INSTITUTO FEDERAL
DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
Bahia | Campus Valença

MINISTÉRIO DA EDUCAÇÃO
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA BAHIA
CAMPUS VALENÇA

ANA CLARA DOS SANTOS NASCIMENTO, EDKELLY CONCEIÇÃO SANTOS, IZABELLY OLIVEIRA DA SILVA, LÁZARO MOTA MARQUES DE JESUS, MATHEUS HENRIQUE MATIAS ANDRADE

Reconhecimento de discursos de ódio em mensagens de texto em língua portuguesa

Valença-BA
2025

ANA CLARA DOS SANTOS NASCIMENTO, EDKELLY CONCEIÇÃO SANTOS, IZABELLY OLIVEIRA DA SILVA, LÁZARO MOTA MARQUES DE JESUS, MATHEUS HENRIQUE MATIAS ANDRADE

Reconhecimento de discursos de ódio em mensagens de texto em língua portuguesa

Trabalho de Conclusão de Curso do Curso Integrado de Informática, do Instituto Federal de Educação, Ciência e Tecnologia da Bahia, como requisito parcial para a obtenção do título de Técnico em Informática.

Orientador: Prof. Me. Rodolfo Costa Cezar da Silva

Valença-BA
2025

ANA CLARA DOS SANTOS NASCIMENTO, EDKELLY CONCEIÇÃO SANTOS, IZABELLY OLIVEIRA DA SILVA, LÁZARO MOTA MARQUES DE JESUS, MATHEUS HENRIQUE MATIAS ANDRADE

Reconhecimento de discursos de ódio em mensagens de texto em língua portuguesa

A banca examinadora, abaixo listada, **aprova** o Trabalho de Conclusão de Curso "Reconhecimento de discursos de ódio em mensagens de texto em língua portuguesa" elaborado por "ANA CLARA DOS SANTOS NASCIMENTO, EDKELLY CONCEIÇÃO SANTOS, IZABELLY OLIVEIRA DA SILVA, LÁZARO MOTA MARQUES DE JESUS, MATHEUS HENRIQUE MATIAS ANDRADE" como requisito parcial para obtenção do grau de Técnico em Informática, pelo Instituto Federal de Educação, Ciência e Tecnologia da Bahia.

Resultado : _____ - **Nota :** _____

Valença-BA, 14/02/2025

Comissão Examinadora

Prof. Me. Rodolfo Costa Cezar da Silva
Instituto Federal de Educação, Ciência e Tecnologia da Bahia - Campus Valença
(Orientador)

Banca 1
IFBA - Campus Valença

Banca2
IFBA - Campus Valença

Reconhecimento de discursos de ódio em mensagens de texto em língua portuguesa

Resumo

resumo do trabalho.

Palavras-chave: Libras. Reconhecimento de Gestos. Visão Computacional. Aprendizado de Máquina. Inclusão Educacional.

NOME DO TRABALHO EM INGLES

Abstract

resumo em ingles

Keywords: palavras-chave

Sumário

1 Introdução

No contexto atual onde as redes sociais se tornaram um grande meio de comunicação na Internet, é possível ver uma vasta quantidade de debates e opiniões pessoais sobre diversos assuntos que cercam nossa sociedade. Por conta disso, é notável disseminação de ódio através desses comentários. Cada vez mais é comum vermos ataques direcionados a pessoas por conta de suas características como raça, gênero, etnia, nacionalidade e outras (PAIVA et al., 2019). Essa disseminação de ódio tem como seu conceito o discurso de ódio, que segundo Cohen-Almagor [2011], se caracteriza por ser um discurso perverso, cruel, hostil e preconceituoso direcionado a grupos específicos por conta de gênero, etnia, religião, nacionalidade, deficiência física ou mental, orientação sexual e condicionamento físico. Uma mensagem com discurso de ódio é definida assim ao possuir palavras de ódio ou cunho depreciativo (SILVA; SERAPIÃO, 2018).

É evidente cada vez mais os discursos de ódio estarem presentes nas redes sociais, que conseguem ter um largo alcance, um volume exacerbado, e uma velocidade que acabam permeando o mundo todo e as pessoas que são alvo desses ataques.

