CBIS'24 XX Congresso Brasileiro de Informática em Saúde
08/10 a 11/10 de 2024 - Belo Horizonte/MG - Brasil

Detecção de discurso de ódio para o apoio à saúde mental

Hate speech detection for mental health support

Detección de discurso de odio para apoyo a la salud mental

Ítalo Santos de Oliveira[1], Rodrigo Rafael Villarreal Goulart[2]

1 Bacharelando, Universidade Feevale, Instituto de Ciências Criativas e Tecnológicas, Universidade Feevale, Novo Hamburgo (RS), Brasil.
2 Pesquisador, Instituto de Ciências Criativas e Tecnológicas, Universidade Feevale, Novo Hamburgo (RS), Brasil.

Autor correspondente: Ítalo Santos de Oliveira
E-mail: italo.oli12@gmail.com
Links (opcional): https://github.com/italosdoliveira/hate-speech-detection-with-lstm

Resumo

Objetivo: Este artigo visa explorar a classificação de textos extraídos de comentários de redes sociais que contêm linguagem ofensiva e discurso de ódio. As interações em redes sociais com este viés podem ter efeitos prejudiciais à saúde mental da população. Método: Utilizamos técnicas de Processamento de Linguagem Natural e Aprendizado de Máquina, aplicando-as a um conjunto de dados brasileiro. Investigamos o uso de embeddings, o emprego de redes neurais Long Short Term Memory (LSTM) e uma abordagem híbrida com Convolutional Neural Network (CNN). A análise inclui a avaliação do desbalanceamento de dados e a aplicação de técnicas de undersampling e oversampling. Resultados e conclusão: A otimização da LSTM resultou em ganhos modestos, sendo mais eficaz quando combinada com a CNN, especialmente com oversampling. No entanto, este último gera preocupações de overfitting. Os resultados indicam que o modelo desenvolvido é mais confiável para a detecção de linguagem ofensiva do que para o discurso de ódio.

Descritores: Ódio, Inteligência Artificial; Processamento de Linguagem Natural

Objective: This article aims to explore the classification of texts extracted from social media comments containing offensive language and hate speech. Interactions on social networks with this bias can have harmful effects on the population's mental health. Method: We used Natural Language Processing and Machine Learning techniques, applying them to a Brazilian dataset. We investigated the use of embeddings, the deployment of Long Short-Term Memory (LSTM) neural networks, and a hybrid approach with Convolutional Neural Network (CNN). The analysis includes evaluating data imbalance and applying undersampling and oversampling techniques. Results and conclusion: LSTM optimization resulted in modest gains, being more effective when combined with CNN, especially with oversampling. However, the latter raises concerns about overfitting. The results indicate that the developed model is more reliable for detecting offensive language than hate speech.
Keywords: Hate, Artificial Intelligence; Natural Language Processing

Resumen

Objetivo: Este artículo tiene como objetivo explorar la clasificación de textos extraídos de comentarios en redes sociales que contienen lenguaje ofensivo y discurso de odio. Las interacciones en redes sociales con este sesgo pueden tener efectos perjudiciales para la salud mental de la población. Método: Utilizamos técnicas de Procesamiento del Lenguaje Natural y Aprendizaje Automático, aplicándolas a un conjunto de datos brasileños. Investigamos el uso de embeddings, el empleo de redes neuronales Long Short-Term Memory (LSTM) y un enfoque híbrido con Convolutional Neural Network (CNN). El análisis incluye la evaluación del desequilibrio de datos y la aplicación de técnicas de submuestreo y sobremuestreo. Resultados y conclusión: La optimización de LSTM resultó en ganancias modestas, siendo más efectiva cuando se combina con CNN, especialmente con sobremuestreo. Sin embargo, este último plantea preocupaciones sobre el sobreajuste. Los resultados indican que el modelo desarrollado es más confiable para detectar lenguaje ofensivo que discurso de odio.
Descriptores: Odio, Inteligencia Artificial; Procesamiento de Lenguaje Natural

Introdução

A expansão constante das redes sociais nos últimos anos contribuiu para que as pessoas falem abertamente o que pensam por meio de comentários em publicações. Infelizmente alguns desses comentários podem ser classificados como discursos de ódio, devido a um conjunto de características e elementos que eles carregam no seu conteúdo, empregando uma linguagem ofensiva e agressiva para com outros participantes, principalmente nas redes sociais. Este tipo de comentário tem se tornado cada vez mais comum, como podemos verificar, por exemplo, no período eleitoral, onde houve uma forte discussão e grande volume de ofensas nas redes sociais, potencializadas, inclusive, por *fake news*.

