficiência ou orientação sexual (COHEN-ALMAGOR, 2013).

Aumento de dados é uma estratégia crucial no aprendizado de máquina e tem importância especial nas tarefas de PLN. A tarefa de detecção de discurso de ódio enfrenta desafios devido à escassez de conjuntos de dados rotulados em português. As técnicas de aumento de dados são vitais para aumentar os dados de treinamento disponíveis (PELLICER *et. al.*, 2023).

O Aprendizado Profundo (AP) permite que modelos de computador com várias camadas de processamento aprendam representações de dados em vários níveis de abstração. Esses métodos avançaram significativamente o reconhecimento de fala, o reconhecimento visual de objetos e a detecção de

objetos. Os algoritmos de AP usam a retropropagação para indicar como uma máquina deve ajustar seus parâmetros internos para calcular a representação em cada camada com base na representação na camada anterior (LECUN, 2015).

Um desenvolvimento inovador na arquitetura de Aprendizado de Máquina é a introdução dos *Transformers*. Essa arquitetura de modelo evita a recursão e depende inteiramente de um mecanismo de atenção para estabelecer dependências globais entre entrada e saída (VASWANI *et. al.*, 2017).

Um aplicativo notável da arquitetura *Transformer* é o *Bidirectional Encoder Representations from Transformers* (BERT), projetado para pré-treinar representações bidirecionais profundas a partir de texto não rotulado, condicionando conjuntamente o contexto esquerdo e direito em todas as camadas (DEVLIN *et. al.*, 2019).

2. METODOLOGIA

Neste trabalho foram utilizados o conjunto de dados de Fortuna (FORTUNA *et. al.*, 2019) e o modelo BERTimbau base (SOUZA *et. al.*, 2020). Os conjuntos de dados foram divididos em 80% para treinamento, 10% para validação e 10% para teste. É importante observar que usamos como hiperparâmetros: 8 para o tamanho do lote, 4 épocas para treinamento, *CrossEntropy* para a função de perda e *AdamW* como otimizador.

Para a realização dos experimentos deste trabalho foi utilizado o conjunto de dados de Fortuna (FORTUNA *et. al.*, 2019), os dados desse conjunto de dados foram coletados da rede social Twitter/X entre janeiro e março de 2017. Os autores usaram a API de pesquisa de perfil do Twitter/X para palavras-chave e hashtags como sapatão ou \#LugarDeMulherENaCozinha. Foram observados 29 perfis específicos, 19 palavras-chave e 10 hashtags. Ao final do processo, foram coletados 42930 tweets. Duzentos tweets foram escolhidos para cada instância de pesquisa, resultando em 5668 dos 42930 tweets. Cada um dos 5668 tweets foi anotado por três anotadores, com 1786 tweets com discurso de ódio e 3882 tweets sem discurso de ódio.

Foram utilizadas as configurações: (i) Original, onde os dados permanecem da maneira que estão no conjunto de dados; (ii) *Oversampling*, que envolve aumentar o número de instâncias ou amostras da classe minoritária até que ela corresponda à classe majoritária, gerando novas instâncias ou repetindo algumas instâncias; (iii) *Undersampling*, é o processo de reduzir o número de instâncias ou amostras da classe majoritária até que ele corresponda ao número da classe minoritária (MOHAMMED *et. al.*, 2020); (iv) Substituição de Sinônimos, seleciona n palavras aleatórias (excluindo *stopwords*) e as troca por sinônimos escolhidos aleatoriamente; (v) *Text Augmentation*, combina substituição de sinônimos, troca de palavras aleatórias, troca de caracteres aleatórios e adição de ruído (WEI *and* ZOU, 2019).

