Detecção automática de linguagem ofensiva: uma análise do uso de algoritmos tradicionais de AM e técnicas de PLN na detecção de linguagem ofensiva em *tweets*

Pesquisador: Renner Carneiro Araújo[1]; Professor-orientador: Me. Jeziel Costa Marinho[2];

RESUMO

Atualmente, as redes sociais são responsáveis por boa parte da interação humana possibilitando o compartilhamento de todos os tipos de opiniões. Contudo, este ambiente proporciona também a propagação de discursos de ódio Parte destes textos contem algum tipo de linguagem ofensiva. Devido ao grande volume de dados presentes nas redes sociais, a detecção automática de textos com discurso de ódio e linguagem ofensiva são de suma importância. Esta pesquisa buscou, utilizando de processamento de linguagem natural (PLN), inteligência artificial (IA) e aprendizado de máquina (AM), implementar diferentes modelos de IA capazes de detectar linguagem ofensiva em textos curtos de *tweets* a partir de um *corpus* anotado e realizar um estudo comparativo entre estes modelos desenvolvidos. Os resultados apontaram que, apesar de haver atualmente técnicas mais avançadas, os modelos tradicionais ainda demostram ter uma boa eficácia em tarefas relacionadas à PLN como a detecção de linguagem ofensiva.

PALAVRAS-CHAVE: Linguagem ofensiva, Redes sociais, Processamento de linguagem natural, Inteligência artificial

FINANCIAMENTO: Esta pesquisa foi realizada com o apoio financeiro do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ) e do Instituto Federal de Educação, Ciência e Tecnologia do Maranhão (IFMA) em forma de bolsa de iniciação científica.

1 Estudante secundarista do Curso Técnico em Informática na modalidade integrada no IFMA-Campus Barra do Corda
Email: renneraraujo@acad.ifma.edu.br

2 Professor EBTT de informática no IFMA-Campus Barra do Corda
Mestre em Ciência da Computação
Email: jeziel.marinho@ifma.edu.br

INTRODUÇÃO

Atualmente, uma grande parte da interação humana acontece por meio das redes sociais que, cada vez mais, se tornam mais presentes no dia a dia das pessoas, possibilitando o compartilhamento de pensamentos, opiniões, ideias e culturas dos mais variados tipos. Contudo, as redes sociais, por meio deste ambiente de partilha ideológica, tem ampliado e potencializado a disseminação do discurso de ódio, muitas vezes por meio do uso de linguagem ofensiva.

De acordo com Zampieri et al (2019), linguagem ofensiva pode ser definida como aquela que contem insultos, ameaças, ofensas, obscenidades ou quaisquer outras palavras de baixo calão. A linguagem ofensiva é utilizada nas redes sociais para tecer comentários ofensivos e promover ataques de intolerância à pessoas, grupos ou instituições. Devido ao grande volume de dados gerados constantemente nas redes sociais, torna-se inviável a análise manual de comentários para averiguação da presença de conteúdo ofensiva e sua eliminação em tempo hábil. Segundo Allan (2017) e Nobata (2016), devido a essa dificuldade, empresas como Facebook e Twitter têm investido bastante na detecção automática de conteúdo ofensivo em suas plataformas.

Conforme afirma Pinheiro (2017), houve um aumento de 67,5% de denúncias de crime de ódio na internet envolvendo racismo, LGBTQIA-fobia, xenofobia, neonazismo, misoginia, apologia a crimes contra a vida e intolerância religiosa. Parte desses crimes são caracterizados pela presença de linguagem ofensiva. Neste cenário, o uso de tecnologias capazes de detectar e eliminar conteúdos abusivos de forma automática e eficiente utilizando Inteligência Artificial (AI), Aprendizado de Máquina (AM) e Processamento de Linguagem Natural (PLN) torna-se muito valiosa (BISPO, 2018).

Além da relevância do tema tratado, esta pesquisa justifica-se pelo fato de que, apesar de existe um grade número de pesquisas voltadas para a detecção de linguagem ofensiva para idiomas como o inglês, para o português esse número ainda é pequeno (DE ALMEIDA e BERTON, 2020).

Esta pesquisa buscou, utilizando um corpus anotado de tweets, algoritmos de IA tradicionais e técnicas de PLN, implementar diferentes modelos de AM para a detecção de linguagem ofensiva e por fim realizar um estudo comparativo entre os modelos implementados.