Fortuna e Nunes[1] trazem a seguinte definição, que tem como base a taxonomia proposta por Salminen et.al [2].:

> "O discurso de ódio é uma linguagem que ataca ou diminuí, que incite violência ou ódio contra grupos, com base em características específicas, como aparência, religião, descendência, nacionalidade ou etnia origem, orientação sexual, identidade de gênero ou outro, e pode ocorrer com diferentes estilos, mesmo em formas sutis ou quando o humor é usado."

De acordo com Nguyen [3], o discurso de ódio pode causar uma série de danos, incluindo mentais, emocionais, sociais e físicos, tanto para indivíduos quanto para a sociedade. Embora possa ser ignorado por membros de grupos majoritários, ele tem um impacto duradouro nos membros de grupos minoritários. O discurso de ódio pode gerar emoções negativas, como raiva, vergonha e medo, e até mesmo levar ao ódio internalizado. As vítimas podem enfrentar problemas de saúde mental, como transtorno de estresse pós-traumático, ansiedade, pensamentos suicidas e depressão. Além disso, pode afetar negativamente a autoestima e a dignidade de uma pessoa.

Além dessas repercussões negativas, o discurso de ódio é particularmente prejudicial nas comunidades universitárias. Saha et al.[4] apresenta um estudo sobre os efeitos psicológicos do discurso de ódio, analisando 6 milhões de comentários do Reddit compartilhados em 174 comunidades universitárias. A pesquisa caracterizou a resistência psicológica ao discurso de ódio, analisando a linguagem, o uso de palavras-chave discriminatórias e os traços de personalidade dos indivíduos.

Os resultados indicaram que o discurso de ódio é 25% mais prevalente em comunidades universitárias do que em comunidades não universitárias. O estudo também revelou que a exposição ao ódio leva a uma maior expressão de estresse. No entanto, nem todos os indivíduos expostos são igualmente afetados. Alguns indivíduos mostram menor resistência psicológica do que outros, sendo que aqueles com baixa resistência são mais vulneráveis a explosões emocionais e tendem a ser mais neuróticos do que aqueles com maior resistência.

Assim como a pesquisa de Saha et al. [4], que investigou as implicações do discurso de ódio em comunidades universitárias, este trabalho tem como objetivo colaborar com a formulação de políticas e esforços de intervenção para combater os efeitos prejudiciais do discurso de ódio online e promover o apoio à saúde mental.

Métodos

Como ponto de partida foi analisado o estudo de Vargas et al. [5] cuja proposta foi criar um *dataset* para treinar modelos nas tarefas de classificação de linguagem ofensiva e discurso de ódio, o *dataset* foi construído por comentários extraídos de perfis políticos brasileiros através da API do Instagram. Os comentários passaram por um processo de classificação manual feito por especialistas com grau superior de diferentes partes do Brasil. O corpus consiste em 7000 linhas ao total, podemos ver algumas linhas deste corpus já rotulado na Tabela 1. Este mesmo corpus construído pelos autores do artigo citado, foi utilizado durante os treinamentos dos modelos do presente artigo.

Das 7000 linhas do dataset, 3500 foram rotuladas com o valor 0 na coluna offensive_language e os restantes 3500 foram rotuladas com o valor 1 na coluna offensive_language. As respectivas linhas rotuladas com 1 em offensive_language também receberam um valor na coluna hate_speech, com esses valores variando entre –1, 1,2,3,4,5,6,7,8. Ao todo, 727 linhas receberam uma classificação de 1 a 8 e 2773 receberam –1, conforme descrito em Vargas et al. [5] as linhas que receberam o valor –1 na coluna hate_speech são linhas que não contém linguagem ofensiva e não contém discurso de ódio.

