Discurso de Ódio no *Twitter* em Língua Inglesa: Detecção e Análise por Meio de Técnicas de *Machine Learning*

Anderson Soares de Santana Júnior[1], **Antônio Carlos Bispo Cunha**[2], **Bruno Henrique Carneiro da Silva**[3], **João Felipe Tunes Oliveira**[4], **José Gustavo Abreu Alves**[5]

[1] [1]Departamento de Computação (DCOMP) – Universidade Federal de Sergipe (UFS)
Av. Marechal Rondon, s/n – Jardim Rosa Elze – CEP 49100-000
São Cristóvão – SE – Brazil

{bruno.carneiro, joao.tunes}@dcomp.ufs.br

{antonio0407, and_soares, keysos}@academico.ufs.br

Resumo. *O estudo investiga a detecção automática de discurso de ódio em tweets em inglês usando Machine Learning e Transfer Learning. Analisa características linguísticas, ironia e abreviações que dificultam a classificação de mensagens ofensivas. Modelos pré-treinados como BERT, DistilBERT, e XLM-R são avaliados para identificar conteúdo discriminatório em larga escala.*
Palavras-chave*: Discurso de ódio, Redes sociais, Machine Learning, Transfer Learning.*

Abstract. *The study investigates automatic detection of hate speech in English tweets using Machine Learning and Transfer Learning. It analyzes linguistic features, irony, and abbreviations that hinder classification of offensive messages. Pre-trained models such as BERT, DistilBERT, and XLM-R are evaluated to identify discriminatory content at large scale.*
Keywords*: Hate speech, Social networks, Machine Learning, Transfer Learning.*

1. Introdução

O avanço das Tecnologias da Informação e Comunicação transformou a forma como informações são produzidas, distribuídas e consumidas, como já argumentava Manuel Castells ao destacar que essas tecnologias reconfiguram os processos de produção e difusão da informação. O discurso de ódio emerge como fenômeno, impulsionado pela disseminação e pelo anonimato parcial nas redes sociais. Estudos mostram que grande parte desse conteúdo ofensivo é produzido em língua inglesa, em contextos de polarização social e eventos globais (Ali et al., 2022; Khanday et al., 2022). Este trabalho concentra-se na investigação da detecção automática de discurso de ódio em inglês, com foco em dados do Twitter.

A questão de pesquisa que orienta este estudo é: como os modelos de Machine Learning identificam o discurso de ódio em língua inglesa no Twitter? Pesquisas apontam que a complexidade linguística e a velocidade de circulação das mensagens aumentam o problema, especialmente durante períodos de instabilidade social, como a pandemia de COVID-19 (Khanday et al., 2022). Usuários empregam ironias, gírias, abreviações e variações ortográficas para burlar sistemas automáticos de moderação, reduzindo o desempenho de métodos tradicionais (Ali et al., 2022). Estudos indicam a necessidade de abordagens baseadas em modelos pré-treinados e arquiteturas profundas para lidar com nuances semânticas.

O tema se justifica pela incidência de discurso de ódio nas redes sociais e pelas consequências sociais, psicológicas e políticas desse conteúdo em grupos vulneráveis (Ali et al., 2022). A disseminação de mensagens discriminatórias reforça estereótipos, aumenta tensões sociais e contribui para práticas de violência simbólica e discursiva em ambientes digitais (Ali et al., 2022). A pandemia aumentou comportamentos hostis e a circulação de narrativas ofensivas, tornando a detecção automática importante para reduzir impactos na sociedade (Khanday et al., 2022). Estudar métodos capazes de identificar e conter o fenômeno apresenta necessidade urgente.