3. RESULTADOS E DISCUSSÃO

No contexto do conjunto de dados de Fortuna (FORTUNA *et. al.*, 2019), como pode ser observado na Tabela 1, apesar de uma diferença marginal, a configuração *OverSampling* superou consistentemente as configurações alternativas em todas as métricas, atingindo 0,94 para acurácia balanceada. Por outro lado, a configuração UnderSampling apresentou resultados inferiores em comparação com suas contrapartes, atingindo 0,87 para acurácia balanceada. A lógica por trás da disparidade está no fato de que a *UnderSampling*, implica a remoção de instâncias do conjunto de dados para corrigir desequilíbrios das classes. Embora tenha como objetivo o equilíbrio, esse processo diminui inadvertidamente o conjunto de exemplos disponíveis para classificação, podendo descartar textos essenciais e cruciais para uma classificação diversificada e abrangente. A consequente perda de diversidade textual pode servir como uma explicação plausível para o desempenho abaixo do ideal associado à configuração *UnderSampling*. Essas descobertas ressaltam a sensibilidade do conjunto de dados às nuances da distribuição de classes e destacam a função essencial de estratégias de amostragem eficazes para melhorar os resultados da classificação.

Tabela 1. Resultados dos experimentos.

Configuração	Acurácia	Acurácia Balanceada	Medida-F
Original	0,93	0,91	0,93
Oversampling	**0,94**	**0,94**	**0,94**
Undersampling	**0,85**	**0,87**	**0,85**
Substituição de Sinônimos	0,92	0,91	0,92
Text Augmentation	0,92	0,90	0,92

4. CONCLUSÕES

Este estudo teve como objetivo detectar discurso de ódio usando o BERTimbau como modelo. Notavelmente, a aplicação do aumento de dados teve impacto mínimo sobre os resultados, indicando que o desbalanceamento de classe não afetou significativamente os resultados. Como trabalhos futuros, pretendemos incorporar a funcionalidade de *back translation* no aumento de dados, que está pendente de implementação no momento. Para isso, estamos considerando aproveitar os modelos *Transformers* para as traduções necessárias no processo de retrotradução. Por fim, planejamos substituir o dicionário atual pelo WordNet para preservar melhor a semântica das frases.

Agradecemos à CNPq, à FAPERGS e à NVIDIA Corporation pelo financiamento parcial deste trabalho.

5. REFERÊNCIAS BIBLIOGRÁFICAS

Bandgar, B. 2014. Role of social networks in recent era. International Journal of Research in Computer Science and Management 1(1):2321–8088.

Mathew, B.; Dutt, R.; Goyal, P.; Mukherjee, A. Spread of hate speech in online social media. In: ACM Conference on Web Science, 11., 2019, Boston, MA, USA. Proceedings of the 11th ACM Conference on Web Science, WebSci 2019. Boston, MA, USA: ACM, 2019. p. 173 - 182.

Firmino, A. A. 2022. Uma abordagem para detecção de discurso de ódio utilizando aprendizado de máquina baseado em cruzamento de idiomas. Ph.D. Dissertation, Universidade Federal de Campina Grande.

Khurana, D.; Koli, A.; Khatter, K.; and Singh, S. 2022. Natural language processing: state of the art, current trends and challenges. Multimedia Tools and Applications 82(3):3713–3744.

Cohen-Almagor, R. Freedom of expression v. social responsibility: Holocaust denial in Canada. Journal of Mass Media Ethics, v. 28, n. 1, p. 42-56, 2013.

Pellicer, L. F. A. O.; Ferreira, T. M.; Costa, A. H. R. Data augmentation techniques in natural language processing. Applied Soft Computing, Amsterdam, v. 132, 2023.

Lecun, Y.; Bengio, Y.; HintpnINTON, G. Deep learning. Nature, Londres, v. 521, n. 7553, p. 436-444, 2015.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention is all you need. Advances in Neural Information Processing Systems, California USA, v. 30, p. 6000-6010, 2017.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Souza, F.; Nogueira, R.; and Lotufo, R. 2020. Bertimbau: pretrained BERT models for Brazilian Portuguese. In Proceedings of the 9th Brazilian Conference on Intelligent Systems.

Fortuna, P.; Rocha da Silva, J.; Soler-Company, J.; Wanner, L.; and Nunes, S. 2019. A hierarchically-labeled Portuguese hate speech dataset. In Proceedings of the 3rd Workshop on Abusive Language Online.

Mohammed, R., Rawashdeh, J. and Abdullah, M.; "Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results," 2020 11th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 2020, pp. 243-248, doi: 10.1109/ICICS49469.2020.239556.

Wei, J., and Zou, K. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing.