# Implementação e Validação de um Sistema Protótipo para a Detecção de Discurso de Ódio Usando Inteligência Artificial

**Gabriel Barreto**[1], **Juan S. Toquica**[1]

[1]Universidade Federal do Ceará, Campus Itapajé, Ceará, Brasil

gabrielbarretogomes01@gmail.com, juan.arenas@ufc.br

***Abstract.*** *This study proposes a prototype system leveraging the BERTimbau model to detect hate speech in chat conversations, specifically tailored to the linguistic and contextual features of Brazilian Portuguese. In this way, the research explores how automated moderation systems can be integrated into digital environments to enhance social interactions. A prototype system was developed and evaluated using a descriptive, quantitative approach, achieving satisfactory performance metrics in identifying offensive language and demonstrating efficient response times. The study highlights the potential of NLP models to improve content moderation in Portuguese, offering valuable insights for academia and the private sector to strengthen digital safety strategies.*

***Resumo.*** *Este estudo propõe um sistema protótipo usando o modelo BERTimbau para detectar discurso de ódio em conversas de chat, especificamente adaptado as características linguísticas e contextuais do português brasileiro. Dessa forma, é analisado como os sistemas de moderação automatizados podem ser integrados em ambientes digitais para aprimorar interações sociais. Um sistema protótipo foi desenvolvido e avaliado usando uma abordagem quantitativa descritiva, alcançando métricas de desempenho satisfatórias na identificação de linguagem ofensiva e demonstrando tempos de resposta eficientes. O estudo destaca o potencial de modelos NLP para melhorar a moderação de conteúdo em português, oferecendo resultados adequados para a academia e o setor privado quer permitam fortalecer estratégias de segurança digital.*

## 1. Introdução

No contexto mundial, assim como no Brasil, a liberdade de expressão nas mídias sociais tem sido alvo de discussões, especialmente em plataformas amplamente utilizadas no país, como Instagram, Facebook e X (anteriormente Twitter). Essas redes promovem a troca de ideias, mas enfrentam o desafio contínuo de moderar o conteúdo de forma eficaz. Embora utilizem sistemas automatizados para identificar e remover conteúdos ofensivos ou de ódio, essas tecnologias frequentemente não cobrem todas as formas de violência e abuso online, deixando muitos usuários expostos a comentários e discursos prejudiciais em uma ampla variedade de tópicos [Llansó et al. 2020].

Modelos de Inteligência Artificial (IA), especificamente modelos de linguagem, como BERTimbau para o português brasileiro, têm sido utilizados para detectar discurso de ódio em textos online. Técnicas de Processamento de Linguagem Natural (em inglês: *Natural Language Processing* - NLP) permitem que sistemas compreendam nuances da linguagem, identificando expressões ofensivas automáticamente. Segundo Nobata et al.

[Nobata et al. 2016], esses sistemas ainda enfrentam desafios, como a dificuldade em lidar com ironias, gírias e o dinamismo da linguagem presente nas redes sociais. No estudo desenvolvido por Schmidt e Wiegand [Schmidt and Wiegand 2017] se destacam a eficácia do aprendizado de máquina na detecção de conteúdo abusivo, mostrando que o uso de modelos de classificação e redes neurais pode otimizar a moderação de comentários. No Brasil, onde o volume de interações online é massivo e a polarização política contribui para a intensificação dos discursos de ódio, essas tecnologias são uma alternativa para criar um ambiente digital mais seguro. Esses sistemas devem permitir uma moderação rápida e precisa, abordando o problema do discurso de ódio de forma eficiente, ao mesmo tempo que consideram as particularidades linguísticas e culturais do país.

O discurso de ódio, frequentemente caracterizado por linguagens que incitam discriminação, violência ou preconceito, tornou-se um problema social crescente nas mídias digitais. Plataformas populares como Facebook, Instagram e X (antigo Twitter) têm implementado políticas de moderação, mas enfrentam dificuldades em detectar todas as particularidades do discurso ofensivo, especialmente as mensagens contextualmente complexas. Segundo [Tufekci 2018], essa lacuna na moderação contribui para a perpetuação de abuso contra as pessoas, prejudicando a experiência de usuários vulneráveis.

