RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 15 Nro. 1 2025
Uni-FACEF
Centro Universitário Municipal de Franca

ANÁLISE DE SENTIMENTOS NO X:
classificação de *cyberbullying* utilizando modelos de *deep learning*

Maria Clara Piola Colmanetti Campos Borges
Graduando em Ciência da Computação – Uni-FACEF
mariaclarapiola@hotmail.com

Jaqueline Brigladori Pugliesi
Doutora em Ciência da Computação – USP São Carlos
jbpugliesi@gmail.com

Resumo
O crescimento das redes sociais, especialmente o X (antigo Twitter), transformou a comunicação online, mas também potencializou problemas como o *cyberbullying*, uma forma de violência virtual. Este trabalho utiliza o *dataset Cyberbullying Classification*, composto por mais de 47.000 *tweets* classificados em seis categorias, para realizar a classificação de textos por meio de Redes Neurais Recorrentes, com arquitetura *Long Short-Term Memory* (LSTM). A metodologia incluiu a exploração dos dados, o pré-processamento e a aplicação de técnicas de Processamento de Linguagem Natural (PLN). O desempenho do modelo foi avaliado com métricas como acurácia, precisão, recall, F1-score e matriz de confusão para identificar padrões de erro. Os resultados mostraram bom desempenho nas categorias relacionadas a etnia, religião e idade, mas dificuldades na detecção de mensagens ofensivas ligadas a gênero e na distinção entre conteúdos ofensivos e não-ofensivos. Como trabalhos futuros, pretende-se investigar abordagens mais sofisticadas, especialmente modelos baseados em *Transformers*. Assim, este estudo contribui para o avanço das pesquisas em detecção de discursos ofensivos e para a promoção de ambientes digitais mais seguros.

Palavras-chave: *Cyberbullying*; Análise de sentimentos; Aprendizado de Máquina. PLN; Redes sociais.

Abstract
The growth of social networks, especially X (formerly Twitter), has transformed online communication but has also intensified problems such as cyberbullying, a form of virtual violence. This study uses the Cyberbullying Classification dataset, which contains over 47,000 tweets classified into six categories, to perform text classification through Recurrent Neural Networks, using the Long Short-Term Memory (LSTM) architecture. The methodology included data exploration, preprocessing, and the application of Natural Language Processing (NLP) techniques. Model performance was evaluated accuracy, precision, recall, F1-score, and confusion matrix to identify error patterns. The results showed good performance in categories related to ethnicity, religion, and age but difficulties in detecting offensive messages related to gender and in distinguishing between offensive and non-offensive content. For future work, more sophisticated approaches are intended to be explored, especially Transformer-based models. Thus, this study contributes to advancing research on the detection of offensive speech and to promoting safer digital environments.

Keywords: *Cyberbullying; Sentiment Analysis; Machine Leaming; NLP; Social Media.*

1 Introdução

Com o avanço acelerado da internet e o uso massivo das Tecnologias da Informação e Comunicação (TICs), a quantidade de dados gerados globalmente tem crescido exponencialmente. Grande parte desses dados está diretamente relacionada à linguagem humana, manifestando-se de diversas formas, como textos, vídeos e imagens. Atualmente, estima-se que a internet contenha dezenas de trilhões de páginas de informações, com um volume de dados que dobra aproximadamente a cada dois anos. Além disso, o número de usuários conectados continua a crescer rapidamente, impulsionado pelo aumento do acesso móvel e da digitalização de diversos setores (Statista, 2025).

Acompanhando o crescimento do acesso à internet, observa-se um acelerado aumento no uso das redes sociais. Essas plataformas permitem o compartilhamento de informações, promovendo debates e conectando pessoas em escala global. Entre os aspectos positivos, destaca-se a facilidade de acesso à informação, a mobilização social para causas relevantes e o fortalecimento do engajamento político e educacional. O X (antigo Twitter), por exemplo, se destacou por sua dinâmica de compartilhamento em massa, na qual qualquer usuário pode publicar e interagir com conteúdos de forma simples e instantânea (Aslam, 2020), tornando-se um dos principais meios para a disseminação de notícias e discussões em tempo real, contribuindo para a amplificação de vozes e perspectivas diversas.

