Revista ARACÊ
ISSN: 2358-2472

DETECÇÃO AUTOMÁTICA DE DISCURSO HOMOFÓBICO UTILIZANDO APRENDIZADO DE MÁQUINA

https://doi.org/10.56238/arev7n5-029

Data de submissão: 02/04/2025 **Data de publicação:** 02/05/2025

Samuel Henrique Santos Silva
Bacharel em Sistemas de Informação
Universidade de Pernambuco
E-mail: samhenriquess@gmail.com

Erika Carlos Medeiros
Doutora em Ciência da Computação
Universidade de Pernambuco
E-mail: erika.medeiros@upe.br
ORCID: 0000-0003-2506-7116
LATTES: 6574506939749437

Patrícia Cristina Moser
Doutora em Ciência da Computação
Universidade de Pernambuco
E-mail: patricia.moser@upe.br
LATTES: 2977508109885476

Bianca Gabriely Ferreira Silva
Doutora em Administração
Universidade de Pernambuco
Caruaru, Pernambuco, Brasil
E-mail: bianca.ferreirasilva@upe.br
ORCID: 0000-0002-7881-398X
LATTES: 0419687984635851

Fernando Ferreira de Carvalho
Doutor em Ciência da Computação
Universidade de Pernambuco
Caruaru, Pernambuco, Brasil
E-mail: fernando.carvalho@upe.br
LATTES: 8491797408318076

Jorge Cavalcanti Barbosa Fonsêca
Doutor em Ciência da Computação
Universidade de Pernambuco
Caruaru, Pernambuco, Brasil
E-mail: jorge.fonseca@upe.br
LATTES: 8075101995480409

Revista ARACÊ
ISSN: 2358-2472

Rômulo César Dias de Andrade
Doutor em Ciência da Computação
Universidade de Pernambuco
Caruaru, Pernambuco, Brasil
E-mail: romulo.andrade@upe.br
LATTES: 1559585906838684

Marco Antônio de Oliveira Domingues
Doutor em Ciência da Computação
Instituto Federal de Pernambuco
Recife, Pernambuco, Brasil
E-mail: marcodomingues@recife.ifpe.edu.br
ORCID: 0000-0002-7579-348
LATTES: 7139685024425123

RESUMO
Esta pesquisa explora modelos de aprendizado de máquina para detectar discursos de ódio com contexto homofóbicos em redes sociais, um problema relevante na era digital devido ao impacto negativo sobre a comunidade LGBTQIA+. O objetivo geral é treinar modelos preditivos capazes de identificar discursos homofóbicos de forma eficiente, contribuindo para o combate ao discurso de ódio e promovendo um ambiente virtual mais seguro. A metodologia CRISP-DM foi utilizada, aplicando cinco fases: entendimento do negócio, entendimento e preparação dos dados, modelagem e avaliação. Foram treinados seis modelos: Decision Tree, Random Forest, Extra Trees, Passive Aggressive, eXtreme Gradient Boosting e Support Vector Machine. A avaliação dos modelos utilizou métricas como acurácia, precisão, recall e F1-Score, além de análise da matriz de confusão e da curva Receiver Operating Characteristic para medir o desempenho de cada modelo. O modelo SVM obteve o melhor desempenho geral, com acurácia de 87,10%, precisão de 79,15% e área sob a curva de 0,9227, destacando sua eficácia em minimizar falsos positivos. Os resultados evidenciam o potencial dos modelos de aprendizado na identificação de discurso de ódio e contribuem para a construção de ambientes digitais mais seguros e inclusivos.

Palavras-chave: Aprendizado de Máquina. Discurso de Ódio. Redes Sociais. Mineração de dados. Processamento de Linguagem Natural.

Revista ARACÊ
ISSN: 2358-2472

1 INTRODUÇÃO

A identificação precoce de discursos de ódio com contexto homofóbicos nas redes sociais é um desafio crescente que tem implicações diretas para a segurança e o bem-estar de pessoas LGBTQIA+. Segundo o relatório do Center for Countering Digital Hate (2022), intitulado *Digital Hate - Social Media's Role in Amplifying Dangerous Lies About LGBTQ+ People*, as plataformas contribuem com propagação do discurso de ódio online, incluindo homofobia, o que transforma plataformas sociais em ambientes hostis para essa comunidade. O impacto do discurso homofóbico, que pode ser classificado como uma forma de discurso de ódio, transcende as fronteiras da liberdade de expressão e alimenta o preconceito e a violência contra pessoas LGBTQIA+ (CHAKRAVARTHI et al., 2021). Detectar e mitigar esse tipo de conteúdo de forma eficaz poderia, então, auxiliar na criação de espaços mais inclusivos e respeitosos para todos os usuários.

Dados fornecidos pelo *National Crime Victimization Survey* (NCVS) (BUREAU OF JUSTICE STATISTICS, 2024) e o pelo trabalho de Flores et al. (2023) revelam que o discurso de ódio homofóbico é mais prevalente em contextos onde a proteção legal para essas comunidades é limitada, especificamente, pessoas LGBTQ+ têm uma probabilidade nove vezes maior de serem vítimas de crimes de ódio violentos em relação a pessoas cisgênero e heterossexuais, e essa diferença é mais acentuada em grupos de baixa renda e áreas urbanas. Detectar esses discursos de forma proativa permitiria não só evitar a escalada de violência verbal, mas também prevenir consequências psicológicas adversas para as vítimas.

Como apontam os trabalhos de Yin e Zubiaga (2021) e Macavaney et al. (2019), a presença de nuances linguísticas e definições ambíguas de discurso de ódio dificultam a precisão dos modelos, uma vez que muitos algoritmos não captam as sutilezas de termos ofensivos em contextos específicos. Além disso, esses trabalhos destacam que, embora os modelos recentes possam alcançar alto desempenho, a falta de interpretabilidade nas decisões ainda é um grande obstáculo, especialmente em métodos baseados em redes neurais, que frequentemente carecem de clareza em relação ao processo de tomada de decisão.

Segundo o relatório *U.S. National Survey on the Mental Health of LGBTQ Young People* (THE TREVOR PROJECT, 2023), há impacto significativo do discurso homofóbico online na saúde mental de jovens LGBTQIA+. Ele revela que ambientes virtuais hostis aumentam sintomas de ansiedade, depressão e o risco de suicídio, especialmente em contextos com políticas limitadas de proteção para essa comunidade. Cerca de 73% dos jovens LGBTQIA+ relataram sintomas de ansiedade, e 58% descreveram sintomas de depressão, com taxas alarmantes de tentativa de suicídio entre aqueles expostos a vulnerabilidades virtuais. Esses dados evidenciam a urgência de mecanismos de detecção

Revista ARACÊ
ISSN: 2358-2472

de discurso de ódio que possam tornar os espaços digitais mais seguros e acolhedores para essa população vulnerável. Nesse contexto, Vianna, Pareto e Bianchi (2025) discutem como o ambiente das redes sociais tem sido manipulado por estratégias de guerra cognitiva, que utilizam tecnologias e narrativas para disseminar desinformação e discursos de ódio, ampliando os riscos à integridade de grupos vulneráveis. Martins e Rodrigues (2024) também destacam que os algoritmos das redes sociais podem reforçar câmaras de eco e bolhas informacionais, favorecendo a disseminação de conteúdos extremistas e discursos de ódio, o que reforça a urgência de mecanismos de detecção automatizados e éticos.

A utilização de técnicas de aprendizado de máquina para identificar precocemente discursos homofóbicos tem se mostrado promissora em pesquisas recentes, demonstrando o potencial da inteligência artificial para combater o discurso de ódio. O estudo de Mcgiff e Nikolov (2024) aponta que a variabilidade nas expressões de ódio, o contexto cultural específico de cada idioma e a dificuldade em detectar discursos sutis e contextuais são obstáculos que ainda necessitam de novas abordagens para uma cobertura mais eficaz. Essa lacuna abre espaço para outras soluções, como a proposta neste trabalho, que explorem novas arquiteturas de modelos e técnicas de mineração de dados adaptadas à língua inglesa, complementando a literatura existente e contribuindo para um sistema mais completo e eficaz na identificação do discurso homofóbico.