No âmbito da NLP pode ser considerada como uma subárea da IA que permite a interação entre computadores e a linguagem de comunicação dos seres humanos. Para realizar essas tarefas complexas, são usados Modelos de Linguagem de Grande Escala (LLM), que aplicam redes neurais profundas para aprender e gerar padrões linguísticos a partir de grandes volumes de texto [Jurafsky and Martin 2008]. Os LLM são ferramentas poderosas no processamento de texto, permitindo que sistemas compreendam, gerem e interajam com a linguagem humana de forma eficiente. Esses modelos utilizam arquiteturas que empregam mecanismos de atenção para analisar diferentes partes do texto simultaneamente [Vaswani 2017]. Cabe destacar que NLP é uma grande área de pesquisa de análise da linguagem, enquanto LLM são modelos de linguagem que fazem parte dessa área.

No contexto da língua portuguesa, o BERTimbau se destaca como um modelo de linguagem baseado na arquitetura BERT, treinado especificamente com dados em português do Brasil. O BERTimbau foi ajustado para diversas tarefas de NLP, incluindo a detecção de discurso de ódio [Souza et al. 2019, Souza et al. 2020]. Existem algumas propostas baseadas em BERTimbau e treinados com dataset específicos de discurso de ódio, como o dataset HateBR [Vargas et al. 2022]. Esse dataset contém múltiplos exemplos de linguagem ofensiva. O treinamento com esse tipo de dataset permite com que os modelos derivados sejam eficazes na identificação de discursos de ódio e linguagem ofensiva, tornando-se uma alternativa valiosa para a moderação de conteúdo em redes sociais, fóruns e plataformas de comunicação, devido à aos resultados promissores que tem alcançado até o momento, com métricas como a acurácia de 85% e um F1-Score de 87%, por exemplo [Souza et al. 2022].

A moderação eficiente de comentários em plataformas online não só reduz a incidência de discurso de ódio, mas também contribui para um ambiente mais saudável e construtivo para os usuários [Ribeiro et al. 2023]. Também se faz necessária a vigilância constante de todos os comentários postados e anteriores frente a implementação de um possível algoritmo, seguindo as normas postas na Lei Geral de Proteção de Da-

dos, popularmente conhecida por LGPD aprovada em 2018 pelo congresso nacional [Taborda et al. 2024], e que estes constem nos termos de uso de cada sistema de interação entre pessoas, possuindo a capacidade para tomar medidas preventivas.

O restante deste manuscrito está organizado da seguinte forma: a Seção 3 descreve a metodologias proposta, as componentes e funcionamento do sistema para a identificação do discurso de ódio, enquanto a Seção 4 apresenta alguns resultados obtidos e faz uma analise sobre o desempenho atingido pelo prototipo desenvolvido. Finalmente, as conclusões e trabalhos futuros estão descritas na Seção 5.

## 2. Metodologia proposta

O sistema protótipo funciona como um chat com capacidade de classificação de mensagens que permita a identificação do discurso de ódio, utilizando NLP e o BERTimbau. Foi implementada uma arquitetura Cliente/Servidor para operar durante a troca de mensagens entre vários usuários, enquanto são analisados e classificados as mensagens compartilhadas. O dataset utilizado foi o 'HateBR', dividido em duas partes: 80% dos dados para treinamento do modelo e 20% para validação. As frases incluíam categorias como "ofensivas" ou "não ofensivas", com um balanceamento adequado das classes para evitar viés no treinamento. Durante os testes, o sistema detectou diversas expressões discriminatórias que, embora sutis, foram classificadas como ofensivas devido ao contexto. Isso evidencia a relevância da aplicação do BERTimbau para identificar discursos de ódio em diferentes contextos e interações linguísticas.