A relevância científica deste estudo está no avanço de técnicas que combinam Machine Learning, Transfer Learning e Processamento de Linguagem Natural para lidar com desafios linguísticos (Ali et al., 2022). Pesquisas indicam que modelos como BERT, DistilBERT e XLM-R apresentam desempenho superior a algoritmos tradicionais, mas ainda enfrentam limitações com ambiguidade, ironia e criatividade linguística dos usuários (Ali et al., 2022). Resultados podem contribuir para sistemas de moderação automática em plataformas digitais, auxiliando na redução de conteúdo ofensivo. Grande parte da literatura trabalha com inglês, espanhol e árabe. Neste estudo, o foco está no inglês devido a três fatores: (1) grande volume de dados disponíveis; (2) ampla presença de pesquisas anteriores, permitindo comparação; (3) acesso a modelos amplamente treinados, com menor necessidade de adaptação. O objetivo do estudo é desenvolver e avaliar modelos de Machine Learning e Transfer Learning para detecção de discurso de ódio no Twitter, com base em tweets em inglês e nas abordagens de Ali et al. (2022) e Khanday et al. (2022). Os objetivos específicos são:

1. Identificar um conjunto de tweets em língua inglesa com discurso de ódio.
2. Aplicar algoritmos de *Machine Learning* para comparação de desempenho inicial.
3. Treinar modelos baseados em *Transfer Learning*, como *BERT* e *DistilBERT*.
4. Comparar os resultados obtidos com os achados de Ali et al. (2022) e Khanday et al. (2022).

2. Referencial Teórico

O referencial teórico apresenta conceitos, definições e abordagens aplicadas à detecção automática de discurso de ódio em tweets em inglês. Inicialmente, discute o contexto das Tecnologias da Informação e Comunicação (TIC) e plataformas digitais, com ênfase no Twitter como espaço de interação e debate público. Em seguida, aborda o discurso de ódio, sua propagação e impacto social. Posteriormente, apresenta técnicas de Machine Learning e Transfer Learning utilizadas na classificação de mensagens. Por fim, descreve desafios relacionados à análise automatizada de textos curtos, como complexidade linguística. O referencial fundamenta métodos aplicáveis à análise de tweets em inglês e à implementação de sistemas de detecção automática.

2.1. Tecnologias da Informação e Comunicação e Redes Sociais

As Tecnologias da Informação e Comunicação estruturam fluxos informacionais que definem modos de produção textual em ambientes digitais (CASTELLS, 1999). O Twitter apresenta organização baseada em mensagens curtas que circulam em ritmo contínuo. A plataforma permite observação direta de postagens públicas e formação de corpus amplo. A limitação de caracteres produz estruturas condensadas que influenciam análises textuais. Usuários escrevem com abreviações, símbolos e hashtags que alteram a composição do texto. A publicação ocorre em sequência rápida e gera volume elevado

de dados. A estrutura digital permite coleta sistemática de mensagens. A dinâmica comunicacional condiciona métodos de preparação dos dados. O ambiente forma base para processos de classificação.

Plataformas digitais operam como espaços de circulação permanente de conteúdo produzido por usuários (JENKINS, 2013). O Twitter organiza linhas temporais abertas e acessíveis para observação de interações públicas. Usuários realizam respostas, menções e compartilhamentos que formam cadeias discursivas. A plataforma registra fluxos contínuos que compõem corpus utilizável em pesquisas. A produção textual ocorre por mensagens breves que se conectam em redes amplas. O volume de dados possibilita identificação de padrões linguísticos. A estrutura facilita coleta e armazenamento de mensagens. O ambiente permite análises voltadas a comportamentos discursivos. O conteúdo disponibilizado configura campo para investigação.

Textos em ambientes digitais apresentam variação linguística significativa. A informalidade do Twitter inclui abreviações, neologismos e emojis. O processamento automático precisa lidar com heterogeneidade textual. A representação adequada de dados melhora desempenho de classificadores. Sistemas dependem de limpeza e normalização antes da análise. Toraman et al. (2022) mostram que variabilidade textual interfere na consistência de resultados. A extração de padrões linguísticos exige métodos robustos. A forma reduzida dos tweets limita contexto e reduz pistas semânticas. A análise precisa identificar relações relevantes entre tokens.

Transfer Learning permite adaptação de modelos pré-treinados a diferentes ambientes digitais. A técnica reduz necessidade de treinar modelos do zero. A abordagem facilita interpretação de textos curtos com padrões variados. Priyadarshini et al. (2023) indicam que ajustes de modelos melhoram estabilidade de classificação. A utilização de embeddings contribui para captação de proximidade semântica entre termos. A adaptação possibilita aplicação em plataformas distintas. O ajuste direciona pesos internos para categorias específicas. Modelos pré-treinados capturam relações sintáticas e semânticas. O processo reduz custo computacional e melhora viabilidade experimental.

