EXPLORANDO OS MODELOS BERTIMBAU E ALBERTINA NA DETECÇÃO DE DISCURSO DE ÓDIO

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

O Discurso de Ódio é caracterizado como um discurso agressivo e prejudicial, impulsionado por preconceitos, que visa prejudicar um indivíduo ou um grupo de pessoas devido a certas características inatas, reais ou percebidas. Ele manifesta atitudes discriminatórias, ameaçadoras, desaprovadoras, hostis e/ou prejudiciais em relação a tais características, que podem incluir gênero, raça, religião, etnia, cor, nacionalidade, deficiência ou orientação sexual. O propósito do discurso de ódio é causar dor, desumanização, assédio, intimidação, depreciação, degradação e vitimização dos grupos alvo, além de promover a insensibilidade e a crueldade contra eles (ALMAGOR, 2013).

As redes sociais desempenham um papel fundamental na disseminação do discurso de ódio online. Esse fato contribui consideravelmente para a dificuldade em detectá-lo, pois as postagens nas redes sociais envolvem elementos paralinguísticos, como *emoticons* e *hashtags*, e seu conteúdo linguístico muitas vezes contém erros de escrita. A tarefa de lidar com o discurso de ódio envolve identificar se diferentes formas de comunicação, como texto, áudio e outras, contêm expressões de ódio ou incitam à violência contra indivíduos, ou grupos específicos. Outra dificuldade reside na natureza contextualmente dependente dessa tarefa e na falta de consenso sobre o que realmente constitui o discurso de ódio, tornando a tarefa desafiadora até mesmo para os seres humanos. Isso dificulta criar grandes conjuntos de dados rotulados e requer um alto consumo de recursos (KOVÁCS *et al.* 2021).

O Processamento de Linguagem Natural (PLN) é uma disciplina que se encontra na interseção da Inteligência Artificial e da Linguística. Seu foco está em capacitar os computadores para compreenderem e interpretarem expressões ou palavras escritas em linguagens humanas, aquelas que ocorrem naturalmente em comunicações humanas (DIKSHA e*t al.* 2022).

O Aprendizado Profundo (AP) possibilita que modelos computacionais, que consistem em várias camadas de processamento, adquiram representações de dados em diferentes níveis de abstração. Essas técnicas tiveram um impacto significativo no avanço do reconhecimento de fala, na identificação visual de objetos e na detecção de objetos (YANN *et al.* 2015).

Transformer é uma técnica de AP que foi apresentada em 2017 e emprega o conceito de autoatenção. BERT, que significa "*Bidirectional Encoder Representations from Transformers*" em inglês, é uma estratégia de treinamento para os modelos *Transformers*. Além disso, BERT é também o nome dos modelos pré-treinados por essa abordagem. Eles alcançaram o desempenho líder quando aplicados em várias tarefas de PLN (VASWANI 2017).

2. METODOLOGIA

O trabalho é composto por quatro passos: (1) os modelos BERTimbau (SOUZA *et al.* 2020) e Albertina (RODRIGUES *et al.* 2023) são utilizados, (2) é aplicado o *fine-tuning* na tarefa de detecção de discurso de ódio no *dataset* ToLD-BR (LEITE *et al.* 2020), (3) os resultados são analisados e (4) os dois modelos são comparados.