### 2.1. Comunicação Cliente/Servidor

Neste tipo de arquiteturas o(s) servidor(es) podem fornecem serviços ou recursos na rede para ser solicitados e consumidos pelo(s) cliente(s), recebendo algum tipo de resposta para cada requisição feita. A arquitetura é amplamente usada em redes institucionais, sejam estas de pequeno, médio ou grande porte [Tanenbaum and Wetherall 2010]. Na figura 1 é apresentada a arquitetura Cliente/Servidor básica e considerada neste trabalho.

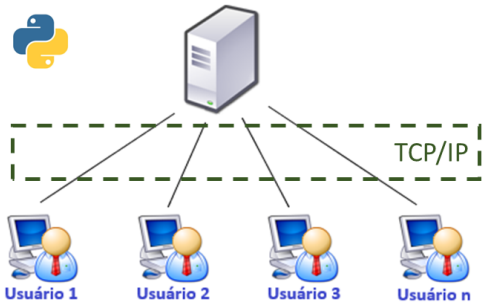


**Figura 1. Sistema baseado na arquitetura Cliente/Servidor.**

Uma das funcionalidades do protótipo proposto é a geração de uma uma lista de clientes banidos quando a comunicação com discurso de ódio é frequente, assim como o registro dos usuários que estão conectados no Chat.

### 2.2. Processo de classificação das mensagens

Foi utilizado o modelo pre-treinado fornecido pela biblioteca *Transformers* [Chaves Rodrigues et al. 2023], previamente treinado com o dataset "*HateBR*" e

baseado no modelo BERTiambau, especializado em identificar discurso ofensivo em português do Brasil, facilitando desta forma a implementação de forma local. A classificação de mensagens é fundamental para garantir que o sistema possa identificar e categorizar o conteúdo de forma adequada. A seguir são descritas as etapas envolvidas na classificação, desde o recebimento inicial da mensagem até a classificação final, assim como o registro num arquivo das mensagens enviadas: Recepção da mensagem; Tokenização; Classificação; e Registro.

### 2.3. Interface gráfica e funcionamento do sistema

A interface permite visualizar a conversa entre os diferentes usuários de forma simultânea e dinâmica, possibilita listar os usuários banidos e também encerrar a comunicação com o servidor. Quando a interação entre os usuário inicia o sistema de forma automática classifica as mensagens e mostra a conversa processada na interface gráfica, além de enviar as mensagens para os demais usuários conectados no servidor. Na figura 2 se apresenta o funcionamento do Chat no sistema proposto.

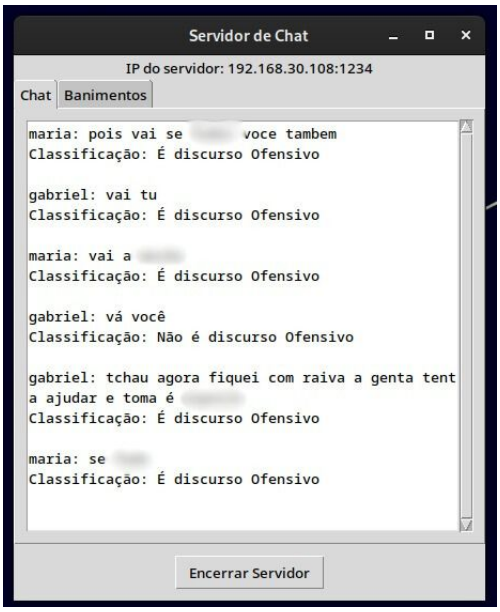


**Figura 2. Exemplo do chat na troca de mensagens entre os usuários**

Foi implementado o método `socket`, responsável por fornecer os mecanismos necessários para a comunicação em rede, assim como a conceito de `threading` que permite gerenciar a execução simultânea de múltiplas tarefas dentro da mesma aplicação.

## 3. Resultados e Discussão

Os testes foram realizados em um ambiente local, simulando uma conversa entre três usuários (clientes) conectados ao servidor via rede LAN, o que permitiu validar o sistema proposto e coletar as informações para a analise respectiva. As interações entre os usuários duravam cerca de trinta minutos e envolveram a troca de mensagens sobre diversos assuntos, incluindo posicionamentos e palavras, algumas vezes camufladas com caracteres especiais.