Diante desse contexto, o objetivo geral desta pesquisa é explorar modelos de aprendizado de máquina capazes de detectar a presença de discurso homofóbico com base em discursos extraídos de redes sociais. Para atingir o objetivo geral, foram delineados quatro objetivos específicos, listados a seguir:

- Realizar o pré-processamento dos dados coletados do Kaggle (KAGGLE, 2024) e Hugging Face (HUGGING FACE, 2024), empregando técnicas estatísticas e de mineração de dados;
- Treinar modelos de aprendizado de máquina utilizando dados de discursos de ódio identificados na etapa anterior;
- Identificar e comparar os modelos de aprendizado de máquina com base em métricas de desempenho, incluindo acurácia (MÔNICO et al., 2009), precisão (MARIANO, 2021), recall (TORGO; RIBEIRO, 2009), F1-Score (ZHANG; WANG; ZHAO, 2015), matriz de confusão (LIANG, 2022) e área sobre a *Receiver Operating Characteristic Curve* (Curva ROC) (NAKAS set al., 2023);

Revista ARACÊ
ISSN: 2358-2472

- Escolher o modelo que apresentar as métricas mais elevadas de desempenho como o modelo final para a detecção de discurso homofóbico com base em discursos extraídos de redes sociais.

Este trabalho está estruturado em cinco seções, começando pela presente seção introdutória, que apresenta o contexto, motivação e objetivos. A segunda seção abordará os trabalhos relacionados, apresentando pesquisas recentes na área de aprendizagem de máquina relacionadas a identificação de discurso de homofóbico. A terceira seção descreverá a metodologia adotada, detalhando as etapas do pré-processamento de dados, treinamento dos modelos, além da avaliação de desempenho. A quarta seção apresentará os resultados obtidos e a análise dos modelos, enquanto a quinta e última seção consolidará as conclusões do estudo, destacando contribuições, limitações e sugestões para pesquisas futuras.

2 TRABALHOS RELACIONADOS

Nesta seção, serão apresentadas pesquisas e estudos relevantes que fundamentam e contextualizam o presente trabalho, oferecendo uma visão abrangente do estado da arte na área de estudo. Serão explorados conceitos-chave, metodologias e resultados de pesquisas anteriores, que servem como base teórica e prática para o desenvolvimento deste projeto. Essa revisão de literatura visa identificar as principais contribuições, lacunas e desafios enfrentados por outros pesquisadores, destacando como os métodos e abordagens escolhidos se relacionam com o objetivo desta pesquisa.

No estudo de Shanmugavadivel et al. (2024), o foco foi a detecção de comentários homofóbicos e transfóbicos em plataformas de mídias sociais, especificamente em comentários do YouTube em língua inglesa. Os autores implementaram diversas técnicas de aprendizado de máquina, incluindo modelos como Random Forest, Decision Tree e Support Vector Machine (SVM), além de um modelo de aprendizado profundo baseado em redes neurais LSTM.

A principal contribuição deste trabalho foi demonstrar que os modelos de aprendizado de máquina, especialmente o Random Forest, superaram os modelos de aprendizado profundo na tarefa de identificação de discursos de ódio, alcançando uma macro F1-Score de 0,369 no conjunto de dados em inglês. No entanto, os autores reconheceram lacunas em sua abordagem, destacando que não utilizaram embeddings contextualizados, como BERT ou GPT, que poderiam aprimorar significativamente o desempenho dos modelos. Além disso, mencionam a ausência de técnicas de transferência de aprendizado em sua análise atual, sugerindo que futuras pesquisas devem explorar essas metodologias para melhorar a detecção de comentários homofóbicos e transfóbicos em plataformas online.

Revista ARACÊ
ISSN: 2358-2472

Em seguida, o estudo de Chakravarthi et al. (2024) aborda a detecção de homofobia e transfobia em comentários do YouTube, com foco em conteúdo direcionado à comunidade LGBTQ+ em inglês, tâmil e em código misto tâmil-inglês. Os autores desenvolveram um novo conjunto de dados anotados por especialistas, contendo 15.141 comentários para treinar modelos de aprendizado de máquina na identificação automática de discurso homofóbico e transfóbico.

O processo envolveu a criação de um corpus multilíngue e a elaboração de uma taxonomia para classificar o conteúdo em três categorias: homofóbico, transfóbico ou não anti-LGBTQ+. Experimentos foram realizados com diversos modelos de aprendizado de máquina e aprendizado profundo, incluindo algoritmos tradicionais, como Regressão Logística, Naive Bayes e SVM, além de modelos baseados em *Deep Neural Networks* (DNN), como BERT e BiLSTM.

Os resultados indicaram que os modelos de aprendizado profundo superaram significativamente os modelos tradicionais, especialmente ao utilizar *embeddings* como BERT. Especificamente, o modelo baseado em BERT alcançou uma F1-Score macro de 0,570 para o tâmil, 0,870 para o inglês e 0,610 para o código misto tâmil-inglês. A principal contribuição deste trabalho é a disponibilização de um novo conjunto de dados anotados para a detecção de homofobia e transfobia em comentários online, bem como a demonstração da eficácia dos modelos de aprendizado profundo nessa tarefa.

No entanto, a detecção de discursos homofóbicos e transfóbicos em contextos multilíngues e multiculturais mostrou-se desafiadora, devido à complexidade linguística e à escassez de dados anotados em línguas com poucos recursos, como o tâmil. Os autores destacam a necessidade de pesquisas futuras para aprimorar os modelos e expandir os recursos linguísticos em línguas sub-representadas, visando melhorar a identificação automática desse tipo de conteúdo prejudicial nas redes sociais (CHAKRAVARTHI et al., 2024)

Ashraf et al. (2022) desenvolveram um modelo baseado em aprendizado de máquina para a detecção automática de homofobia e transfobia em comentários de mídias sociais. O estudo utilizou a técnica *Term Frequency-Inverse Document Frequency Vectorizer* (TF-IDF) (JOACHIMS, 1997) para a vetorização dos textos e implementou o algoritmo SVM como principal classificador. Os autores avaliaram o modelo em conjuntos de dados em inglês, tâmil e uma combinação de tâmil-inglês, alcançando F1-Scores ponderados de 0,91, 0,92 e 0,88, respectivamente.

A principal contribuição deste trabalho é a aplicação eficaz do SVM combinado com TF-IDF na detecção de discursos homofóbicos e transfóbicos em múltiplos idiomas, incluindo línguas de baixo recurso e textos em código misto. No entanto, os autores reconheceram lacunas, como a possibilidade de melhorar o desempenho por meio de sistemas mais complexos, como modelos baseados em

Revista ARACÊ
ISSN: 2358-2472

aprendizado profundo. Além disso, sugerem que o uso de representações de palavras mais avançadas, como *embeddings*, poderia aprimorar a representação textual e, consequentemente, a performance do classificador (Ashraf et al., 2022)

Arcila-Calderón et al. (2021) exploraram a detecção automática de discursos de ódio motivados por gênero e orientação sexual no Twitter (X, 2024) em espanhol. Os autores desenvolveram um corpus de treinamento específico, anotado manualmente, para treinar modelos de aprendizado de máquina supervisionados. Utilizaram tanto algoritmos de aprendizado raso, como Naive Bayes, Regressão Logística e SVM, quanto de aprendizado profundo, especificamente Recurrent Neural Network (RNN). Os resultados mostraram que os modelos de aprendizado profundo superaram significativamente os modelos de aprendizado raso. A Regressão Logística apresentou o melhor desempenho entre os modelos rasos, mas o modelo baseado em RNN obteve métricas superiores, com acurácia e Área Sob a Curva (AUC) significativamente melhores. Contudo, o F1-Score diminuiu consideravelmente no modelo de aprendizado profundo, indicando a necessidade de aprimoramento. A principal contribuição deste trabalho é a criação do primeiro protótipo para detecção automática de discurso de ódio motivado por gênero e orientação sexual em língua espanhola, além de demonstrar a vantagem do aprendizado profundo nessa tarefa.