As TIC influenciam formas de circulação de conteúdo e estruturam ambientes digitais propícios à análise discursiva (CASTELLS, 1999). O Twitter apresenta configuração que possibilita observação contínua de mensagens. A escrita breve gera formatos de alta densidade textual. A dinâmica da plataforma oferece corpus adequado para investigações linguísticas. A organização do ambiente determina estratégias de coleta e representação. O fluxo informacional orienta delimitação dos processos analíticos. A produção acelerada de mensagens sustenta modelos supervisionados e pré-treinados. O conjunto estrutural da plataforma fornece campo para aplicação de técnicas de classificação. A base teórica fundamenta procedimentos metodológicos.

2.2. Discurso de Ódio no Twitter

O Twitter apresenta fluxo intenso de mensagens que incluem ataques direcionados relacionados a eventos públicos. A plataforma opera com postagens abertas que tornam visíveis práticas discursivas ofensivas. Usuários produzem mensagens curtas que facilitam formulações de conteúdo hostil (ALI et al., 2022). A estrutura reduzida dos tweets gera expressões condensadas que dificultam interpretação automática. A divulgação ocorre em ritmo acelerado e amplia alcance textual. O formato da plataforma permite acesso contínuo ao conteúdo publicado. O corpus resultante possibilita análise de manifestações ofensivas. A composição dos tweets influencia padrões de escrita. A circulação do conteúdo cria cenário adequado para estudo da linguagem hostil.

Picos de mensagens ofensivas surgem em períodos de mobilização pública e produzem volume expressivo de ataques discursivos. A plataforma reúne publicações que refletem tensões sociais representadas linguisticamente. A curta extensão dos tweets concentra referências identitárias e termos agressivos. Usuários utilizam estratégias de abreviação que alteram sinais de hostilidade. A postagem ocorre de modo contínuo e gera sequências de insultos (KHANDAY et al., 2022). A estrutura digital favorece difusão de conteúdos direcionados a grupos específicos. O conjunto de mensagens compõe material de análise para categorização. A linguagem empregada evidencia padrões recorrentes. A organização dos dados permite classificação por modelos.

Eventos envolvendo disputas discursivas geram volume elevado de conteúdo hostil que se registra em intervalos curtos (ALI et al., 2022). Usuários reagem a acontecimentos públicos com mensagens ofensivas estruturadas por termos repetidos. A dinâmica do Twitter concentra respostas rápidas que intensificam circulação textual. A produção de insultos ocorre por uso de palavras-chave que se replicam em diferentes postagens. A plataforma favorece formação de temas por hashtags que agrupam discursos. A presença contínua de ataques produz amostras que permitem estudos quantitativos. A organização da timeline registra padrões recorrentes. O fluxo textual apresenta elementos relevantes para identificação automática. O ambiente expõe formas de hostilidade verbal.

A linguagem do Twitter inclui usos irônicos, termos abreviados e alterações ortográficas que dificultam reconhecimento de hostilidade (KHANDAY et al., 2022). Usuários empregam sinais gráficos que deslocam sentidos aparentes. A escrita curta limita contexto e reduz pistas semânticas. A variação textual interfere no desempenho de classificadores tradicionais. A presença de códigos informais modifica estrutura do texto. A produção heterogênea exige técnicas de representação adequadas. A identificação automática depende de mecanismos que reconheçam padrões implícitos. A plataforma reúne conteúdo que desafia métodos estatísticos. A variabilidade transforma análise em processo complexo.

A ampla disponibilidade de dados no Twitter sustenta criação de bases rotuladas utilizadas em estudos sobre discurso de ódio. O acesso ao conteúdo permite formação de conjuntos destinados a treinamento e teste. A produção contínua de tweets gera material extenso que possibilita aplicação de técnicas de classificação. A coleta ocorre por procedimentos que extraem mensagens com termos específicos. A organização do corpus facilita identificação de categorias de hostilidade. O ambiente oferece volume suficiente para comparação de algoritmos. A produção textual permite análises replicáveis. O conjunto de dados se torna fundamento operacional para modelos automáticos (ALI et al., 2022). A plataforma fornece insumos para pesquisas empíricas.