A decisão de utilizar uma detecção automática para reconhecer discursos de ódio e não fazer um trabalho manual, se consolida pelo fato de que o volume e velocidade de informações que são trocadas na internet, se faz necessário o uso de técnicas informatizadas. Com isso em vista, esse trabalho utilizará da detecção automática por meio de Algoritmos de Aprendizagem de Máquina. Esses algorítimos têm obtido êxito em cenários desse tipo, devido à capacidade de "aprender" a partir de exemplos reais obtidos do ambiente em que irão atuar, de forma que padrões possam ser reconhecidos (PAIVA et al., 2019). Apesar de que os modelos criados a partir desses algorítimos não serem cem por cento precisos, eles conseguem lidar com esse volume exacerbado de dados na internet sem muita intervenção humana.

Elaborar discussão sobre limites de discurso de ódio e liberdade de expressão....

Nesse trabalho, utilizamos diversas coleções de dados em língua portuguesa disponíveis na literatura, as quais variam em número de classes, quantidade de instâncias e balanceamento entre as categorias. Para cada método de classificação aplicado a esses conjuntos de dados, avaliamos o desempenho por meio das métricas de acurácia e F1-score.

O trabalho é organizado da seguinte forma: No Capítulo 2 é apresentado uma breve revisão da literatura. No Capítulo 3 são apresentados os métodos e metodologias utilizadas no desenvolvimento do trabalho. No Capítulo 4 são apresentados os resultados obtidos, seguidos por uma conclusão e apresentação de trabalhos futuros no Capítulo 5.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Trabalhos Relacionados

Os autores em (SILVA; SERAPIãO, 2018) utilizaram técnicas de aprendizado profundo, mais especificativamente, Redes Neurais Convolucionais (CNN) para a detecção de discurso de ódio em língua portuguesa, se embasando nas seguintes bases de dados: OffComBr, OffComBr-3 (PELLE; MOREIRA, 2017a), OffComBr-2 (PELLE; MOREIRA, 2017a), Hate Speech Dataset (FORTUNA, 2017). Além das técnicas de classificação, empregaram diferentes técnicas de vetorização de palavras, tais como: Wang2Vec e GloVe (ambos com dimensões de: 0, 100 e 300) combinando quatro diferentes métodos de otimização (RMSprop, Adagrad, Adadelta e Adam). Executando 84 cenários, cada um com configuração única, propuseram analisar as coleções utilizando as métricas: F-score, Acurácia, respectivamente, de até 82,64% e 0.89 para a coleção de dados OffComBr-2, 92,82% e 0.96 para OffComBr-3 e 92,74% e 0.96 HSD(Hate Speech Dataset).

(HELDE et al., 2025) , os autores utilizaram técnicas de aprendizado profundo, mais especificativamente, LSTM (Long Short- Term Memory) e RNNs (Recurent Neural Networks) para analisar e classificar diferentes tipos de discurso de ódio, se embasando nas seguintes bases de dados: Weitzel (WEITZEL et al., 2023), ToLD-Br (LEITE et al., 2020), HateBR (VARGAS et al., 2022) e Jigsaw Toxic Comment Classification Dataset[1], utilizando classificadores Multilabel e Multiclasse para coleções de dados em português, denominados MultilabelPT e MulticlassePT, respectivamente. Foram desenvolvidos também classificadores multilabel e multiclasse para as coleções de dados em inglês, denominados MiltilabelEN e MulticlasseEN, respectivamente. Esse trabalho foi avaliado utilizando as métricas: acurácia, F1-Score e AUC-ROC (Area Under Curve –Receiver Operating Characteristic), com o objetivo de avaliar, sob diferentes óticas, os modelos. Por fim, obtiveram como resultados principais: 71% de acurácia e 70,4% de F1-Score no MultilabelPT; 90%(acurácia) e 83% (F1-Score) no MultilabelEN; 68% (acurácia)e 64,6% (F1-Score) em MulticlassePT; e 71% (acurácia) e 68,7% (F1-Score) em MulticlasseEN.

Em (PALMEIRA, 2023), os autores implementaram e compararam algoritmos de aprendizado de máquina e processamento de linguagem natural que são capazes de identificar discursos de ódio na Rede Social Twitter. Para isso, os autores selecionaram os algoritmos Máquina Vetor de Suporte (SVM), Regressão Logística, *Random Forest* e Naive Bayes e criaram uma matriz TFIDF (Term Frequency–Inverse Document Frequency) para o treinamento dos algoritmos. Ess matriz é utilizada para representação de documentos, e utiliza a frequência de um termo (*Term Frequency* - TF) e combina com a frequência

[1] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data

inversa do documento (*Inverse Document Frequency* - IDF), e indica, para cada termo, a importância naquele documento. Utilizando as coleções de dados OffComBR (PELLE; MOREIRA, 2017a) (para o treino dos algoritmos) e API do X[2] e a métrica acurácia, os autores registraram os resultados de acurácia de 78% com Regressão Logística, 77% com *Random Forest*, 68% com Naive Bayes e 79% com SVM.