As lacunas identificadas pelos autores incluem a necessidade de aprimorar o corpus de treinamento para melhorar a precisão dos modelos, especialmente o F1-Score. Além disso, sugerem que estudos futuros devem realizar validação externa para confirmar o desempenho do modelo e expandir o corpus de treinamento para superar limitações, como o tamanho limitado da amostra e a coleta de dados em um único momento no tempo (Arcila-Calderón et al., 2021)

Por fim, Nirmal et al. (2020) concentraram-se na detecção automatizada de *cyberbullying* em mídias sociais, especificamente no Twitter (X, 2024), utilizando técnicas de Processamento de Linguagem Natural (PLN) e algoritmos de aprendizado de máquina, como Naive Bayes, SVM e DNN. O estudo destacou a crescente preocupação com o *cyberbullying* e propôs uma metodologia que envolve a coleta e pré-processamento de dados, vetorização TF-IDF e aplicação de modelos de classificação para identificar automaticamente instâncias de cyberbullying.

A principal contribuição deste trabalho é a implementação de um sistema que combina abordagens textuais e não textuais para a detecção de *cyberbullying*, reconhecendo a complexidade da tarefa devido à necessidade de contexto para uma classificação precisa. No entanto, os autores reconheceram lacunas em sua pesquisa, como a necessidade de aprimorar a precisão dos modelos por meio do uso de conjuntos de dados mais robustos e diversificados. Além disso, sugerem futuras extensões, como a expansão da detecção para além de vítimas e agressores, a determinação do estado

Revista ARACÊ
ISSN: 2358-2472

emocional da vítima após um incidente e a detecção em tempo real de dados em fluxo. A ausência de técnicas mais avançadas, como *embeddings* contextuais e modelos de aprendizado profundo mais sofisticados, indica espaço para melhorias adicionais na abordagem proposta (Nirmal et al., 2020)

Em síntese, os estudos apresentados evidenciam avanços significativos na detecção automática de discursos de ódio direcionados à comunidade LGBTQ+ em diferentes idiomas e contextos culturais. A utilização de modelos de aprendizado profundo, especialmente aqueles baseados em *embeddings* contextuais como BERT, mostrou-se promissora na melhoria do desempenho dos classificadores. Contudo, desafios persistem, especialmente relacionados à escassez de dados anotados em línguas com poucos recursos, à necessidade de aprimorar a precisão dos modelos e à adaptação a contextos multilíngues e multiculturais.

A próxima seção descreve a metodologia usada neste trabalho para treinamento do modelos de aprendizado de máquina.

3 METODOLOGIA

A metodologia utilizada para o desenvolvimento deste trabalho é baseada no modelo *Cross Industry Standard Process for Data Mining* (CRISP-DM) (SCHRÖER; KRUSE; GÓMEZ, 2021), que consiste em uma abordagem consolidada para projetos de mineração de dados. Este processo é dividido em seis fases: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e implantação. No contexto deste trabalho, serão executadas as cinco primeiras fases. A fase de implantação, que envolveria a implementação do modelo em um ambiente de produção, estará fora do escopo deste trabalho. A seguir, são detalhadas as atividades desenvolvidas em cada fase:

a) Entendimento do negócio: No entendimento do negócio, a detecção automática de discurso homofóbico utilizando aprendizado de máquina foi analisada a partir de uma revisão da literatura, conforme apresentado na seção 2. Este levantamento buscou identificar o impacto da inteligência artificial no combate ao discurso de ódio, destacando seu papel na automação da análise de grandes volumes de dados textuais. Foram observados através dos estudos apresentados as lacunas existentes, bem como a necessidade de aprimorar a precisão dos modelos em identificar nuances culturais e linguísticas específicas ao discurso homofóbico.

b) Entendimento dos dados: Na fase de entendimento dos dados, é essencial compreender a composição e estrutura do conjunto de dados que será utilizado para a detecção automática de discurso homofóbico. O conjunto de dados utilizado neste trabalho é composto por um total de 5.576 registros de discursos, dos quais 3.699 são classificados como discursos homofóbicos, enquanto 1.877 são

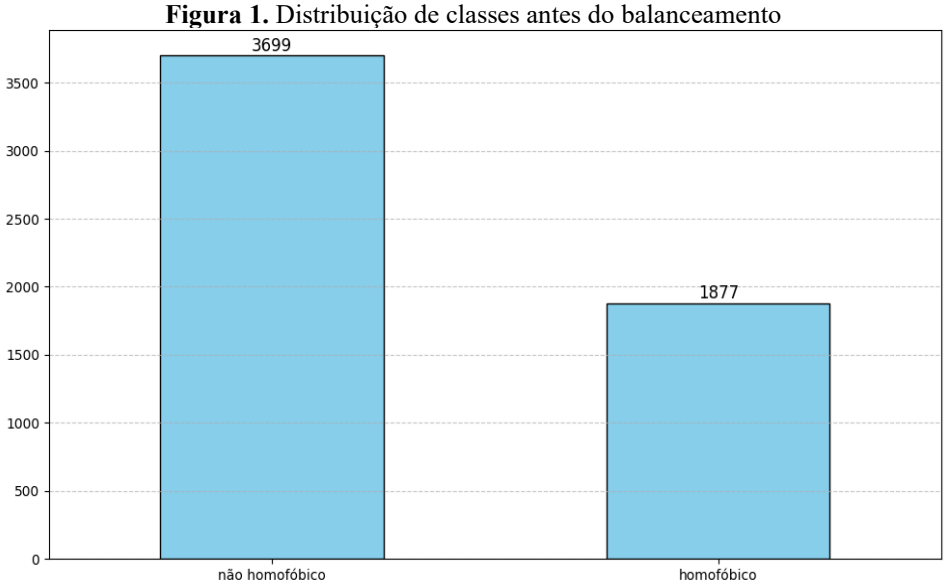
Revista ARACÊ
ISSN: 2358-2472

discursos não homofóbicos, como ilustrado na Figura 1. Este conjunto de dados será a base para o treinamento e avaliação dos modelos de aprendizado de máquina, permitindo a identificação de padrões linguísticos específicos associados ao discurso homofóbico.

Os dados utilizados neste trabalho foram coletados de dois conjuntos de dados distintos e estão escritos na língua inglesa. O primeiro conjunto foi refeito por Wood (2023), com base no conjunto de dados originalmente coletado por Sachdeva et al. (2022) a partir das plataformas *X* (anteriormente conhecido como Twitter) (X, 2024), Reddit (Reddit, 2024) e YouTube (Youtube, 2024).

O segundo conjunto de dados foi coletado por Mcgiff e Nikolov (2024), também a partir de discursos extraídos da plataforma *X*. Esses dois conjuntos de dados foram combinados para compor o corpus final utilizado neste trabalho, permitindo uma análise abrangente de discursos homofóbicos em diferentes contextos e ambientes digitais.