No entanto, o uso dessas plataformas também apresenta desafios significativos, especialmente no que diz respeito ao *cyberbullying*. A estrutura do X, que permite a rápida viralização de conteúdos e o anonimato de muitos usuários, facilita a existência de discursos de ódio e ataques virtuais. Muitas vezes, indivíduos são expostos a assédios, insultos e ameaças, que resultam em impactos profundos na saúde mental das vítimas levando a consequências psicológicas graves, como ansiedade, depressão e baixa autoestima, especialmente entre adolescentes e jovens adultos (Hinduja; Patchin, 2019). A dificuldade de monitoramento e a moderação limitada desses conteúdos agravam ainda mais o problema, tornando essencial a implementação de estratégias para identificar e mitigar esse tipo de comportamento prejudicial nas redes sociais.

Diante desse contexto, esta pesquisa propõe a análise de *tweets* coletados da plataforma X, com o objetivo de classificá-los de acordo com diferentes tipos de *cyberbullying*, tais como ataques relacionados a idade, etnia, gênero, religião, outras formas de ofensa e a categoria “não é *cyberbullying*”. Para isso, é utilizado o modelo de Redes Neurais Recorrentes, com foco na arquitetura *Long Short-Term Memory* (LSTM) aprimorado com um mecanismo de atenção. A proposta visa contribuir para o desenvolvimento de uma ferramenta automática de detecção de discursos ofensivos, auxiliando na construção de ambientes digitais mais seguros e incentivando um uso mais consciente das redes sociais.

Para alcançar esse objetivo geral, a pesquisa foi conduzida em etapas. Primeiramente, foi realizada uma exploração do *dataset* “*Cyberbullying Classification*”, que contém mais de 47.000 *tweets* rotulados conforme diferentes categorias de *cyberbullying*. Esse conjunto de dados foi balanceado para conter aproximadamente 8.000 exemplos em cada classe, garantindo uma distribuição equilibrada para o treinamento dos modelos. Após a análise inicial, os dados foram pré-processados com

a remoção de ruídos, como *stopwords* e caracteres especiais, e submetidos a técnicas de normalização textual. Em seguida, foram aplicadas estratégias de Processamento de Linguagem Natural (PLN) para converter os textos em representações numéricas adequadas à modelagem.

Após essa etapa, implementou-se o modelo de Redes Neurais *Long Short-Term Memory* fazendo uso de mecanismo de atenção, com o intuito de avaliar o desempenho desses classificadores na identificação automática de postagens ofensivas. O algoritmo foi testado na tarefa de classificar os *tweets* de acordo com as categorias previamente rotuladas no *dataset*, e a performance do modelo foi analisada por meio de métricas amplamente utilizadas em tarefas de classificação, como acurácia, precisão, *recall* e *F1-score*.

Por fim, esta pesquisa também buscou explorar possíveis aprimoramentos, discutindo a viabilidade de técnicas mais avançadas para aplicações futuras. Os resultados obtidos serão úteis no avanço das pesquisas sobre *cyberbullying* e servirão como base para o desenvolvimento de ferramentas mais eficazes de monitoramento e combate ao discurso de ódio nas redes sociais.

2 Referencial Teórico

Nesta seção, será exposta a fundamentação teórica para o desenvolvimento deste trabalho: os principais conceitos e as ferramentas que foram utilizadas para obter os resultados.

2.1 Inteligência Artificial e Aprendizado de Máquina

A Inteligência Artificial (IA), conceito formalizado por John McCarthy (1956), representa um vasto campo da Ciência da Computação dedicado ao desenvolvimento de sistemas que buscam simular a inteligência humana. Conforme detalhado por Russell e Norvig (2009), o objetivo é capacitar máquinas a realizar tarefas que tipicamente exigiriam cognição humana, como o raciocínio lógico, a aprendizagem a partir da experiência, a tomada de decisões complexas e o reconhecimento de padrões. O avanço expressivo da IA nas últimas décadas é indissociável do aumento exponencial do poder computacional e da crescente disponibilidade de grandes volumes de dados (*Big Data*), que servem como combustível para os algoritmos.