METODOLOGIA

A primeira etapa da pesquisa consistiu no estudo dos principais conceitos relacionados à inteligência artificial, ao aprendizado de máquina, ao processamento de linguagem natural a fim definir qual a melhor abordagem para a resolução da problemática proposta.

A partir da análise da problemática, definir em um texto a presença ou não de linguagem ofensiva, e dos dados disponíveis, estabeleceu-se que a abordagem a ser seguida seria utilizar algoritmos supervisionados de classificação. Algoritmos de AM Supervisionados, de acordo com Gama et al. (2021), são utilizados em tarefas de aprendizado preditivo com o qual busca-se encontrar uma função que a partir dos dados de treinamento possa ser utilizada para prever um rótulo ou valor que caracterize um novo exemplo. Estes algoritmos podem ser utilizados para problemas de classificação onde o objetivo é prever um rótulo de classe, que é uma escolha de uma lista predefinida de possibilidades.

Os algoritmos utilizados nesta pesquisa foram o SGD[3], DecisionTree[4], RandomForest[5], SVM linear[6], NaiveBayes[7] e Multi-layer Perceptron[8]. Todos eles foram implementados na linguagem Python com a biblioteca Scikit-leaning (GERON, 2019).

A Tabela 1 apresenta os hiperparâmetros de cada algoritmo de classificação utilizado nos experimentos. A definição dos valores dos hiperparâmetros foi feita utilizando o método GridSearchCV da biblioteca sckit-learn que realiza uma busca exaustiva a fim de definir quais os melhores valores para os parâmetros de um algoritmo e com isso otimizar o seu desempenho.

Tabela 1 – Parâmetros de cada classificador usado nos experimentos

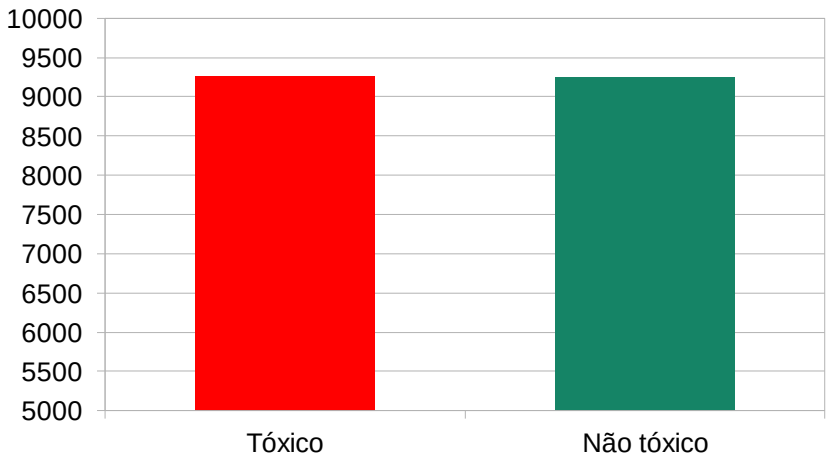
Classificador	Hiperparâmetros
SGD	et0 =0.01, learning_rate = ‘adaptive’, penalty = ‘l1’
DecisionTree	max_depth = 200, max_leaf_nodes= 260, min_samples_split = 3, splitter = 'random'

3 https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html#sklearn.linear_model.SGDClassifier
4 https://scikit-learn.org/stable/modules/tree.html#tree-classification
5 https://scikit-learn.org/stable/modules/ensemble.html#forest
6 https://scikit-learn.org/stable/modules/svm.html
7 https://scikit-learn.org/stable/modules/naive_bayes.html
8 https://scikit-learn.org/stable/modules/neural_networks_supervised.html#multi-layer-perceptron

RandomForest	bootstrap: False, max_depth = None, min_samples_leaf = 2, min_samples_split = 2, n_estimators = 200
SVM linear	Padrão
NaiveBayes	alpha = 2.0, fit_prior = False
MLP	hidden_layer_sizes=(100, 50), max_iter=1000, random_state= 42, solver='sgd'

Sardinha (2004) define Corpus como sendo um conjunto de dados linguísticos criteriosamente sistematizado, que podem ser processados por computadores para uso em pesquisas linguísticas. Para o treinamento dos modelos foi utilizado o corpus ToLD-Br (LEITE, 2020), este corpus é formado por um conjunto de dados com 18.510 tweets em português brasileiro anotados de acordo com diferentes aspectos tóxicos. Os tweets foram coletados de forma automática entre julho e agosto de 2019 e anotados de forma manual por 129 voluntários que classificaram a linguagem contida nos textos como tóxica, anotados com 1, ou não tóxica, anotados com 0, levando em considerando insultos como LGBTQIA-fobia, obscenidade, insultos, racismo, misoginia e xenofobia. Ao todo foram anotados 9.248 tweets com linguagem não tóxica e 9.262 com linguagem tóxica. A Figura 1 apresenta a distribuição balanceada das classes tóxico e não tóxico dos *tweets* presentes no corpus ToLD-Br.