O trabalho descrito em (SILVA; SANTOS, 2024), analisou publicações na rede social X destinadas a candidatos que disputaram o segundo turno das eleições municipais de 2020 no Brasil, para criar um classificador *Naïve Bayes* através do processo CRISP-DM (CRoss Industry Standard Processfor Data Mining) e identificar discursos de ódio. Nele, foi utilizado para treinamento e avaliação do modelo, os dados coletados a partir de um levantamento de publicações realizadas na rede social X e o repositório Kaggle. A metodologia adotada foi a quantitativa, a partir de métricas estatísticas para assertividade, em particular acurácia, precision, recall e f1-score. Por fim, como resultado obteve resultados de 72,38% em acurácia, 70,73% em precision, 70,46% recalle e 70,74% em f1-score.

O autor em (BISPO, 2018) identifica o impacto da barreira linguística em tarefas de Processamento de Linguagem Natural (PLN) e propõe soluções para o problema identificado. O autor coleta comentários das redes sociais Facebook e Twitter e constrói um dataset em inglês, após essa etapa, ele treina um modelo deep cross-lingual com os métodos Long Short-Term Memory (LSTM) e Gradient Boosting Decision Tree (GBDT), utilizando bases de dados em inglês. Ao todo ele utiliza seis bases de dados para treinamento do modelo, sendo elas: discursos_votado (BISPO, 2018) ; discursos_votados_en (BISPO, 2018) ; NAACL_SRW_2016 (WASEEM; HOVY, 2016) , esta base corresponde a um conjunto de tweets rotulados com as três classes "sexism"(seximo/misoginia), "racism"(racismo) e "none"(nenhuma das classes anteriores), o autor converte a base originalmente multiclasse em uma base de classificação binária ao reduzir os comentários classificados em "sexism"e "racism"a comentários que contém discurso de ódio; NAACL_SRW_2016_pt (WASEEM; HOVY, 2016), é a forma traduzida através da API do Google Translate da base de dados "NAACL_SRW_2016"; NAACL_SRW_2016_cleaned_pt (WASEEM; HOVY, 2016), corresponde a base de dados "NAACL_SRW_2016" pré-processada e traduzida; dataset_portugues(FORTUNA, 2017). O autor obteve resultados de até 81,8% de precisão e 80,7% de cobertura e medida F.

O trabalho realizado em (PLATH et al., 2022) investigou a anonimização de perfis, que causa a sensação de liberdade irrestrita e o sentimento de impunidade que têm tornado as redes sociais um ambiente propício para o surgimento de discursos de ódio. Através de um processo de criação da base MINA BR, com mensagens retiradas das redes sociais: X, Facebook, Instagram e Youtube. Os dados coletados foram refinados pelos algoritmos

[2] https://x.com/

Máquina Vetor de Suporte SVM, Random Forest e Naive Bayes (Inserir quais foram as formas de vetorização palavras). A partir da métrica F-1 foi estabelecido como melhor resultado 0.57 obtido por SVM + TF IDF.

A proposta dos autores em (OLIVEIRA et al., 2023)

3 PROCEDIMENTOS METODOLÓGICOS

3.1 Sistema Proposto

Desenhar o pipeline para cada etapa

3.2 Pré-Processamento

Para o pré-processamento da parte textual, foi realizada uma estrategia de processamento genérica para o amplo funcionamento, isso é, para que funcionasse em todas as coleções de dados escolhidas. Porém, o processo manteve seu rigor envolvendo o alinhamento às práticas, já consolidadas, do processo conhecido pela comunidade acadêmica como linguagem natural (NLP). Primeiramente, foi feita uma normalização textual das coleções de dados a partir da conversão das letras às suas versões minúsculas, caso ainda não estivessem; remoção de acentos gráficos via decomposição da tabela Unicode; eliminação de caracteres alfabéticos, simbólicos e *hyperlinks*; por fim, reduziu-se as palavras às suas formas raízes através de lematização.