Para a realização do trabalho foi utilizado o *dataset* ToLD-BR (LEITE et al. 2020). Esse *dataset* incluiu *tweets* coletados durante os meses de Julho e Agosto de 2019, por meio da ferramenta de coleta do Twitter *GATE Cloud*. Duas estratégias distintas foram empregadas para a coleta desses *tweets*. A primeira estratégia consistiu em buscar palavras-chave específicas e *hashtags* predefinidas, tais como "gay", "mulherzinha" e "nordestino". Já a segunda estratégia envolveu a coleta de *tweets* que mencionam figuras influentes, como o ex-presidente do Brasil, Jair Bolsonaro, e o jogador de futebol Neymar Jr. Nesse método, não houve restrições quanto a palavras-chave ou *hashtags*, resultando na obtenção de mais de 10 milhões de *tweets* únicos. A partir desse conjunto, 21.000 *tweets* foram selecionados aleatoriamente para compor o *dataset*. Importante ressaltar que a primeira estratégia representou 60% dos dados coletados. Para a anotação deste *dataset*, contou-se com a participação de 42 anotadores, os quais classificaram 1.500 *tweets* como LGBTQ+fobia, obscenos, insultantes, racistas, misóginos ou xenofóbicos. Ao final, o *dataset* continha 9.245 *tweets* ofensivos e 11.693 *tweets* não ofensivos, cada um deles classificado por três anotadores.

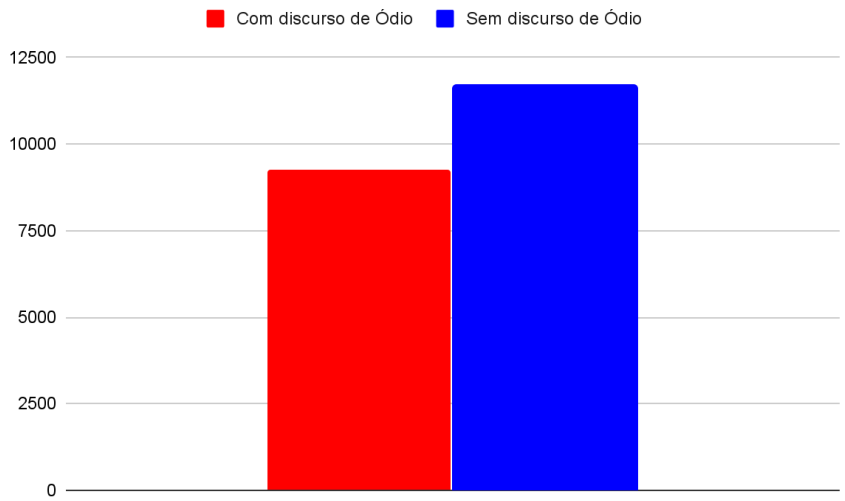

Figura 1. Gráfico de distribuição das classes do *dataset* ToLD-BR.

Também foram utilizados dois modelos: BERTimbau (SOUZA *et al.* 2020) e Albertina (RODRIGUES *et al.* 2023).

O BERTimbau desenvolvido pela NeuralmindAI, é um modelo líder em desempenho quando se trata de avaliação de Similaridade Textual de Sentenças (STS), Reconhecimento de Atributos Textuais e Identificação de Entidades Nomeadas no contexto da língua portuguesa do Brasil (SOUZA *et al.* 2020). O Albertina é uma versão avançada do modelo de linguagem BERT, especialmente projetada para o idioma português. Pertencente à família BERT e fundamentada nos *Transformers*, Albertina é uma evolução baseada no modelo DeBERTa (HE *et*

al. 2020), oferecendo um desempenho altamente competitivo para o português. Essa variante do Albertina foi treinada especificamente com dados do português do Brasil, utilizando o conjunto de dados brWaC (RODRIGUES *et al.* 2023).

Os hiperparâmetros utilizados neste trabalho foram definidos como: 4 épocas para o treinamento, 2e-5 para a taxa de aprendizagem, *crossentropy* como função de *loss* e Adamw como otimizador para o BERTimbau Base e Large; 8 de *batch*, 3 épocas para o treinamento, 1e-5 de taxa de aprendizagem, *crossentropy* como função de *loss* e Adamw como otimizador para o Albertina Base, 2 de *batch* e 3 épocas para o treinamento, 1e-5 de taxa de aprendizagem, *crossentropy* como função de *loss* e Adamw como otimizador para o Albertina Large, 32 de *batch.*

3.RESULTADOS E DISCUSSÃO

Os resultados obtidos com os experimentos deste trabalho estão descritos na Tabela 1, os resultados atingidos com a utilização do BERTimbau foram melhores tanto para o modelo Base quanto para o modelo Large. O BERTimbau Large atingiu os melhores resultados com 0,89 de acurácia, 0,90 de precisão, 0,89 de revocação e 0,89 de medida-f, e o Albertina Large atingiu os piores resultados com 0,58 de acurácia, 0,34 de precisão, 0,58 de revocação e 0,43 de medida-f.Uma coisa que foi observada durante os experimento foi que os modelos erravam mais quando os textos estavam mal escritos ou abreviados.