Tabela 1 - Resultados originais obtidos por Vargas et al.[5]

instagram_comments	offensive_language	offensiveness_levels	hate_speech
Este lixo ...	1	1	-1
Vagabunda. Comunista. Mentirosa. O povo chileno não merece uma desgraça desta.	1	3	"5,8"
Quem tem pena é galinha, mas ela é uma VACA LOUCA.	1	3	8
Essa foto ficou bonita	0	0	0

Para os experimentos a equipe utilizou as bibliotecas Scikit-learn e Pandas. Utilizou como *features* o resultado do emprego das técnicas de pré-processamento de textos TF-iDF e Unigramas, e para cada *feature* foram aplicados 4 algoritmos de Machine Learning (ML): Naive Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR) e Multi Layer Perceptron (MLP). Em relação aos dados no treinamento eles foram divididos da seguinte forma: 80% treinamento, 10% teste, 10% validação. Na tarefa de detecção de discurso de ódio, para passar os dados pelos algoritmos a equipe converteu os valores que eram menores ou iguais a 0 para 0 na coluna hate_speech e todos os valores maiores ou iguais a 1 viraram 1. Algumas linhas foram rotuladas com mais de um valor na coluna hate_speech, como o exemplo da Tabela 1 que o valor era "5,8", estas linhas com mais de um valor foram removidas e então foram geradas duas novas linhas com o mesmo comentário desta linha, neste caso uma linha com o valor 5 e outra com o valor 8, com isso o total de linhas ao final ficou 7025.

Como foco deste trabalho, os experimentos que foram realizados utilizaram recursos que não foram empregados por Vargas et al.[5], como por exemplo o uso da

feature Word of Embeddings. Além disso, o estudo propôs explorar técnicas de Aprendizado de Máquina, usando de Deep Learning por meio do emprego da rede neural LSTM, assim como foi feito por Fortuna et.al[6] e no trabalho de Badjatiya et al. [7]. Os autores apresentam um estudo que investiga a aplicação de métodos de aprendizado profundo para a tarefa de detecção de discurso de ódio em Tweets utilizando várias *features* e as redes neurais LSTM e Convolutional Neural Network (CNN). Este trabalho também desenvolveu um experimento empregando uma abordagem híbrida com as redes neurais CNN e LSTM, de acordo com o proposto por Garg et.al [8] em seu estudo e por Badjatiya et al. [7].

Este trabalho foi dividido em 4 experimentos, implementados na plataforma Codespace do Github, usando a linguagem Python com as bibliotecas Tensorflow, Keras e Scikit-learn.

O primeiro experimento teve como objetivo reproduzir o experimento de Vargas et al.[5] que após a construção do *dataset* HateBr, aplicaram os 4 algoritmos Naive Bayes, Logistic Regression, Support Vector Machine e Multi Layer Perceptron. A Tabela 2 apresenta os resultados que os autores obtiveram nos experimentos. Ela contém um recorte dos dados, trazendo apenas o F1-score do *average*, a média de todas as classificações, tanto para a categoria 0 quanto para categoria 1. Também apresenta os resultados obtidos para as tarefas de detecção de linguagem ofensiva e detecção de discurso de ódio. Para a primeira tarefa o melhor resultado foi obtido com o emprego da feature TF-iDF e o algoritmo Support Vector Machine. Para a segunda tarefa o melhor resultado também foi obtido com a feature TF-iDF, mas com o emprego do algoritmo Naive Baiyes.

O segundo experimento teve como objetivo utilizar uma técnica de balanceamento dos dados diferente da utilizada por Vargas et al.[5]. O primeiro experimento utilizou a técnica *undersampling*, que seleciona as linhas com a classe em maior quantidade e aleatoriamente exclui essas linhas até que esta tenha o mesmo tamanho da classe com menos linhas. O segundo experimento empregou a técnica de *oversampling*, que seleciona as linhas da classe em menor quantidade e aleatoriamente as duplica até que tenha a mesma quantidade das linhas da classe com maior quantidade.