Podemos incluir, para exemplificar, uma Tabela mostrando textos da coleção de dados, e como ele fica após o pré-processamento.

Texto	**Texto após pré-processamento**
Boicote mediático ao PNR: crime no jornalismo https://t.co/GEobmnYdOD https://t.co/QBWrB6ByVy	boicote mediatico pnr crime jornalismo
Por isto ele usa o pseudonimo de SURREAL bidu	usa pseudonimo surreal bidu
’ temporadas pelo RALA e titulos kkkkkkkkkk Messi em anos de Barca tem titulos’	temporadas rala titulos kkkkkkkkkk messi ano barca titulos

Tabela 1 – Exemplo de tabela a ser criada na seção de pré-processamento

3.3 Coleções de dados

Inicialmente, as coleções de dados pré-selecionadas incluíram: ToLDBr (LEITE et al., 2020); ToLDBr-Binário (adaptado de (LEITE et al., 2020)); OLID-BR Está sem um citing em BIB na página oficial, o que é que eu faço?; Offcom2 (PELLE; MOREIRA, 2017b); OffComBR-3 (PELLE; MOREIRA, 2017b); hateBR (VARGAS et al., 2022); Eu ainda vou usar alguns conectivos e sinônimos, mas é assim que é pra ser feito? (com exeplo da ToLDBr...) O uso da coleção ToLDBr (coleção de dados de mensagens em português que contém 7 colunas, 21.000 instâncias, 6 colunas de classificação de categorias como categorias de discurso ofensivo: homophobia, obscene, insult, racism, misogyny e xenophobia) O uso da coleção OLID-BR [...] O uso da coleção OffcomBR-2 [...] O uso da coleção OffcomBR-3 [...] O uso da coleção hateBR [...]

Explicar todas as coleções de dados, falar sobre tamanho, classes, quantidade de instâncias por classe, binária/multiclasse, foi feito balanceamento? (como?)

Para cada coleção d dados vamos informar o nome, quantidade total de mensagens, distribuição de classes, quantidade de instâncias por classe (talvez não), se é binária ou não

	# de Instâncias	**Tamanho médio de Instância**	**# de Classes**	**Distribuição de Classes**	**Balanceamento Artificial?**
Coleção de Dados 1					
Coleção de Dados 2					
Coleção de Dados 3					
Coleção de Dados 4					
Coleção de Dados 5					
Coleção de Dados 6					

3.4 Métodos de classificação

Em termos de métodos de classificação, foram utilizados X classificadores : X, Y, ... Todo o código foi desenvolvido utilizando a biblioteca scikit-learn[1] e linguagem de programação Python.
O motivo da escolha desses classificadores foi a grande popularidade dos mesmos, e o uso deles em diversos trabalhos da literatura(Citar todos trabalhos que usam ALGUM desses métodos). Além disso, como as coleções de dados utilizadas não continham muitas instâncias, foi escolhido não trabalhar com algoritmos de redes neurais, que atualmente dominam o estado da arte na tarefa de classificação. O uso desses tipos de algoritmos, bem com estratégias para obter mais instâncias (*Oversampling*) serão exploradas em trabalhos futuros.

	Modelo 1	Modelo 2	Modelo 3
Acurácia	3.5	23	34
F1-Score	4	36	34

Tabela 2 – Esse é um teste de legenda

3.5 Métricas

[1] http://scikit-learn.org/

4 RESULTADOS

Para cada coleção de dados vamos utilizar uma tabela desta e uma matriz de confusão.

	Modelo 1	**Modelo 2**	**Modelo 3**	**Modelo 4**	**Modelo 5**
Acurácia					
F1-Score					

Tabela 3 – Resultados para a Coleção de Dados XXX(citação)

5 CONCLUSÃO

5.1 Limitações e Trabalhos Futuros

- GridSearch (Não utilizamos os melhores hiperparâmetros)
- uso de redes neurais (Técnicas mais utilizadas no estado da artes - motivo : poucas instâncias)
- oversampling - Aumentar coleção de dados
- cross-validation (Melhor avaliação do modelo de classificação)

Referências

BISPO, T. D. Arquitetura lstm para classificação de discursos de ódio cross-lingual inglês-ptbr. Pós-Graduação em Ciência da Computação, 2018.