Figura 1. Distribuição de classes antes do balanceamento



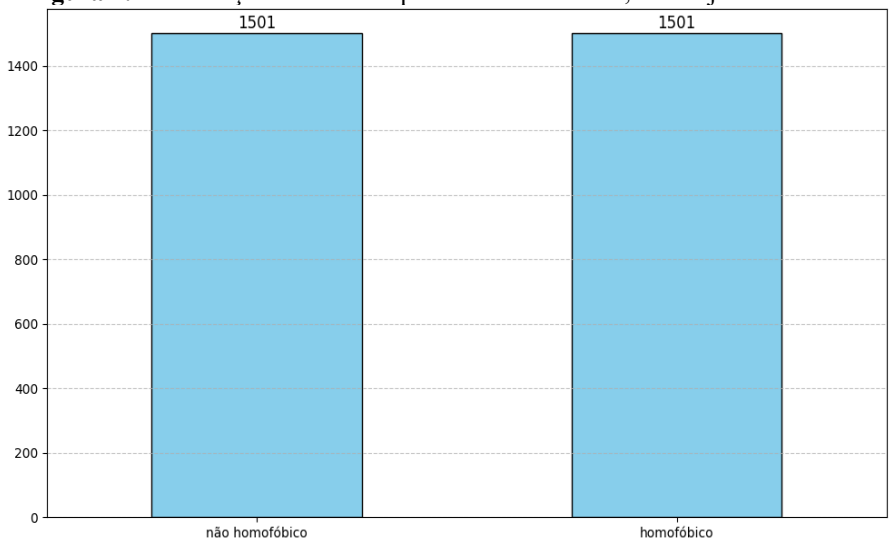
Fonte: Os Autores

c) Preparação dos dados: Na fase de preparação dos dados, foram realizadas diversas etapas para garantir a qualidade e a consistência do conjunto de dados utilizado para a detecção de discurso homofóbico. Inicialmente, foi feita a identificação de duplicatas, não havendo duplicatas no conjunto de dados. Em seguida, os textos foram convertidos para letras minúsculas para padronizar o formato e facilitar a análise. Emojis foram substituídos por seu texto correspondente para que suas nuances fossem corretamente interpretadas pelos modelos de aprendizado de máquina. Também foram removidas *stop words*, que são palavras comuns e pouco informativas, e marcações de usuários, *Uniform Resource Locators* (URLs) e caracteres não textuais foram eliminados para limpar o texto e reduzir o ruído. Essas etapas de pré-processamento são essenciais para melhorar a qualidade dos dados

Revista ARACÊ
ISSN: 2358-2472

e assegurar que o modelo possa focar nos aspectos relevantes do discurso durante o treinamento e a avaliação. Foi utilizado o TF-IDF para a tokenização dos dados. Por fim, foi aplicada a técnica de *undersampling* (MOHAMMED et al., 2020) para balanceamento das classes no conjunto de treino, reduzindo a classe majoritária para o mesmo valor da classe minoritária, como ilustrado na Figura 2.

Figura 2. Distribuição de classes após o balanceamento, no conjunto de treino



Fonte: Os Autores

d) Modelagem: Na fase de modelagem, o conjunto de dados foi dividido em conjuntos de treino e teste na proporção de 80% para treinamento e 20% para teste. Foram utilizados os modelos: Random Forest, Decision Tree, SVM, Passive Aggressive, Extra Trees e eXtreme Gradient Boosting (XGBoost). Para otimizar o desempenho dos modelos, foi realizado uma busca em grade (KIM, 1997) para ajuste fino dos hiperparâmetros da biblioteca scikit-learn (scikit-learn, 2024) dos modelos. A busca em grade permitiu testar uma gama de valores para os hiperparâmetros de cada modelo, buscando as melhores combinações para melhorar a precisão e a eficácia dos modelos.

Tabela 1. Hiperparâmetros da biblioteca scikit-learn testados na busca em grade

Modelo	Hiper Parâmetro Ajustado	Valores Testados	Valor Escolhido
Decision Tree	max_depth	3, 5, 10, None	None
	min_samples_split	2, 10, 20	10
Random Forest	n_estimators	50, 100, 200	200
	max_depth	3, 5, 10, None	None
Extra Trees	n_estimators	50, 100, 200	200
	max_depth	3, 5, 10, None	None
Passive Agressive	C	0.001, 0.01, 0.1, 1, 10	0.01
	max_iter	1000, 2000	1000
SVM	C	0.1, 1, 10	1
	Kernel	linear, rbf	linear

Revista ARACÊ
ISSN: 2358-2472

XGBoost	n_estimators	50, 100, 200	100
	learning_rate	0.01, 0.1, 0.2	0.2
	max_depth	3, 5, 10	10

Fonte: Os Autores

Os valores dos hiperparâmetros testados e as respectivas combinações são detalhados na Tabela 1, oferecendo uma visão completa das configurações exploradas para alcançar o melhor desempenho possível na tarefa de classificação.

e) Avaliação: Na fase de avaliação, os modelos foram testados e comparados com base seguintes métricas de desempenho: acurácia, precisão, recall, F1-Score, além da análise da matriz de confusão e da curva ROC.

Na próxima seção, são discutidos os resultados obtidos a partir da avaliação dos modelos de aprendizado de máquina. Será realizada uma análise do desempenho de cada modelo, considerando como eles se comportaram na tarefa de detecção de discurso homofóbico. Essa análise fornecerá uma visão sobre qual modelo oferece os melhores resultados.

4 RESULTADOS

Nesta seção, serão apresentados os resultados da avaliação de desempenho dos modelos de aprendizado de máquina aplicados à tarefa de avaliação preditiva de identificação de discurso homofóbico. Os modelos avaliados incluem Random Forest, Decision Tree, SVM, Passive Aggressive, Extra Trees e XGBoost. As métricas utilizadas para avaliar o desempenho dos modelos são acurácia, precisão, recall, F1-Score, matriz de confusão e curva ROC que medem a proporção de previsões corretas em relação ao total de previsões.

Tabela 2. Hiper parâmetros da biblioteca scikit-learn testados na busca em grade

Modelos	Acurácia	Precisão	Recall	F1-Score
Decision Tree	80,82%	68,16%	80,85%	73,97%
Extra Trees	83,60%	73,14%	81,12%	76,92%
Random Forest	84,14%	72,56%	85,11%	78,34%
Passive Aggressive	86,20%	77,07%	84,04%	80,41%
XGBoost	86,20%	78,32%	81,65%	79,95%
SVM	87,10%	79,15%	83,78%	81,40%

Fonte: Os Autores

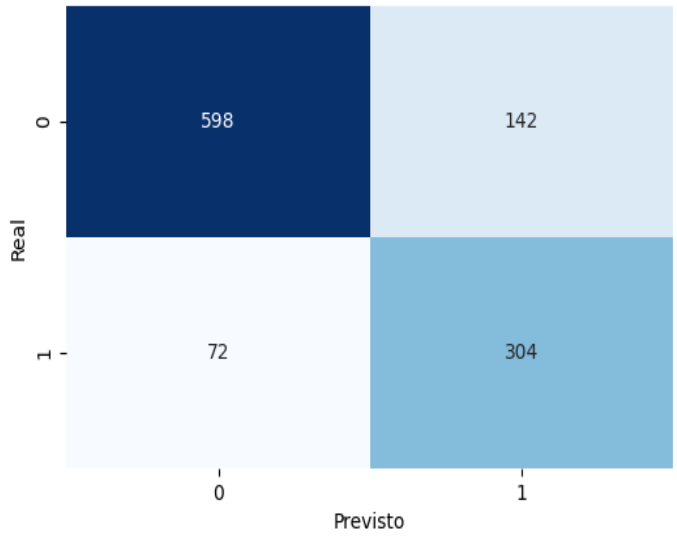
A Tabela 2 apresenta a acurácia de teste, precisão, recall e F1-Score dos diferentes modelos de aprendizado de máquina utilizados. A seguir segue a análise de como cada modelo se comportou na tarefa de detecção automática de discurso homofóbico.