Tabela 2 - Resultados originais obtidos por Vargas et al.[5]

Tarefa	Feature	Algoritmos			
		NB	SVM	MLP	LR
Detecção de Linguagem ofensiva	ngrams	0.75	0.80	0.84	0.84
	tfidf	0.77	**0.85**	0.84	0.84
Detecção de discurso de ódio	ngrams	0.76	0.69	0.77	0.77
	tfidf	0.78	0.76	0.77	0.77

O terceiro e quarto experimentos tiveram como objetivo o emprego da rede neural LSTM e da rede CNN somada à rede LSTM. Como ponto de partida para a implementação destes experimentos foi utilizada a proposta apresentada por Fortuna et.al[6], que determina uma estrutura composta uma camada LSTM com 64 neurônios, rodando em 10 épocas, com a função de ativação sigmoid, otimizador adam. Como função de perda este trabalho utilizou uma função diferente da utilizada por Fortuna et.al[6]. Uma função de perda de entropia cruzada binária foi utilizada em razão da forma que os dados foram rotulados. Os dados contêm dois rótulos, um para a tarefa de detecção de linguagem ofensiva com os valores 0 e 1 e outro rótulo para a tarefa de detecção de discurso de ódio também com os valores 0 e 1 e por estes valores que a entropia cruzada binária foi utilizada. Estes dois últimos experimentos além de fazerem uso de redes neurais também tiveram o uso da feature Word of Embeddings com tokens em português que foram extraídas utilizando o algoritmo Global Vector (GloVe) que foram resultado do trabalho de Hartmann et.al [9].

Foram também empregadas otimizações na implementação da Long Short Term Memory (LSTM) seguindo os valores de referência propostos por, que foram baseadas no estudo de Rajalaxami et.al [10].Por fim, a estrutura ficou com uma camada LSTM com 128 neurônios, uma *hidden layer* com 512 neurônios e 20 épocas, a segunda estrutura derivada desta ficou semelhante, mas antes de passar pela camada LSTM de 128 neurônios, os dados terão que passar por uma camada CNN com 64 neurônios e com a função de ativação relu,.

Resultados e Discussão

Na Tabela 3 podemos observar os resultados da reprodução dos experimentos de Vargas et al.[5] neste trabalho. As Tabelas 3, 4 e 5 trazem apenas o resultado F1-score do média de todas as classificações de 0 e 1.

Tabela 3 - Resultados da reprodução dos experimentos do artigo base

Tarefa	Feature	Algoritmos			
		NB	SVM	MLP	LR
Detecção de Linguagem ofensiva	ngrams	0.74	**0.80**	0.79	0.86
	tfidf	0.79	0.84	0.84	0.85
Detecção de discurso de ódio	ngrams	0.70	**0.69**	0.78	0.75
	tfidf	0.73	0.70	**0.77**	**0.77**

Observa-se na Tabela 3, que os resultados destacados em negrito com amarelo são resultados que chegaram próximos dos resultados obtidos pelos autores. No entanto, ainda há uma pequena diferença em relação aos resultados anteriores, com uma variação de 0.01 a 0.02, o que neste trabalho foi considerado admissível e dentro do resultado esperado. A separação aleatória dos dados para o processo de treinamento e validação pode gerar essa pequena variação. Por outro lado, os resultados na tarefa de detecção de discurso de ódio, utilizando o algoritmo NB, produziu um F1-score com uma diferença de 0.05 a 0.06 do que foi obtido

originalmente. Os resultados destacados em verde e sublinhados são os resultados que tiveram um resultado igual ao do artigo base.

Na Tabela 4 é possível encontrar os resultados do segundo experimento que teve como propósito balancear os dados do treinamento com *oversampling* e submeter esses dados pelos mesmos 4 algoritmos utilizados anteriormente.

Estes resultados apresentam uma melhora em comparação aos obtidos no experimento 1, se olharmos com atenção podemos ver que o algoritmo que teve o pior resultado foi o Naive Bayes, enquanto o que teve melhor resultado foi o do Support Vector Machines.