2.3. *Machine Learning* aplicado à detecção de discurso de ódio

Modelos supervisionados utilizam conjuntos de dados rotulados para identificar padrões linguísticos em mensagens curtas. O texto é convertido em representações numéricas que permitem análise computacional. A classificação depende de vetores que registram distribuição de termos. A etapa de treinamento ajusta parâmetros para reduzir erros. O modelo organiza mensagens em categorias definidas no corpus. O processamento envolve segmentação e normalização do texto. A estrutura estatística orienta previsões futuras. KHANDAY et al. (2022) indicam que o aprendizado supervisionado identifica relações internas entre tokens. O procedimento gera resultados consistentes dentro do conjunto analisado.

O uso de TF-IDF produz vetores que indicam importância relativa de palavras em relação ao corpus total. Os valores representam presença e relevância das palavras em cada documento. A técnica opera sobre frequência e distribuição. A representação permite comparação entre mensagens curtas. A transformação remove dependência de texto bruto. A matriz resultante organiza dados para etapas posteriores. A entrada vetorial estrutura classificadores lineares ou não lineares. A abordagem permite processamento eficiente de grandes volumes de dados. Yuan et al. (2023) mostram que TF-IDF fornece base sólida para classificação de textos.

Embeddings geram representações densas que registram relações semânticas entre tokens. Os vetores aproximam palavras com funções similares e separam termos com usos distintos. A técnica reduz perda de informação presente em métodos esparsos. A representação permite captação de padrões não detectáveis por frequência simples. A estrutura vetorial facilita operação de classificadores profundos. O modelo utiliza aspectos contextuais incorporados no embedding. A abordagem cria espaço contínuo de relações entre termos. Priyadarshini et al. (2023) indicam que embeddings melhoram interpretação de variações de escrita. A representação orienta procedimentos de aprendizado automático.

Modelos supervisionados ajustam parâmetros internos comparando previsões com rótulos verdadeiros. A atualização ocorre por processos iterativos que minimizam erro. O mecanismo permite identificação de padrões recorrentes em mensagens curtas. A operação depende de cálculo de gradientes que alteram pesos internos. O sistema responde à estrutura do corpus e organiza categorias de saída. Yuan et al. (2023) indicam que a repetição de ciclos aprimora desempenho do modelo. A técnica estabelece base para tarefas posteriores. O modelo aprende relações internas de tokens nos tweets. O ajuste fino orienta classificações consistentes em novos conjuntos.

A aplicação de Machine Learning possibilita análise de grandes coleções de tweets com procedimentos padronizados. O processamento automatizado organiza mensagens por regras estatísticas. A escala permite identificação de padrões não perceptíveis em análise manual. O corpus rotulado orienta classificação dos elementos analisados. A operação produz categorias aplicáveis a processos de moderação. O método reduz necessidade de intervenção humana. KHANDAY et al. (2022) indicam que o procedimento fornece base para replicação em diferentes conjuntos. A abordagem fundamenta sistemas capazes de identificar hostilidade. O processamento estruturado permite consistência em avaliações subsequentes.

2.4. *Transfer Learning* e Modelos Pré-Treinados

Transfer Learning reutiliza modelos pré-treinados para aplicar conhecimento aprendido em grandes corpora a novas tarefas. O fine-tuning modifica parâmetros específicos do modelo para adaptação aos dados analisados, permitindo reconhecimento de padrões complexos em mensagens curtas. Representações semânticas previamente aprendidas ajudam na interpretação de relações entre tokens e na classificação de conteúdo. A técnica reduz a necessidade de treinar modelos do zero, tornando o processo mais eficiente em termos de tempo e recursos computacionais (ALI et al., 2022). Embeddings densos registram relações entre termos e orientam ajustes finos necessários para otimização da detecção.