Figura 1 – Distribuição das classes no corpus ToLD-Br



Para os experimentos realizados nesta pesquisa foram definidos conjuntos estratificados de

treinamento e teste levando em consideração a proporção e distribuição da quantidade de tweets com linguagem tóxica e não tóxica. Para o conjunto de treinamento foram utilizados 90% do total de tweets, já os 10% restantes foram utilizados para testar o desempenho dos modelos treinados.

O corpus foi preprocessado para remover *hashtags*, *retweets*, símbolos e caracteres desnecessários que poderiam prejudicar a classificação das frases. Em seguida, foi utilizado um tokenizador de *tweets* da biblioteca NLTK[9] para tokenizar as frases, separando símbolos e abreviações junto com os termos, e removendo espaços desnecessários.

De acordo com a Comissão Especial de Processamento de Linguagem Natural da Sociedade Brasileira de Computação (SBC)[10], o Processamento de Linguagem Natural (PLN) trabalha com problemas relacionados à automação da interpretação e da geração da língua humana. Esse processamento geralmente envolve a tradução da língua natural em dados (números) que um computador pode usar para aprender sobre o mundo. Para a conversão dos textos dos tweets em padrões numéricos foram utilizados os modelos *bag of words* (SARKAR, 2019) que transforma documentos em uma coleção de palavras, desconsiderando a gramática e a ordem das palavras e *Term Frequency Inverse Document Frequency* (TF-IDF) (SARKAR, 2019) que computa uma pontuação de cada palavra para significar a sua importância ao longo do corpus.

Por meio da análise dos resultados obtidos através dos testes realizados com os modelos de IA implementados para a tarefa de detecção de linguagem ofensiva, realizou-se um estudo comparativo entre estes modelos a fim de avaliar quais são mais eficientes na execução da tarefa para o qual foram treinados.

RESULTADOS E DISCUSSÃO

Os resultados obtidos durantes os testes com os algoritmos de AM foram organizados e tabelados para que fossem analisados. A Tabela 2 apresenta os resultados obtidos com cada algoritmo de AM testado utilizando o *Bag of words*. Já na Tabela 3 estão apresentados os valores obtidos com o TF-IDF. Em ambas as tabelas é possível observar

9 https://www.nltk.org/
10 https://sites.google.com/view/ce-pln/

a precisão na predição de cada uma das classes, 0 para não tóxica e 1 para tóxica, e a acurácia do modelo treinado.

Tabela 2 – Resultados obtidos com cada modelo utilizando *Bag of words*

Classificador	Classes	Precisão	Acurácia
SGD	0 1	0.81 0.71	**0.75**
DecisionTree	0 1	**0.84** 0.67	0.74
RandomForest	0 1	0.81 0.71	**0.75**
SVM linear	0 1	0.76 **0.72**	0.74
NaiveBayes	0 1	0.80 0.64	0.70
MLP	0 1	0.75 0.70	0.73

Tabela 3 – Resultados obtidos com cada modelo utilizando TF-IDF

Classificador	Classes	Precisão	Acurácia
SGD	0 1	0.76 0.73	0.74
DecisionTree	0 1	**0.82** 0.67	0.73
RandomForest	0 1	0.81 0.71	**0.75**
SVM linear	0 1	0.76 **0.72**	0.74
NaiveBayes	0 1	0.78 0.66	0.72
MLP	0 1	0.68 0.71	0.69

Ao analisar os valores de precisão e acurácia dos modelos treinados com *Bag of Words,* pode-se perceber que o modelo implementado com o algoritmo *DecisionTree* obteve uma precisão de 84% na predição de tweets com linguagem não tóxica, contudo a precisão para os tweets com linguagem tóxica foi de apenas 62% fazendo com que a acurácia deste modelo alcançasse os 74% de acerto. Olhando apenas para a precisão de

tweets com linguagem tóxica, o modelo baseado no algoritmo SVM linear foi o que teve uma maior precisão com 72% de acerto.