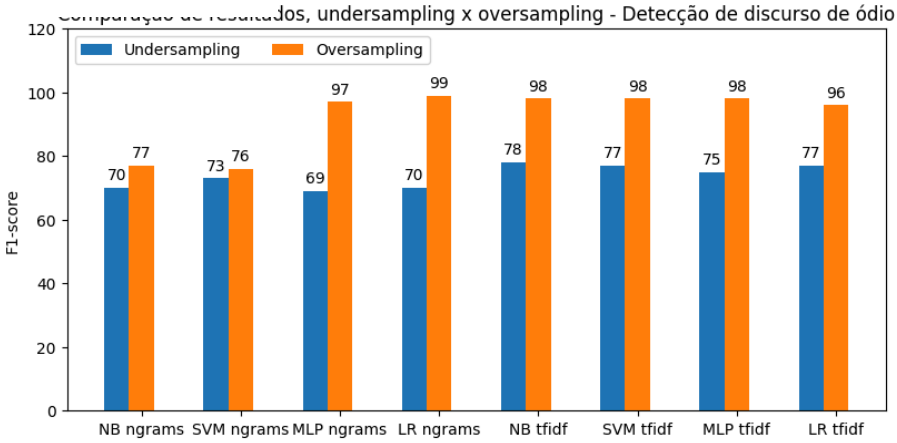
Tabela 4 - Resultados da reprodução dos experimentos do artigo base, utilizando *oversampling*

Tarefa	Feature	Algoritmos			
		NB	SVM	MLP	LR
Detecção de discurso de ódio	ngrams	0.77	0.97	0.98	0.98
	tfidf	0.76	**0.99**	0.98	0.96

No entanto existem questões a serem discutidas sobre a implementação da *oversampling*, já que ela apenas faz uma cópia das linhas que estão com menor quantidade em relação a outra classe, isso faz com o modelo seja treinado com dados duplicados o que pode fazer com que este venha a ter algum viés na sua tarefa de classificação na ausência de uma diversidade maior de dados.

O Gráfico 1 apresenta a comparação dos resultados na tarefa de detecção de discurso de ódio a qual foi necessário um balanceamento, mostrando uma comparação de usar *undersampling* e *oversampling*. Para os experimentos da tarefa de classificação de linguagem ofensiva não foi necessário aplicar nenhuma técnica de balanceamento porque os dados já estavam perfeitamente balanceados.

Gráfico 1 - Comparação resultados dos algoritmos base com *undersampling* e *oversampling*



Na comparação das duas técnicas de balanceamento apresentadas no Gráfico 1, a *oversampling* apresenta o melhor resultado. No entanto, existem questionamentos sobre sua aplicação num treinamento em razão da dúvida se houve ou não um *overfiting* durante o treinamento. O *dataset* na tarefa de discurso de ódio apresenta três vezes mais casos da classe 0 que a classe 1. Os resultados utilizando a técnica de *undersampling* se mostram um pouco mais consistentes onde os resultados dos treinamentos chegam a resultados muito parecidos em alguns casos. Uma avaliação da existência ou não do viés poderá ser analisada em trabalhos futuros, com a inclusão de mais dados no *dataset*.

Por fim, os resultados dos experimentos 3 e 4, utilizando a rede neural LSTM e depois uma abordagem híbrida de CNN com LSTM, são apresentados na Tabela 5 e no Gráfico 2.