Foram comparados algumas técnicas de extração de características na literatura para classificação de texto, especificamente com um dataset tradicionalmente usado na identificação de discurso de ódio em idioma Inglês. Alguns dessas técnicas são o *Bag of Words* (BoW) e o TF-IDF, sendo técnicas tradicionais de representação de texto, onde o modelo BoW conta as ocorrências de palavras e TF-IDF as pondera pela importância no texto, mas esses métodos não conseguem capturar o contexto. Por outro lado, as RNNs (Redes Neurais Recorrentes) lidam com sequências, mas têm dificuldades com dependências de longa duração. LSTMs (Memória de Longo Curto Prazo) e GRUs (Unidade Recorrente Gated) melhoram as RNNs ao gerenciar contextos mais longos com mecanismos de portão especializados. Enquanto BoW e TF-IDF são mais simples e precisam menor consumo computacional, LSTMs e GRUs são superiores em tarefas que exigem compreensão contextual profunda, como em Modelos de Linguagem Grande (LLMs). A tabela 1 apresenta as métricas de desempenho de alguns métodos descritos, assim como duas derivações do modelo BERT.

**Tabela 1. Comparação de desempenho entre técnicas para classificação de texto**

| Método/Modelo | Acurácia | Precisão | Recall | F1 Score |
|---|---|---|---|---|
| Bag of Words | 0.878026 | 0.833370 | 0.878026 | 0.838570 |
| TF-IDF | 0.874829 | 0.765325 | 0.874829 | 0.816422 |
| RNN | 0.700777 | 0.800512 | 0.700777 | 0.739817 |
| LSTM | 0.820466 | 0.820810 | 0.820466 | 0.820015 |
| GRU | 0.799452 | 0.818090 | 0.799452 | 0.806804 |
| bert-base-uncased | 0.880000 | 0.808163 | 0.880000 | 0.842553 |
| bert-base-cased | 0.900000 | 0.881304 | 0.900000 | 0.890549 |

A partir dos resultados apresentados pode se observar que o desempenho dos modelos analisados é semelhante, mas existe uma leve melhoria nos dois modelos baseados no BERT, em comparação aos métodos tradicionais de vetorização e os modelos de redes neurais. Posteriormente, foram analisados os resultados obtidos usando o sistema proposto através de vários testes de troca de mensagens, simulando uma conversa entre três usuários conectados ao servidor numa rede local. Na Tabela 2 se apresentam os resultados de um dos teste realizados via chat, onde a implementação do modelo baseado em BERTimbau permite a classificação de mensagens no sistema proposto.

**Tabela 2. Informações básicas derivadas da execução do sistema**

| Identificação das Mensagens | Resultado |
|---|---|
| Número de mensagens classificadas | 235 |
| Número de mensagens não ofensivas | 166 |
| Número de mensagens ofensivas | 69 |
| Tempo total de classificação | 71.26s |
| Tempo médio de classificação | 0.30s |

Para ter uma noção mais clara do desempenho do sistema, foram registradas as mensagens num arquivo e classificadas pelo mesmo modelo utilizado no Chat, mas dessa vez no ambiente Google Colab, devido ao fato de facilitar o uso de bibliotecas especificas para analisar o desempenho da classificação. Na tabela 3 são apresentados os resultados

das métricas resultantes da classificação feita nas mensagens registradas usando a inferencia do modelo BERTimbau:

**Tabela 3. Desempenho do sistema protótipo baseado em BERTimbau**

| Modelo | Sistema protótipo baseado em BERTimbau |
|---|---|
| Acurácia | 0.705882 |
| Precisão | 0.750000 |
| Recall | 0.818182 |
| F1 Score | 0.782609 |

Mesmo que o presente trabalho não considere a analise aprofundada do desempenho da aplicação em relação aos outros trabalhos na área, a proposta descrita neste manuscrito é uma contribuição significativa para o aprimoramento de modelos baseados no português do Brasil em situações especificas. Aproveitando outro teste feito durante os experimentos deste trabalho foram analisadas uma menor quantidade de mensagens resultantes de uma outra conversa via Chat, no caso, ao redor de 80 mensagens, o que permitiu gerar uma matriz de confusão a partir do mesmo modelo no sistema proposto.