O modelo Decision Tree alcançou uma acurácia de 80,82%, sendo o modelo de menor desempenho em termos de acurácia entre todos os testados. A precisão foi de 68,16%, o que significa

Revista ARACÊ
ISSN: 2358-2472

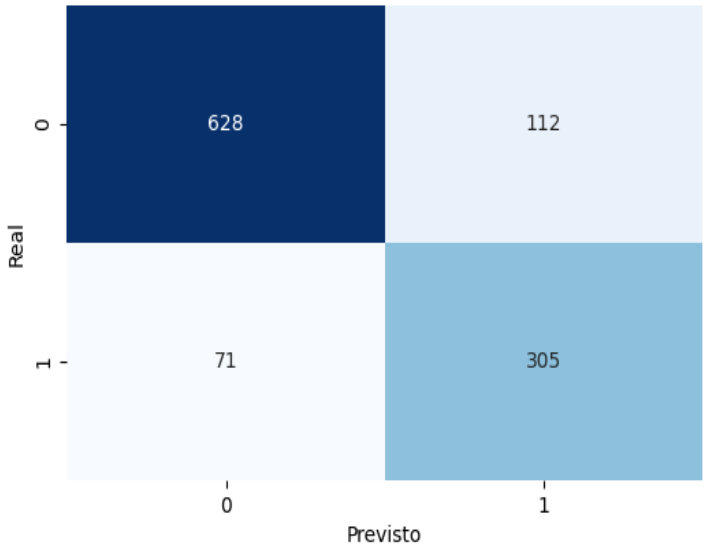
que o modelo teve dificuldade em identificar corretamente discursos homofóbicos, resultando em uma maior quantidade de falsos positivos. O recall de 80,85% indica que o Decision Tree conseguiu identificar uma proporção considerável de discursos homofóbicos, mas ainda deixou alguns casos de lado, possivelmente devido a uma limitação na generalização do modelo. O F1-Score de 73,97% reflete o equilíbrio entre a precisão e o recall, mas ainda mostra limitações na generalização do modelo. A matriz de confusão (Figura 3) mostra que o modelo confundiu 142 discursos não homofóbicos como homofóbicos e classificou incorretamente 72 discursos homofóbicos como não homofóbicos, o que corresponde a taxa de 19,18% de erros.

Figura 3. Matriz de confusão do modelo Decision Tree



Fonte: Os Autores

Figura 4. Matriz de confusão do modelo Extra Trees



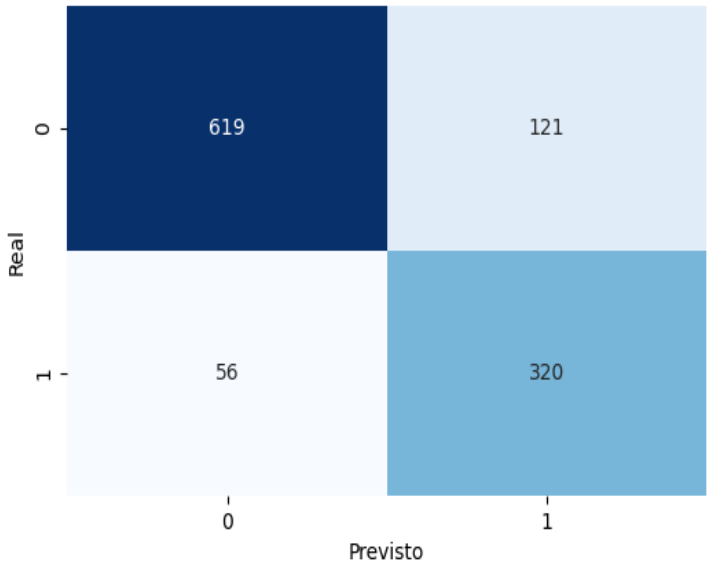
Fonte: Os Autores

Revista ARACÊ
ISSN: 2358-2472

O modelo Extra Trees, com uma acurácia de 83,60%, apresentou um desempenho semelhante ao Random Forest, mas com um ligeiro ganho de desempenho. A precisão de 73,14% e o recall de 81,12% refletem uma melhoria marginal na capacidade de identificar discursos homofóbicos corretamente. O F1-Score de 76,92% indica que o modelo conseguiu manter um bom equilíbrio entre as duas métricas, o que o torna um concorrente direto do Random Forest. A matriz de confusão (Figura 4) é bastante semelhante à do Random Forest, com 112 discursos não homofóbicos e 71 discursos homofóbicos classificados incorretamente, o que corresponde a taxa de 16,40% de erros. Esse resultado sugere que a aleatoriedade adicional do Extra Trees pode ter ajudado o modelo a generalizar ligeiramente melhor que o Random Forest, embora o ganho não tenha sido expressivo.

O Random Forest apresentou uma melhora em relação ao Decision Tree, com uma acurácia de 84,14%. A precisão de 72,56% e o recall de 85,11% mostram que o modelo foi mais eficiente em detectar discursos homofóbicos, reduzindo a quantidade de falsos positivos em comparação com o Decision Tree.

Figura 5. Matriz de confusão do modelo Random Forest



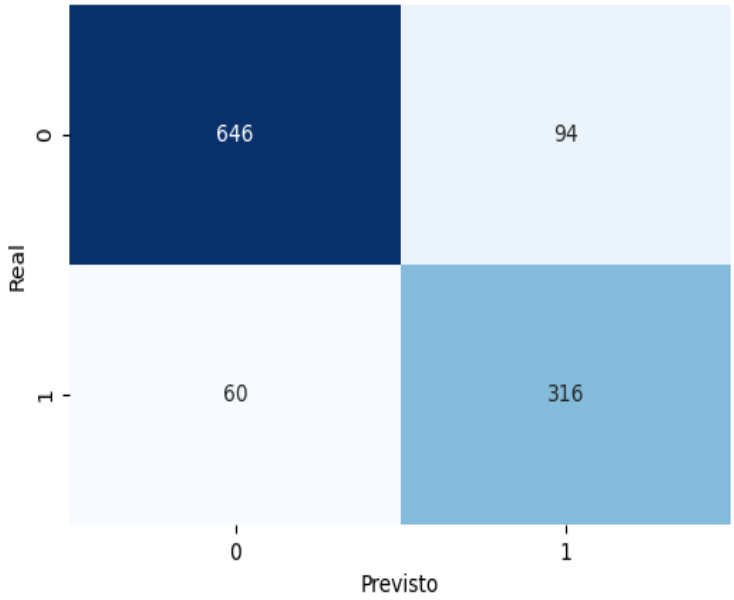
Fonte: Os Autores

O F1-Score de 78,34% sugere um equilíbrio mais adequado entre precisão e recall, indicando que o Random Forest conseguiu capturar mais nuances dos dados. A matriz de confusão (Figura 5) revela uma diminuição significativa de falsos positivos e falsos negativos, com 121 discursos não homofóbicos classificados incorretamente e 56 discursos homofóbicos mal classificados, o que corresponde a taxa de 15,86% de erros. Isso destaca a eficácia do Random Forest na tarefa de classificação, beneficiando-se do ensemble de múltiplas árvores de decisão.

Revista ARACÊ
ISSN: 2358-2472

O Passive Aggressive foi um dos modelos de maior desempenho, com uma acurácia de 86,20%. Sua precisão de 77,07% e recall de 84,04% indicam que o modelo foi eficiente na identificação de discursos homofóbicos, com menos falsos positivos em comparação aos modelos anteriores. O F1-Score de 80,41% sugere que o Passive Aggressive conseguiu um equilíbrio significativo entre precisão e recall, particularmente eficaz em conjuntos de dados desbalanceados. Na matriz de confusão (Figura 6), o modelo classificou incorretamente 94 discursos não homofóbicos e 60 discursos homofóbicos, o que corresponde a taxa de 13,79% de erros, mostrando uma redução considerável de falsos positivos e falsos negativos em comparação com os modelos baseados em árvores. Esse modelo demonstrou ser robusto para dados textuais, especialmente em classificações com grandes volumes de dados.