Tabela 5 - Resultados dos experimentos utilizando redes neurais

Tarefa	Feature	LSTM	LSTM + CNN
Detecção de Linguagem ofensiva	embeddings	0.86	0.96
Detecção de discurso de ódio sem balancear o dataset	embeddings	0.57	0.88
Detecção de discurso de ódio com o dataset balanceado com undersampling	embeddings	0.69	1
Detecção de discurso de	embeddings	0.96	1

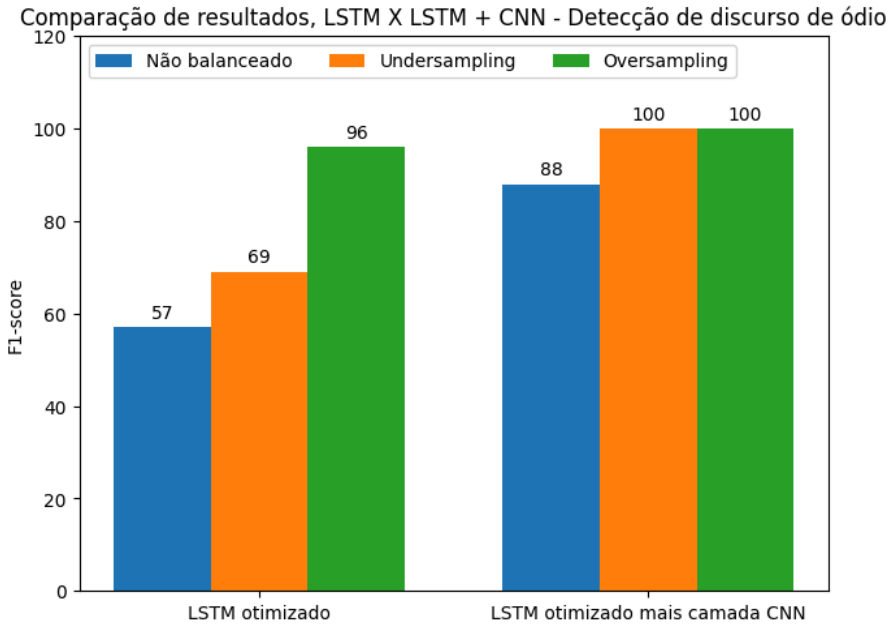
ódio com o dataset balanceado com oversampling

Percebe-se que os melhores resultados obtidos são sempre encontrados no segundo grupo, que diz respeito aos resultados obtidos em um experimento onde primeiro os dados passam por uma camada CNN, só depois indo para a camada LSTM. Mesmo com os dados desbalanceados como na tarefa de discurso de ódio, foi encontrado um resultado melhor do que o experimento que emprega somente uma camada LSTM.

Vale notar que quando não se utiliza essa abordagem híbrida o resultado de empregar apenas LSTM é semelhante ao que foi obtido nos outros 4 algoritmos.

Mesmo a implementação da LSTM tendo sido submetida por um processo de otimização, o resultado só se mostrou mais vantajoso quando os dados foram balanceados utilizando a técnica de *oversampling*.

Gráfico 2 - Comparação de resultados da LSTM com 3 tipos de balanceamento



Apesar dos resultados serem satisfatórios nos experimentos, ainda é necessária atenção a duas questões que precisam ser consideradas referentes aos dados que foram extraídos. A primeira questão é o contexto dos dados. Estes foram

extraídos do perfil do Instagram de políticos brasileiros. Avaliando alguns dos comentários contidos no *dataset* é possível questionar se os modelos gerados no treinamento vão ser mais eficientes em contextos de textos que envolvam política e menos eficiente em textos que são extraídos de outros contextos. Esta é uma das questões que pode ser explorada em trabalhos futuros.

A segunda questão é o desbalanceamento dos dados rotulados com alguma categoria que indique que ele contém ou não algum tipo de discurso de ódio, ao todo 727 comentários continham algum tipo de discurso de ódio e 2773 não continham. Isso faz com que haja uma probabilidade maior de que os modelos que foram gerados para identificar linguagem ofensiva tenham mais confiança do que os modelos construídos para identificar discurso de ódio. O desbalanceamento nesta categoria que levou a utilizar as técnicas *undersampling* conforme os experimentos de Vargas et al.[5] e *oversampling*, que foi proposta neste trabalho. A técnica de *oversampling* talvez não tenha sido o melhor neste caso, já que a quantidade de comentários que de fato continham algum discurso de ódio eram cerca de três vezes menores que as quantidade de linhas que não continham algum discurso de ódio, gerando a possibilidade da técnica *oversampling* ter duplicado pelo menos três vezes os dados.