O desempenho não foi adequado com menor número de mensagens, mas, o sistema proposto é funcional e consegue fazer a classificação de discurso de ódio, derivando a necessidade de melhorar o desempenho deste tipo de modelos focados em discurso de ódio e nas possíveis atributos que possam aprimorar a classificação final. Os resultados indicam que o sistema protótipo baseado em Bertimbau é uma proposta significativo em direção ao desenvolvimento de ferramentas eficazes para a detecção e moderação de discurso de ódio em ambientes de comunicação digital, mostrando-se competitivo em comparação com outros modelos na área de inteligência artificial.

## 4. Conclusões

Os resultados mostraram que os algoritmos atuais baseados em NLP têm desempenho promissor na distinção entre discursos ofensivos e não ofensivos, especialmente ao lidar com frases complexas que exigem uma compreensão contextual mais aprofundada. O anterior evidencia limitações importantes, mas também oportunidade na identificação do discurso de ódio, como a falta capacidade de captar variantes fundamentais, assim como o impacto de termos isolados que desqualificam indivíduos, levando a classificações inoportunas e/ou interpretações confusas.

Foi desenvolvido um sistema protótipo para a classificação de mensagens enviadas através de chat, com desempenho semelhante em comparação a modelos propostos na literatura, mas em outro tipo de aplicações. A identificação do discurso de ódio num ambiente real utilizando o modelo BERTimbau foi atingida, porém é necessário melhorar o despenho dos modelos nesse tipo de cenários. Estudos futuros podem contribuir na construção de datasets e *benchmarks* que incluam expressões regionais e culturais, devido à diversidade da população e território brasileiro. Além disso, são necessários dataset que abranjam a comunicação intra-comunidades quando for mediada por redes sociais, fortalecendo a inclusão e proteção de comunidades tradicionalmente excluídas. Trabalhos futuros também podem considerar a elaboração e adoção de modelos multimodais, combinando texto, imagens e vídeos, pode surgir como uma metodologia propicia para capturar as dinâmicas das interações online de forma holística.

## Referências


Chaves Rodrigues, R., Tanti, M., and Agerri, R. (2023). Evaluation of Portuguese Language Models.

Jurafsky, D. and Martin, J. H. (2008). *Speech and Language Processing:* An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall, 2 edition.

Llansó, E., Hoboken, J., Leerssen, P., and Harambam, J. (2020). Artificial intelligence, content moderation, and freedom of expression. *Institute for Information Law. https://www. ivir. nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020. pdf*.

Nobata, C., Tetreault, J., Thet, T., and Choi, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, pages 145–153. International World Wide Web Conferences Steering Committee.

Ribeiro, M. H., Santos, C. R., and Silva, A. L. (2023). Moderating hate speech in online platforms: Approaches and challenges. *Journal of Digital Communication*, 15(2):123–139.

Schmidt, A. and Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

Souza, F., Nogueira, R., and Lotufo, R. (2019). Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv:1909.10649*.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In Cerri, R. and Prati, R. C., editors, *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

Souza, M., Oliveira, F., and Gomes, R. (2022). Aplicação de bertimbau para detecção de discurso de Ódio em redes sociais. *Revista Brasileira de Inteligência Artificial*, 23(3):45–62.

Taborda, L. E., de Melo, M. H. C., Luiz, D. E. C., and de Resende Miranda, J. I. (2024). Lei geral de proteção de dados como instrumento de políticas públicas. *OBSERVATÓRIO DE LA ECONOMÍA LATINOAMERICANA*, 22(5):e4908–e4908.

Tanenbaum, A. S. and Wetherall, D. J. (2010). *Computer Networks*. Prentice Hall, 5th edition.

Tufekci, Z. (2018). *Twitter and Tear Gas:* The Power and Fragility of Networked Protest. Yale University Press.

Vargas, F., Carvalho, I., Rodrigues de Góes, F., Pardo, T., and Benevenuto, F. (2022). HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183, Marseille, France. European Language Resources Association.

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.