Figura 6. Matriz de confusão do modelo Passive Agressive



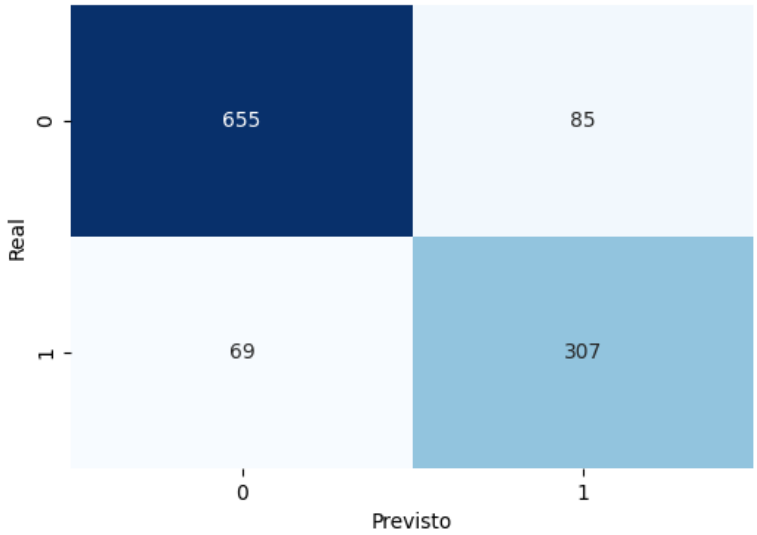
Fonte: Os Autores

O XGBoost apresentou uma acurácia de 86,20%, igualando o desempenho do Passive Aggressive. No entanto, sua precisão foi de 78,32%, levemente inferior à do SVM, mas ainda assim muito competitiva. O recall de 81,65% sugere que o XGBoost teve uma leve dificuldade em detectar alguns discursos homofóbicos em comparação com os outros modelos de maior precisão. A F1-Score de 79,95% reflete o equilíbrio entre suas métricas. A matriz de confusão (Figura 7) mostra que o XGBoost cometeu 85 erros ao classificar discursos não homofóbicos e 69 ao classificar discursos homofóbicos, o que corresponde a taxa de 13,79% de erros. Embora tenha um desempenho robusto, ele ficou levemente atrás do SVM, discutido a seguir, em termos de precisão. No entanto, o XGBoost

Revista ARACÊ
ISSN: 2358-2472

continua a ser uma excelente escolha para problemas com dados textuais e desbalanceados, especialmente devido à sua flexibilidade e capacidade de ajuste fino.

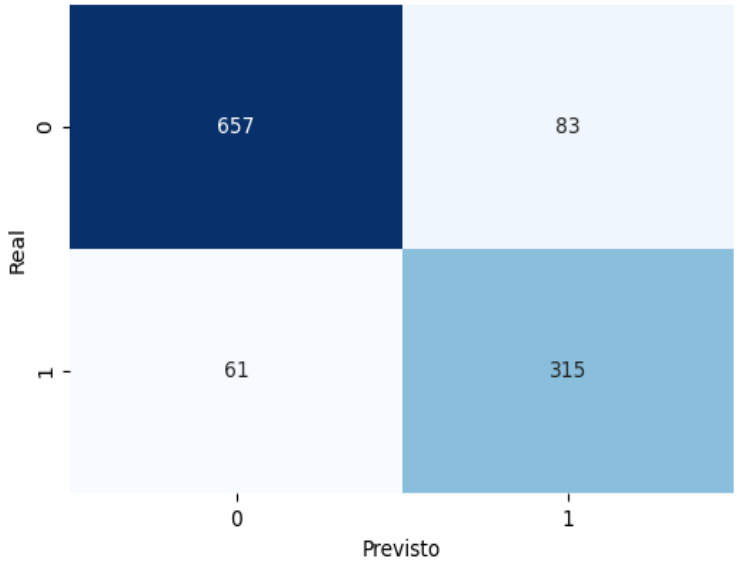
Figura 7. Matriz de confusão do modelo XGBoost



Fonte: Os Autores

O SVM foi o modelo de melhor desempenho geral, com uma acurácia de 87,10%. A precisão de 79,15% foi a maior entre todos os modelos testados, indicando que o SVM foi o mais eficiente em minimizar falsos positivos. O recall de 83,78% mostra que ele também conseguiu identificar a maioria dos discursos homofóbicos corretamente. O F1-Score de 81,40% confirma o equilíbrio robusto entre as métricas, tornando o SVM uma escolha forte para essa tarefa.

Figura 8. Matriz de confusão do modelo SVM



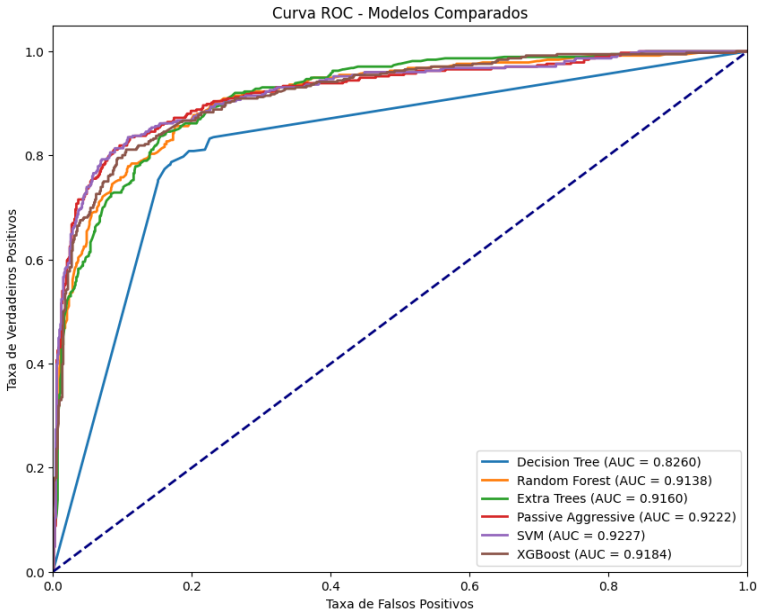
Fonte: Os Autores

Revista ARACÊ
ISSN: 2358-2472

A matriz de confusão (Figura 8) revela que o modelo cometeu 83 erros ao classificar discursos não homofóbicos e 61 ao classificar discursos homofóbicos, o que corresponde a taxa de 12,90% de erros, apresentando a menor quantidade de falsos positivos entre todos os modelos. Esse resultado evidencia a eficiência do SVM na detecção de fronteiras claras entre as classes.

Por ter obtido todas as métricas avaliadas superiores aos demais modelos, o SVM é o modelo escolhido para previsão automática de discurso de ódio homofóbico.

Figura 9. Curva ROC dos modelos



Fonte: Os Autores

A curva ROC (Figura 9) apresentada compara o desempenho de seis modelos de classificação utilizados na detecção de discurso de ódio, avaliados pelo valor da AUC. O modelo SVM se destaca, com a maior AUC de 0,9227, seguido de perto pelos modelos XGBoost e Passive Aggressive, que também obtiveram bons desempenhos, com AUCs de 0,9184 e 0,9222, respectivamente. Os modelos de Random Forest e Extra Trees tiveram desempenhos similares, com AUCs de 0,9138 e 0,9160, enquanto o modelo Decision Tree apresentou o pior desempenho relativo, com uma AUC de 0,8260. Esses resultados indicam que, embora todos os modelos tenham apresentado uma boa capacidade de discriminar entre classes, os modelos de SVM, XGBoost e Passive Aggressive são os mais eficazes para esta tarefa, mostrando-se superiores na taxa de verdadeiros positivos em relação à taxa de falsos positivos. O modelo de Decision Tree, por outro lado, tem um desempenho mais modesto, o que pode ser atribuído à sua maior simplicidade comparado aos demais algoritmos.

Revista ARACÊ
ISSN: 2358-2472

Tabela 3. Tempo de execução dos modelos

Modelo	Tempo de Execução
Decision Tree	28s
Random Forest	21.1s
Extra Trees	32.4s
Passive Aggressive	0.2s
SVM	16.6s
XGBoost	1m18.1s

Fonte: O autor

Como visto na Tabela 3, os tempos de execução dos modelos variam consideravelmente, com o XGBoost sendo o mais demorado, levando 1m18.1s, enquanto o Passive Aggressive é o mais rápido, com apenas 0.2s. Apesar dessas diferenças, todos os modelos têm tempos de execução suficientemente pequenos para que o tempo não seja um fator determinante na escolha do melhor algoritmo para esta tarefa específica de detecção de discurso de ódio. Isso porque, em um contexto de produção, a execução de qualquer um dos modelos dentro de poucos segundos a poucos minutos é aceitável, especialmente considerando que as diferenças de desempenho em termos de precisão, recall e AUC são mais significativas do que o tempo de execução. Portanto, a decisão de qual modelo, no contexto deste trabalho, utilizar deve priorizar a acurácia e a robustez do modelo em detrimento do tempo de execução.