Dentro deste universo, o Aprendizado de Máquina (AM), ou *Machine Learning*, emerge como um subcampo fundamental. O AM foca especificamente na criação de algoritmos que permitem aos computadores aprenderem padrões diretamente dos dados e, com base nesse aprendizado, fazer previsões ou tomar decisões sem a necessidade de programação explícita para cada cenário específico (Alpaydin, 2010). Utilizando modelos estatísticos e matemáticos, esses sistemas ajustam seus parâmetros internos e aprimoram seu desempenho à medida que são expostos a mais informações. Contudo, é crucial reconhecer as limitações inerentes ao AM: sua performance é fortemente dependente da qualidade e representatividade dos dados de entrada, e existe o risco de que vieses presentes nos dados sejam aprendidos e, consequentemente, perpetuados pelos modelos, caso não haja um tratamento cuidadoso durante o desenvolvimento.

O Aprendizado de Máquina se manifesta em diversas abordagens, sendo as três principais o aprendizado supervisionado, o não supervisionado e o por reforço (Pedregosa *et al.*, 2011). O aprendizado supervisionado opera com base em dados previamente rotulados, ou seja, exemplos onde a resposta correta é conhecida. Por exemplo, para treinar um sistema a diferenciar imagens de gatos e cães, é fornecido um conjunto de imagens onde cada uma está explicitamente marcada como "gato" ou "cachorro". O modelo aprende as características distintivas de cada classe para poder classificar novas imagens. Algoritmos como Regressão Logística, Máquinas de Vetores de Suporte (SVM) e Árvores de Decisão são exemplos clássicos desta abordagem. Os rótulos são, portanto, essenciais, indicando a categoria ou valor associado a cada dado de treinamento.

Em contraste, o aprendizado não supervisionado lida com dados brutos, sem rótulos ou respostas predefinidas. A tarefa do algoritmo é descobrir de forma autônoma padrões, similaridades ou estruturas latentes nos dados. Isso pode envolver agrupar dados semelhantes, como segmentar clientes com perfis de consumo parecidos utilizando algoritmos como K-Means ou Agrupamento Hierárquico, ou reduzir a dimensionalidade dos dados para visualização ou processamento mais eficiente, utilizando técnicas como a Análise de Componentes Principais (PCA). Esta abordagem é valiosa para a exploração inicial de dados e a descoberta de *insights* inesperados.

Já o aprendizado por reforço adota uma dinâmica de aprendizado por interação. Um agente (o sistema) aprende tomando ações em um ambiente e recebendo *feedback* na forma de recompensas (positivas) ou punições (negativas), dependendo se a ação contribui ou não para atingir um objetivo predefinido. O agente busca desenvolver uma estratégia (política) que maximize a recompensa acumulada ao longo do tempo. Esta modalidade é amplamente utilizada no desenvolvimento de IA para jogos, em robótica para controle de movimentos, na otimização de sistemas autônomos e em sistemas de recomendação personalizados.

No âmbito do aprendizado supervisionado, a tarefa de classificação ocupa um papel central, tendo como objetivo principal a atribuição de instâncias a categorias discretas e previamente definidas. A classificação pode assumir a forma binária, quando lida com duas classes mutuamente exclusivas, como prever se haverá “chuva” ou “não chuva”, ou identificar se uma transação é “fraudulenta” ou “legítima” (Provost; Fawcett, 2013). Alternativamente, pode ser uma classificação multi-classe, quando existem três ou mais categorias possíveis, como categorizar notícias em “esportes”, “política” e “economia”, ou ainda reconhecer diferentes tipos de objetos em imagens (Aly, 2005).

É importante destacar a diferença entre classificação e regressão, outra tarefa supervisionada. Enquanto a classificação foca em prever rótulos discretos, a regressão tem como objetivo a predição de valores contínuos, como a estimativa do preço de um imóvel ou da temperatura de uma região (Bishop, 2006; Shalev-Shwartz; Ben-David, 2014). No presente estudo, a abordagem adotada é a classificação supervisionada de texto, aplicada especificamente à análise de *tweets*, com o propósito de identificar e categorizar diferentes formas de *cyberbullying* em textos online.