O desempenho dos modelos treinados com TF-IDF foram ligeiramente menores , contudo, como apresenta a Tabela 3, os algoritmos treinados com TF-IDF que obtiveram os melhores resultados foram os mesmo que obtiveram os melhores desempenhos quando treinados com *Bag of Words*.

Ao analisar apenas a acurácia dos modelos, aquele implementado com o algoritmo *RandonForest* mostrou-se ter os melhores resultados, tanto com *Bag of Words* como com TF-IDF. Tal resultado pode-se justificar pelo fato deste algoritmo ajustar uma série de classificadores de árvore de decisão em várias subamostras do conjunto de dados e usa a média para melhorar a precisão preditiva e controlar o sobreajuste fazendo com que o seu desempenho seja melhor do que os outros testados.

Em todos os experimentos, os modelos implementados obtiveram um melhor desempenho da predição de tweets sem linguagem tóxica no corpus de teste e, tendo em vista que o número de amostras com linguagem tóxia e não tóxica, tanto no corpus de treinamento quanto no corpus de teste, é consideralvemente bem balanceado, a capacidade de predição destes modelos pode ser vista como boa.

CONCLUSÃO

Apesar de atualmente existirem tecnologias que utilizam aprendizado de máquina profundo, ou *deep learning*, com desempenho que atingem o estado da arte em tarefas de PLN, esta pesquisa evidenciou que os modelos tradicionais de AM ainda apresentam um bom desempenho em tarefas complexas como a identificação de linguagem tóxica em textos na internet. O uso destes modelos torna-se vantajoso pois são leves e rápidos em relação ao grande modelos de linguagem que utilizam aprendizado de máquia profundo e podem ser uma alternativa para pequenas aplicações web que buscam trazer uma maior segurança em sessões de comentários para seus usuários. Contudo, vale ressaltar que os resultados obtidos aqui podem vir a ser melhorados por meio da melhoria dos hiperparâmetros dos modelos experimentados ou o uso de outros algoritmos tradicionais de IA.

AGRADECIMENTOS

Agradecemos ao Instituto Federal de Educação, Ciência e Tecnologia da Maranhão (IFMA) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ) pelo apoio financeiro para a realização desta pesquisa.

REFERÊNCIAS

ALLAN, t. B. R. Hard Questions: Hate Speech. 2017. Disponível em: <https://newsroom.fb.com/news/2017/06/hard-questions-hate-speech/>. Acessado em: 04/04/2023

BISPO, T. D. Arquitetura LSTM para classificação de discursos de ódio cross-lingual Inglês-PtBR. 2018. 73 f. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Sergipe, São Cristóvão, SE, 2018.

DE ALMEIDA, M. e BERTON, L., 2020. Detecção Automática de Discurso de Ódio em Redes Sociais. Trabalho de Conclusão de Curso. Universidade Federal de São Paulo.

GAMA, J. et al. Inteligência Artificial - Uma Abordagem De Aprendizado de Máquina. [S.l.]: LTC, 2021.

GÉRON, A. Mãos à Obra: Aprendizado de Máquina com Scikit-Learn e TensorFlow. [S.l.]: Alta Books, 2019.

LEITE, J. A. et al. "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* (pp. 914–924). Association for Computational Linguistics, 2020.

NOBATA, C. et al. Abusive language detection in online user content. In: INTERNATIONAL WORLD WIDE WEB CONFERENCES STEERING COMMITTEE. Proceedings of the 25th International Conference on World Wide Web. [S.l.], 2016. p. 145–153

PINHEIRO, R. Crimes de ódio na internet tiveram aumento de quase 70% no primeiro semestre. Disponível em: <https://www12.senado.leg.br/radio/1/noticia/2022/10/10/crimes-de-odio-na-internet-tiveram-aumento-de-quase-70-no-primeiro-semestre>. Acessado em: 05/04/2023

SARDINHA, T. Linguística de corpus. [S.l.]: Manole, 2004.

SARKAR, D. Text Analytics with Python: A Practitioner's Guide to Natural Language Processing. 2nd. ed. [S.l.]: APress, 2019.

ZAMPIERI, Marcos et al. Predicting the Type and Target of Offensive Posts in Social Media. In: PROCEEDINGS of NAACL. [S.l.: s.n.], 2019.