5 CONSIDERAÇÕES GERAIS

A identificação de discursos homofóbicos nas redes sociais é um desafio crescente que possui implicações diretas para a segurança e o bem-estar de pessoas LGBTQIA+. O relatório do Center For Countering Digital Hate (2022) evidencia o papel das plataformas na disseminação de discurso de ódio, incluindo a homofobia, transformando-as em ambientes hostis para essa comunidade. Esse tipo de discurso transcende o direito à liberdade de expressão, reforçando o preconceito e promovendo a violência contra pessoas LGBTQIA+ (CHAKRAVARTHI, 2024). O orgão NCVS nos Estados Unidos revela que pessoas LGBTQ+ têm nove vezes mais chances de serem vítimas de crimes de ódio violentos em comparação com pessoas heterossexuais e cisgênero, com uma prevalência ainda maior em grupos de baixa renda e em áreas urbanas.

Para alcançar o objetivo principal de detectar precocemente discursos homofóbicos em redes sociais utilizando aprendizado de máquina, foram definidos quatro objetivos específicos. O primeiro objetivo específico foi o pré-processamento dos dados coletados do Kaggle e do Hugging Face, aplicando técnicas estatísticas e de mineração de dados. Nesse processo, realizou-se a limpeza e normalização dos textos para garantir a integridade e relevância dos dados, eliminando ruídos e

Revista ARACÊ
ISSN: 2358-2472

preparando as informações para a análise posterior. Também foi aplicada a técnica de *undersampling* na classe majoritária do conjunto de treino.

O segundo objetivo consistiu no treinamento de modelos de aprendizado de máquina utilizando os dados preparados na etapa anterior, a fim de identificar discursos homofóbicos em postagens de redes sociais. Durante essa fase, aplicou-se diferentes algoritmos para explorar suas capacidades de identificar padrões de discriminação homofóbica em textos, ajustando parâmetros e selecionando o modelo mais adequado para a tarefa.

O terceiro objetivo foi a análise e comparação dos modelos de aprendizado de máquina com base em métricas de desempenho. Avaliou-se acurácia, precisão, recall, F1-Score, matriz de confusão e AUC de cada modelo, usando essas métricas para determinar a eficácia de cada modelo em detectar discursos homofóbicos. Essa análise permitiu identificar os modelos mais robustos e adequados para a tarefa, destacando aqueles que apresentaram melhor desempenho nas métricas avaliadas.

O quarto e último objetivo específico foi a escolha do modelo com as métricas mais elevadas de desempenho para a detecção de discursos homofóbicos. Ao final dessa fase, selecinou-se o modelo que melhor equilibrava acurácia e precisão, oferecendo uma solução otimizada para a detecção proativa de discurso homofóbico nas redes sociais.

Com a conclusão e resolução de cada um dos objetivos específicos, foi possível alcançar o objetivo geral deste trabalho: implementar um modelo eficaz de aprendizado de máquina capaz de identificar discursos homofóbicos em redes sociais.

Para a tarefa de detecção de discurso homofóbico, foram explorados seis modelos de aprendizado de máquina, cada um com resultados variados, refletidos nas métricas de desempenho e tempo de execução.

O Decision Tree obteve a menor acurácia (80,82%) e precisão (68,16%) entre os modelos testados, indicando uma limitação na capacidade de generalização e uma maior quantidade de falsos positivos e negativos. O Extra Trees apresentou um aumento na acurácia (83,60%) e manteve um bom equilíbrio entre precisão e recall, sugerindo uma leve melhoria em relação ao Decision Tree, mas com um ganho marginal sobre o Random Forest.

O Random Forest, com acurácia de 84,14%, mostrou-se mais eficiente em reduzir falsos positivos e falsos negativos em comparação com o Decision Tree, refletindo um aumento na capacidade de capturar nuances nos dados com uma abordagem de ensemble. O Passive Aggressive Classifier alcançou uma acurácia de 86,20%, revelando um desempenho considerável na detecção de discursos homofóbicos e um bom equilíbrio entre precisão e recall, destacando-se na análise de dados desbalanceados.

Revista ARACÊ
ISSN: 2358-2472

O XGBoost teve desempenho competitivo com o Passive Aggressive, com uma acurácia de 86,20% e uma leve superioridade em precisão, sendo vantajoso pela flexibilidade e ajuste fino em dados textuais complexos. O SVM obteve a maior acurácia (87,10%) e melhor precisão (79,15%), com a menor taxa de falsos positivos, destacando-se como a opção mais robusta para esta tarefa.

Considerando a precisão e as demais métricas, o SVM foi selecionado como solução final, proporcionando um equilíbrio ideal entre desempenho e aplicabilidade na tarefa de identificação automática de discurso homofóbico.

O presente estudo propôs uma abordagem que explora a aplicação de modelos como Random Forest, Decision Tree, SVM, Passive Aggressive, Extra Trees e XGBoost para a identificação de discurso homofóbico, obtendo resultados comparáveis ou superiores em termos de acurácia e F1-Score aos trabalhos relacionados. Diferente de alguns trabalhos relacionados, que utilizaram conjunto de dados amplamente conhecidos, este trabalho desenvolveu usou dois conjuntos de dados tornando o corpus mais adequado ao contexto do estudo, evitando uma possível limitação de abrangência linguística e cultural.

Apesar dos resultados promissores na identificação de discurso homofóbico, algumas limitações foram observadas no trabalho e merecem consideração. A abordagem textual, embora eficaz, deixa de capturar nuances de discursos multimodais, comuns nas redes sociais, onde texto e imagem frequentemente se combinam para transmitir mensagens odiosas. A adaptabilidade do modelo também representa uma limitação, uma vez que o discurso de ódio evolui rapidamente, exigindo atualizações periódicas para garantir a eficácia na detecção de novas expressões e gírias. Por fim, o foco em métricas tradicionais, como precisão e F1-Score, poderia ser complementado por métricas que considerem o desbalanceamento das classes e ofereçam uma avaliação mais detalhada do desempenho na identificação de discursos menos comuns. Esses pontos sugerem direções para aprimoramentos futuros na área.

Trabalhos futuros poderão expandir esta pesquisa por meio de várias direções promissoras. Primeiramente, incluir dados multimodais, como imagens e vídeos, poderá aprimorar a compreensão dos contextos nos quais o discurso homofóbico ocorre, oferecendo uma visão mais abrangente das expressões de ódio em plataformas visuais. Além disso, adaptar o modelo para incorporar técnicas de aprendizado contínuo e atualização automática seria essencial para acompanhar a rápida evolução das linguagens e gírias que caracterizam discursos odiosos. Outro aspecto relevante é a aplicação de métodos mais sofisticados de balanceamento de classes, como técnicas avançadas de *oversampling* ou algoritmos de aprendizado em desequilíbrio, para mitigar os impactos de classes minoritárias nos resultados finais. Finalmente, a realização de estudos de caso que avaliem a eficácia do modelo em

Revista ARACÊ
ISSN: 2358-2472

diferentes plataformas de mídias sociais pode ajudar a calibrar a solução para contextos variados, contribuindo para uma implementação prática mais robusta. Esses avanços podem enriquecer a pesquisa e ampliar a eficácia da detecção automática de discurso de ódio.

Revista ARACÊ
ISSN: 2358-2472

REFERÊNCIAS

ARCILA-CALDERÓN, Carlos**;** AMORES, Javier J.; SÁNCHEZ-HOLGADO, Patricia; BLANCO-HERRERO, David. Using shallow and deep learning to automatically detect hate motivated by gender and sexual orientation on Twitter in Spanish. **Multimodal Technologies and Interaction**, Basel, v. 5, n. 10, p. 63, out. 2021. Disponível em: https://www.mdpi.com/2414-4088/5/10/63. Acesso em: 5 out. 2024. DOI: 10.3390/mti5100063.