Uma grande dificuldade da detecção de discurso de ódio é identificar o contexto, muitos dos textos são mal escritos, como: “Ui Noooooooofa que lindo fofa Nosso galao e mara ne amiga rajkazblanks”, outro fator que afetou os resultados atingidos foram os hiperparâmetros utilizados nos experimentos, que foram reduzidos pela limitação de recursos de máquina utilizados neste trabalho.

Tabela 1. Resultados dos experimentos.

Modelo	Tipo	Acurácia	Precisão	Revocação	Medida-F
BERTimbau	Base	0,88	0,89	0,88	0,88
Albertina	Base	0,78	0,72	0,77	0,74
BERTimbau	Large	**0,89**	**0,90**	**0,89**	**0,89**
Albertina	Large	**0,58**	**0,34**	**0,58**	**0,43**

4. CONCLUSÕES

Neste trabalho foram utilizados os modelos BERTimbau e Albertina, Base e Large para a realização da detecção de ódio na língua portuguesa, após a realização dos experimentos foi observado uma grande diferença de performance entre os modelos BERTimbau e Albertina , com o BERTimbau apresentando melhores resultados tanto para a versão Base tanto quanto para a versão Large para a tarefa de detecção de discurso de ódio. O BERTimbau Large atingiu resultados um pouco melhores que o Base com acurácia de 0,89 e 0,88, respectivamente, o que não justificaria a utilização da versão Large para a

realização dos experimentos, uma vez que, ela utiliza mais recursos computacionais. O Albertina Base alcançou resultados bem superiores que o Large para esta tarefa com acurácia de 0,78 e 0,58, respectivamente..

Como trabalhos futuros pretendemos utilizar a técnica de *data augmentation* e também a conversão para a técnica de perguntas e respostas para verificar se há algum ganho nos resultados dos experimentos.

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

5. REFERÊNCIAS BIBLIOGRÁFICAS

ALMAGOR, R. C. (2013). Freedom of Expression v. Social Responsibility: Holocaust Denial in Canada. Journal of Mass Media Ethics 28 (02 2013), 42–56. https://doi.org/10.1080/08900523.2012.746119.

HANS-DIETER Wehle. (2017). Machine Learning, Deep Learning, and AI: What's the Difference?

DIKSHA Khurana, ADITYA Koli, KIRANI Khatter, and SUKHDEV Singh. (2022). Natural
language processing: state of the art, current trends and challenges. Multimedia Tools and Applications 82, 3 (jul 2022), 3713–3744. https://doi.org/10.1007/s11042-
022-13428-4

HE, P., LIU, X., GAO, J., and CHEN, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. CoRR, abs/2006.03654.

KOVÁCS, G., ALONSO, P., and SAINI, R. (2021). Challenges of hate speech detection in social media. SN Computer Science, 2.

LEITE, J. A., Silva, D. F., BONTCHEVA, K., and SCARTON, C. (2020). Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis. In Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing.

RODRIGUES, J.; GOMES, L.; SILVA, J.; BRANCO, A.; SANTOS, R.; CARDOSO, H. L.; and OSÓRIO, T. (2023). Advancing neural encoding of portuguese with transformer albertina pt-*.

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

VASWANI, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need.

YANN LeCun, YOSHUA Bengio, and GEOFFREY Hinton. (2015). Deep learning. nature 521, 7553 (2015), 436.