2.2 Processamento de Linguagem Natural

Processamento de Linguagem Natural (PLN) é uma área interdisciplinar da ciência da computação que busca desenvolver sistemas capazes de compreender, interpretar e gerar linguagem humana de maneira natural. Refere-se por “linguagem natural” uma linguagem que é usada para comunicações do dia a dia feitas por humanos; línguas como Português, Inglês ou Mandarim. Em contraste às linguagens artificiais, como linguagens de programação e notações matemáticas, as linguagens naturais têm evoluído à medida que passam de geração para geração, e são difíceis de definir com regras explícitas (Bird; Klein; Loper, 2009). Fundamentado em estatística, aprendizado de máquina e linguística computacional, o PLN tem aplicações que variam desde a tradução automática até a análise de sentimentos em textos (Manning; Schütze, 1999).

Os autores Nitin Indurkhya e Fred Damerau (2010) expõem como o PLN lida com situações complexas, como anáforas e ambiguidades. O Processamento de Linguagem Natural opera por meio de diversas representações de conhecimento, como léxicos, que associam palavras aos seus significados, regras gramaticais da linguagem, sinônimos, abreviações e ontologias que mapeiam entidades e ações. Diferentemente de tratar o texto apenas como uma sequência de caracteres, o PLN reconhece a estrutura hierárquica da linguagem, permitindo uma análise mais profunda do significado das palavras em seus contextos.

Contudo, a linguagem humana, frequentemente falada ou escrita de maneira imperfeita, apresenta um desafio significativo. Para que um sistema compreenda a linguagem natural, ele precisa ir além da interpretação literal das palavras e compreender os conceitos subjacentes e suas relações, que em conjunto formam os significados.

Por fim, a ambiguidade semântica, as variações culturais e o uso de gírias ou expressões idiomáticas tornam essa área de estudo um desafio ainda mais complexo. Sistemas de PLN precisam lidar com diferentes contextos e intenções, o que exige o uso de técnicas avançadas, como modelos baseados em aprendizado profundo, redes neurais e algoritmos de *machine learning*. Esses métodos permitem que o sistema aprenda padrões a partir de grandes volumes de dados, melhorando sua capacidade de interpretação e resposta. Assim, o PLN evolui continuamente, tornando-se uma ferramenta fundamental para aplicações em áreas como saúde, educação, negócios e, especialmente, na análise de sentimentos, onde a compreensão precisa das emoções humanas é essencial para gerar entendimentos relevantes.

2.2.1 Pré-processamento de texto

Muitas vezes, dados em formato de texto contém ruídos, como erros de grafia, pontuação e símbolos, que podem atrapalhar os algoritmos de AM. Assim, a primeira etapa do pipeline dos modelos que lidam com linguagem natural, é o pré-processamento de texto (Jurafsky; Martin, 2019). O objetivo é transformar o texto para um formato que pode ser facilmente analisado e processado por algoritmos de AM, e inclui várias etapas, como:

- **Letras minúsculas:** converter todo o texto em letras minúsculas para garantir a consistência do texto e remover a distinção entre letras maiúsculas e minúsculas.

- **Remoção de pontuação:** Remoção de todos os sinais de pontuação do texto para simplificá-lo e otimizar seu processamento.
- **Tokenização:** quebrar o texto em unidades menores, como palavras ou frases individuais, chamados de *tokens*. Esta é uma etapa importante, visto que a maneira mais comum de processar o texto bruto ocorre no nível do *token*.
- **Remoção *stopwords*:** remoção de palavras muito frequentes ou que não agregam significativamente ao texto, e podem afetar os resultados da análise.
- **Stematização e Lematização:** reduzir as palavras à sua forma raiz, diminuindo a dimensionalidade dos dados.
- **Remoção de caracteres especiais:** remoção de quaisquer caracteres especiais, como números e símbolos, que podem ser considerados como ruído para os modelos de AM.

É importante ressaltar que não há um pré-processamento padrão para todos os problemas, e a escolha de quais etapas usar está relacionada aos objetivos da tarefa em questão, dependendo de outros fatores como o tipo de dados e a capacidade de processamento computacional.