ASHRAF, Nsrin; TAHA, Mohamed; ABD ELFATTAH, Ahmed; NAYEL, Hamada. NAYEL @LT-EDI-ACL2022: Homophobia/Transphobia Detection for Equality, Diversity, and Inclusion using SVM. In: **LT-EDI 2022 - 2nd Workshop on Language Technology for Equality, Diversity and Inclusion**, Proceedings of the Workshop, 2022, p. 287–290. Disponível em: https://aclanthology.org/2022.ltedi-1.42/. Acesso em: 10 dez. 2024. DOI: 10.18653/v1/2022.ltedi-1.42.

BUREAU OF JUSTICE STATISTICS. *National Crime Victimization Survey (NCVS)*. Washington, D.C.: U.S. Department of Justice, 2024. Disponível em: https://bjs.ojp.gov/data-collection/ncvs. Acesso em: 12 nov. 2024.

CHAKRAVARTHI, Bharathi Raja et al. Dataset for identification of homophobia and transphobia in multilingual YouTube comments. **arXiv preprint**, 2021. Disponível em: https://arxiv.org/abs/2109.00227. Acesso em: 16 nov. 2024.

CHAKRAVARTHI, Bharathi Raja et al. Detection of homophobia and transphobia in YouTube comments. **International Journal of Data Science and Analytics**, [S.l.], v. 18, n. 1, p. 49–68, jun. 2024. Disponível em: https://link.springer.com/article/10.1007/s41060-023-00400-0. Acesso em: 24 out. 2024. DOI: 10.1007/s41060-023-00400-0.

CENTER FOR COUNTERING DIGITAL HATE. Digital Hate - Social Media's Role in Amplifying Dangerous Lies About LGBTQ+ People. 10 set. 2022. Disponível em: https://counterhate.com/research/digital-hate-lgbtq/. Acesso em: 17 set. 2024.

FLORES, Andrew R**.** et al. Violent victimization at the intersections of sexual orientation, gender identity, and race: National Crime Victimization Survey, 2017–2019. **PLOS ONE**, [S.l.], v. 18, n. 2, p. e0281641, fev. 2023. Disponível em: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0281641. Acesso em: 18 nov. 2024. DOI: 10.1371/journal.pone.0281641.

MÔNICO, João Francisco Galera et al. Acurácia e precisão: revendo os conceitos de forma acurada. *Revista Brasileira de Ensino de Ciência e Tecnologia*, v. 2, n. 3, p. 107–121, 2009. DOI: 10.17616/R31NJN.

HUGGING FACE. Hugging Face. Disponível em: https://huggingface.co/. Acesso em: 2 dez. 2024.

JOACHIMS, Thorsten. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In: **ICML**, 1997, p. 143–151.

MCGIFF, Josh; NIKOLOV, Nikola S. Bridging the Gap in Online Hate Speech Detection: A Comparative Analysis of BERT and Traditional Models for Homophobic Content Identification on

Revista ARACÊ
ISSN: 2358-2472

X/Twitter. **arXiv preprint**, mai. 2024. Disponível em: https://arxiv.org/abs/2405.09221v1. Acesso em: 3 jan. 2025.

KAGGLE. Kaggle. Disponível em: https://www.kaggle.com/. Acesso em: 3 ago. 2024.

WOOD, Kate. Anti-LGBT Cyberbullying Texts. **Kaggle**, 2023. Disponível em: https://www.kaggle.com/datasets/kw5454331/anti-lgbt-cyberbullying-texts. Acesso em: 25 ago. 2024.

KIM, Jinhyo**.** Iterated grid search algorithm on unimodal criteria. Virginia Polytechnic Institute and State University, 1997.

LIANG, Jingsai. Confusion Matrix: Machine Learning. **POGIL Activity Clearinghouse**, v. 3, n. 4, dez. 2022. Disponível em: https://pac.pogil.org/index.php/pac/article/view/304. Acesso em: 6 nov. 2024.

MACAVANEY, Sean et al. Hate speech detection: Challenges and solutions. **PLOS ONE**, [S.l.], v. 14, n. 8, p. e0221152, ago. 2019. Disponível em: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0221152. Acesso em: 12 nov. 2024. DOI: 10.1371/journal.pone.0221152.

MARIANO, Diego. *Métricas de avaliação em machine learning*. 25 abr. 2021. Disponível em: https://diegomariano.com/metricas-de-avaliacao-em-machine-learning/. Acesso em: 17 nov. 2024.

MARTINS, Kennedy da Nobrega; RODRIGUES, Alexandre Manuel Lopes. **Democracia em rede: o papel dos algoritmos na liberdade de expressão e no pluralismo político**. *Revista Aracê*, São José dos Pinhais, v. 6, n. 3, p. 10785–10805, 2024. DOI: 10.56238/arev6n3-384. Acesso em: 19 abr. 2025.

MOHAMMED, Roweida et al. *Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results*. In: **2020 11th International Conference on Information and Communication Systems (ICICS)**, 2020. p. 243–248. DOI: https://doi.org/10.1109/ICICS49469.2020.239556.

NAKAS, Christos T. et al. *ROC Analysis for Classification and Prediction in Practice*. [S.l.]: CRC Press, maio 2023. DOI: https://doi.org/10.1201/9780429170140.

NIRMAL, Niraj et al. *Automated Detection of Cyberbullying Using Machine Learning*. 2020.

SACHDEVA, Pratik et al. *Constructing interval variables via faceted Rasch measurement and multitask deep learning: a hate speech application*. HuggingFace, 2022. Disponível em: https://huggingface.co/datasets/ucberkeley-dlab/measuring-hate-speech. Acesso em: 12 dez. 2024.

REDDIT. 2024. Disponível em: https://www.reddit.com. Acesso em: 16 nov. 2024.

SCHRÖER, Christoph et al. *A Systematic Literature Review on Applying CRISP-DM Process Model*. **Procedia Computer Science**, v. 181, jan. 2021, p. 526–534. DOI: https://doi.org/10.1016/j.procs.2021.01.199.

Revista ARACÊ
ISSN: 2358-2472

SCIKIT-LEARN. *Scikit-learn: Machine Learning in Python*. 2024. Disponível em: https://scikit-learn.org/. Acesso em: 12 nov. 2024.

SHANMUGAVADIVEL, Kogilavani et al. *KEC-AI-NLP@LT-EDI-2024: Homophobia and Transphobia Detection in Social Media Comments using Machine Learning*. In: **LT-EDI 2024 -** Language Technology for Equality, Diversity and Inclusion, 2024. p. 200–205. Disponível em: https://aclanthology.org/2024.ltedi-1.23. Acesso em: 4 out. 2024.

THE TREVOR PROJECT. *2023 U.S. National Survey on the Mental Health of LGBTQ Young People*. 2023. Disponível em: https://www.thetrevorproject.org/survey-2023/. Acesso em: 6 nov. 2024.

TORGO, Luis; RIBEIRO, Rita. *Precision and Recall for Regression*. In: **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, v. 5808 LNAI, 2009. p. 332–346. DOI: https://doi.org/10.1007/978-3-642-04747-3_26.

VIANNA, Ana Maria dos Santos; PARETO, Eduardo Luiz; BIANCHI, José Mauro Batista. **Guerra cognitiva nas redes sociais: ameaças, desafios e estratégias de mitigação**. *Revista Aracê*, São José dos Pinhais, v. 7, n. 3, p. 14287–14303, 2025. DOI: 10.56238/arev7n3-240. Acesso em: 19 abr. 2025. X (anteriormente Twitter)**.** 2024. Disponível em: https://x.com. Acesso em: 5 ago. 2024.

YIN, Wenjie**;** ZUBIAGA, Arkaitz. *Towards generalisable hate speech detection: a review on obstacles and solutions*. **PeerJ Computer Science**, v. 7, fev. 2021. p. 1–38. DOI: https://doi.org/10.7717/peerj-cs.598.

YOUTUBE. 2024. Disponível em: https://www.youtube.com. Acesso em: 10 set. 2024.

ZHANG, Dell et al. *Estimating the uncertainty of average F1 scores*. In: **ICTIR 2015 - Proceedings of the 2015 ACM SIGIR International Conference on the Theory of Information Retrieval**, set. 2015. p. 317–320. DOI: https://doi.org/10.1145/2808194.2809488.