FORTUNA, P. Automatic detection of hate speech in text: an overview of the topic and dataset annotation with hierarchical classes. **PQDT-Global**, Universidade do Porto (Portugal), 2017.

HELDE, R. V.; RICO, M.; PACHECO, R. G.; COUTINHO, E. R.; GUARDA, G. F.; WEITZEL, L. Análise e classificação de discurso de ódio online: uma abordagem multilabel e multiclasse com deep learning. **Brazilian Journal of Development**, v. 11, n. 5, p. e79603, May 2025. Disponível em: <https://ojs.brazilianjournals.com.br/ojs/index.php/BRJD/article/view/79603>.

LEITE, J. A.; SILVA, D.; BONTCHEVA, K.; SCARTON, C. Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In: WONG, K.-F.; KNIGHT, K.; WU, H. (Ed.). **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**. Suzhou, China: Association for Computational Linguistics, 2020. p. 914–924. Disponível em: <https://aclanthology.org/2020.aacl-main.91>.

OLIVEIRA, C. M. de; ARO, R. G. de; DARAKJIAN, V. M. Protótipo de sistema moderador de conteúdo com interações por deep learning. **RETEC-Revista de Tecnologias**, v. 16, n. 2, p. 31–46, 2023.

PAIVA, P. D.; SILVA, V. M. da; MOURA, R. S. Detecção automática de discurso de ódio em comentários online. In: SBC. **Escola Regional de Computação Aplicada à Saúde (ERCAS)**. [S.l.], 2019. p. 157–162.

PALMEIRA, W. W. d. A. **Um estudo comparativo de algoritmos de aprendizado de máquina na detecção de discurso de ódio na rede social Twitter**. Dissertação (B.S. thesis), 2023.

PELLE, R. P. D.; MOREIRA, V. P. Offensive comments in the brazilian web: a dataset and baseline results. In: SBC. **Brazilian Workshop on Social Network Analysis and Mining (BRASNAM)**. [S.l.], 2017. p. 510–519.

PELLE, R. P. de; MOREIRA, V. P. Offensive comments in the brazilian web: a dataset and baseline results. 2017.

PLATH, H. O.; PAIVA, M. E. O.; PINTO, D. L.; COSTA, P. D. P. Detecção de discurso de Ódio contra mulheres em textos em português brasileiro: Construção da base mina-br e modelo de classificação. **Revista Eletrônica de Iniciação Científica em Computação**, v. 20, n. 3, jul. 2022. Disponível em: <https://journals-sol.sbc.org.br/index.php/reic/article/view/2696>.

SILVA, B. P.; SANTOS, F. Identificação de discurso de ódio nas eleições municipais de 2020. **Boletim de Conjuntura (BOCA)**, v. 19, n. 55, p. 227–247, 2024.

SILVA, S.; SERAPIãO, A. Detecção de discurso de ódio em português usando cnn combinada a vetores de palavras. In: **Anais do VI Symposium on Knowledge Discovery, Mining and Learning**. Porto Alegre, RS, Brasil: SBC, 2018. p. 1–8. ISSN 2763-8944. Disponível em: <https://sol.sbc.org.br/index.php/kdmile/article/view/27378>.

SILVA, S. C.; SERAPIÃO, A. B. Detecção de discurso de ódio em português usando cnn combinada a vetores de palavras. In: SBC. **Symposium on Knowledge Discovery, Mining and Learning (KDMiLe)**. [S.l.], 2018. p. 1–8.

VARGAS, F.; CARVALHO, I.; GÓES, F. Rodrigues de; PARDO, T.; BENEVENUTO, F. HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In: **Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)**. Marseille, France: European Language Resources Association, 2022. p. 7174–7183. Disponível em: <https://aclanthology.org/2022.lrec-1.777>.

WASEEM, Z.; HOVY, D. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In: ANDREAS, J.; CHOI, E.; LAZARIDOU, A. (Ed.). **Proceedings of the NAACL Student Research Workshop**. San Diego, California: Association for Computational Linguistics, 2016. p. 88–93. Disponível em: <https://aclanthology.org/N16-2013/>.

WEITZEL, L.; DAROZ, T. H.; CUNHA, L. P.; HELDE, R. V.; MORAIS, L. M. de. Investigating deep learning approaches for hate speech detection in social media: Portuguese-br tweets. In: IEEE. **2023 18th Iberian Conference on Information Systems and Technologies (CISTI)**. [S.l.], 2023. p. 1–5.