Modelos como BERT, DistilBERT e XLM-R aplicam Transfer Learning em textos curtos, ajustando pesos internos para adaptar representações pré-treinadas ao domínio

dos dados. O modelo captura dependências entre tokens e relações contextuais, permitindo classificação precisa de mensagens ofensivas mesmo em presença de ironia ou abreviações. A reutilização do conhecimento prévio reduz o tempo de treinamento e a necessidade de grandes bases rotuladas, preservando aprendizado adquirido em corpora amplos. Ajustes estratégicos em camadas críticas aumentam a precisão do modelo e permitem análise consistente em diferentes conjuntos de dados (ALI et al., 2022).

O pré-treinamento em grandes corpora fornece representações ricas que registram relações semânticas e sintáticas entre palavras, e o fine-tuning permite que apenas parte dos parâmetros seja modificada, mantendo o aprendizado geral do modelo. Essa abordagem possibilita lidar com variações linguísticas típicas de redes sociais, como gírias e abreviações, sem comprometer a capacidade de classificação. Representações densas captam proximidade semântica entre tokens e distâncias entre termos distintos, organizando os dados para que o modelo identifique padrões complexos de discurso ofensivo (PRIYADARSHINI et al., 2023).

Modelos multilíngues como XLM-R permitem análise de textos provenientes de diferentes idiomas por compartilhamento de parâmetros (YUAN et al., 2023). A estrutura utiliza corpus amplo para aprender relações linguísticas variadas. O procedimento fornece vocabulário expandido que inclui tokens provenientes de múltiplos contextos. O processo permite aplicação em mensagens curtas com características distintas. O ajuste final orienta o modelo para categorias específicas. A operação utiliza embeddings interoperáveis entre idiomas. O mecanismo facilita comparação entre diferentes conjuntos textuais. O modelo permite análise ampla em ambientes digitais. A técnica fornece suporte para tarefas de detecção.

O uso de Transfer Learning reduz necessidade de bases rotuladas extensas e permite adaptação rápida a novos contextos (ALI et al., 2022). O processo aproveita parâmetros aprendidos previamente. A aplicação permite interpretação de mensagens com variação estrutural. O ajuste do modelo concentra etapas em poucas camadas. A técnica estrutura procedimentos de classificação em fluxos organizados. O método viabiliza análises replicáveis em diferentes corpus. O sistema utiliza padrões internos para identificação de categorias. O processo amplia alcance da detecção automática. A abordagem compõe base metodológica para estudos de hostilidade verbal.

3. Referências Bibliográficas

ALI, Raza; FAROOQ, Umar; ARSHAD, Umair; SHAHZAD, Waseem; BEG, Mirza Omer. Hate speech detection on Twitter using transfer learning. Computer Speech & Language, v. 74, p. 101365, jul. 2022. DOI: 10.1016/j.csl.2022.101365.

KHANDAY, Akib Mohi Ud Din; RABANI, Syed Tanzeel; KHAN, Qamar Rayees; MALIK, Showkat Hassan. Detecting Twitter hate speech in COVID-19 era using machine learning and ensemble learning techniques. International Journal of Information Management Data Insights, v. 2, n. 2, p. 100120, nov. 2022. DOI: 10.1016/j.jjimei.2022.100120.

PRIYADARSHINI, I.; SAHU, S.; KUMAR, R. A transfer learning approach for detecting offensive and hate speech on social media platforms. Multimedia Tools and Applications, 2023.

DEVIANTY, Fairuz Astari. Transfer Learning Methods for Hate Speech Detection in Bahasa Indonesia. Journal of Information Technology and Computer Science, 2025.

YUAN, Lanqin; WANG, Tian; FERRARO, Giovanni et al. Transfer learning for hate speech detection in social media. Journal of Computational Social Science, v. 6, p. 1081–1101, 2023.

CASTELLS, Manuel. A sociedade em rede. 2. ed. São Paulo: Paz e Terra, 1999. (A Era da Informação: Economia, Sociedade e Cultura; v. 1).

Jenkins, Henry; Ford, Sam; Green, Joshua. Spreadable Media: Creating Value and Meaning in a Networked Culture. New York: New York University Press, 2013.