Conclusão

O crescimento da internet contribuiu para que mais pessoas compartilhassem suas ideias que atacam grupos específicos, estas ideias podem ser classificadas como discurso de ódio. O discurso de ódio, definido como comentários odiosos dirigidos a um grupo ou alvo específico, tem reflexo na saúde mental da população e é proibido por lei no Brasil. Com o crescimento desse tipo de interação em redes sociais são necessários meios para classificar textos que contenham discurso de ódio e linguagem ofensiva. Desse contexto surge a necessidade de a Inteligência Artificial ser utilizada como uma ferramenta para este trabalho.

Os autores dos artigos analisados destacam a dificuldade de encontrar *datasets* em português que possam ser utilizados para o treinamento de modelos de *Machine Learning*. Desta forma os autores se dedicaram a construir seu próprio *dataset* extraindo textos do Instagram e do Twitter. Com este corpus anotado os

autores e submeteram a uma classificação binária, utilizando vários algoritmos diferentes de ML.

Este trabalho teve como objetivo reproduzir os experimentos já feitos usando um *dataset* brasileiro rotulado com textos que contenham ou não linguagem ofensiva e discurso de ódio. Foram discutidas algumas questões comparativas entre as duas técnicas, como a forma que a *oversampling* faz o balanceamento dos dados. Devido à quantidade de dados disponíveis, talvez os algoritmos que utilizam redes neurais não tenham atingido seu melhor resultado.

Assim utilizando a CNN junto da LSTM é possível obter um resultado melhor que os algoritmos mais tradicionais, mesmo sem uma quantidade massiva de dados. Embora uma quantidade maior de dados possa se mostrar muito mais vantajosa quando falamos do contexto de redes neurais e talvez venha a alterar os resultados que foram obtidos ao longo do treinamento.

Referências

1. Fortuna P, Nunes S. A Survey on Automatic Detection of Hate Speech in Text. ACM Computing Surveys. 2019;51(4):1-30.
2. Salminen J, Almerekhi H, Milenković M, Jung S-G, An J, Kwak H, et al. Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media. Proceedings of the International AAAI Conference on Web and Social Media. 2018;12(1).
3. Nguyen T. Merging public health and automated approaches to address online hate speech. AI and Ethics. 2023.
4. Saha K, Chandrasekharan E, Choudhury MD. Prevalence and Psychological Effects of Hateful Speech in Online College Communities. Proceedings of the 10th ACM Conference on Web Science; Boston, Massachusetts, USA: Association for Computing Machinery; 2019. p. 255–64.
5. Vargas F, Carvalho I, Rodrigues de Góes F, Pardo T, Benevenuto F, editors. HateBR: A Large Expert Annotated Corpus of Brazilian Instagram Comments for Offensive Language and Hate Speech Detection2022 June; Marseille, France: European Language Resources Association.
6. Fortuna P, Rocha Da Silva J, Soler-Company J, Wanner L, Nunes S, editors. A Hierarchically-Labeled Portuguese Hate Speech Dataset. Proceedings of the Third Workshop on Abusive Language Online; 2019 2019-01-01: Association for Computational Linguistics.

7. Badjatiya P, Gupta S, Gupta M, Varma V, editors. Deep Learning for Hate Speech Detection in Tweets. Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion; 2017 2017-01-01: ACM Press.
8. Garg M, Saxena C, Saha S, Krishnan V, Joshi R, Mago V, editors. CAMS: An Annotated Corpus for Causal Analysis of Mental Health Issues in Social Media Posts2022 June; Marseille, France: European Language Resources Association.
9. Hartmann N, Fonseca E, Shulby C, Treviso M, Rodrigues J, Aluisio S. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. arXiv preprint arXiv:170806025. 2017.
10. Rajalaxmi RR, Prasad LVN, Janakiramaiah B, Pavankumar CS, Neelima N, Sathishkumar VE. Optimizing Hyperparameters and Performance Analysis of LSTM Model in Detecting Fake News on Social media. ACM Trans Asian Low-Resour Lang Inf Process. 2022.