2.2.2 Representação Textual

A representação de atributos é uma etapa essencial no PLN. O objetivo é converter dados textuais não estruturados em formatos numéricos ou vetoriais que possam ser processados computacionalmente (Manning; Schütze, 1999). Essa transformação é indispensável, pois algoritmos de aprendizado de máquina não operam diretamente com texto bruto, exigindo uma representação matemática que preserve as características semânticas e sintáticas relevantes do conteúdo original, influenciando diretamente a capacidade do modelo em identificar padrões linguísticos associados a comportamentos ofensivos.

Entre as técnicas mais empregadas para a representação de atributos em PLN, destacam-se o Saco de Palavras (*Bag-of-Words - BoW*), os Bigramas e o TF-IDF (*Term Frequency-Inverse Document Frequency*). Cada uma dessas abordagens possui particularidades que as tornam mais ou menos adequadas dependendo do tipo de análise textual e dos objetivos da classificação (Jurafsky; Martin, 2019).

O *Bag-of-Words* é uma técnica fundamental que transforma textos em vetores numéricos, representando documentos como uma coleção não ordenada de palavras. A representação é feita através de uma matriz onde as colunas correspondem a cada termo do vocabulário do corpus e as linhas representam as frases, com os valores indicando a frequência ou presença de cada palavra. Embora simples e eficaz para identificar padrões baseados na frequência, o *BoW* apresenta limitações significativas, pois não preserva a ordem das palavras nem as relações gramaticais, gera representações esparsas e de alta dimensionalidade, e não captura a correlação entre palavras com significados semelhantes ou compostas, o que é um desafio em tarefas que dependem do contexto e da ordem (Manning; Schütze, 1999).

Para mitigar algumas das limitações do *BoW*, utilizam-se os Bigramas. Essa técnica captura informações sobre a relação entre palavras adjacentes, permitindo uma representação mais rica do texto. Similar ao *BoW*, um vocabulário de

bigramas é construído a partir de todos os pares de palavras adjacentes no corpus, e cada documento é representado por um vetor que indica a frequência de ocorrência desses bigramas (Bird; Klein; Loper, 2009). A representação por bigramas oferece a vantagem de preservar parcialmente a ordem das palavras e capturar relações locais, sendo útil para identificar expressões idiomáticas e negações. Contudo, ela aumenta significativamente a dimensionalidade do espaço vetorial, o que pode levar a problemas de esparsidade e maior custo computacional. Além disso, assim como o *BoW*, os bigramas não capturam relações semânticas complexas.

O *Term Frequency-Inverse Document Frequency* é uma técnica de ponderação que considera não apenas a frequência de um termo em uma frase, mas também sua importância relativa em todo o corpus (Ribeiro, 2015). Termos muito frequentes em todo o corpus recebem um peso menor, pois são menos distintivos, enquanto termos que aparecem poucas vezes recebem um peso maior, indicando maior relevância descritiva. O TF-IDF é calculado multiplicando-se o TF pelo logaritmo da razão entre o número total de frases no corpus e a frequência do termo no corpus. Essa abordagem melhora a capacidade do modelo em diferenciar entre classes de documentos, sendo útil na detecção de *cyberbullying* ao atribuir maior peso a palavras distintivas. Ainda assim, TF-IDF não captura a ordem das palavras nem relações semânticas complexas, sendo frequentemente combinado com outras técnicas de pré-processamento, como *stemming* e lematização, e aplicável a n-gramas.

Mais recentemente, técnicas de *Word Embeddings*, como *Word2Vec* e *GloVe*, revolucionaram a representação de atributos ao mapear palavras para vetores densos em um espaço contínuo. Diferente das outras abordagens, os *embeddings* capturam relações semânticas e sintáticas entre as palavras, onde palavras com significados semelhantes ou que aparecem em contextos parecidos são representadas por vetores próximos. Isso permite que os modelos compreendam o significado contextual das palavras, superando as limitações das técnicas tradicionais que não consideram a ordem ou o sentido das palavras, e oferecendo uma representação mais rica e eficiente para tarefas complexas (Mikolov *et al.*, 2013).

2.3 Análise de Sentimentos

A análise de sentimentos mescla conceitos de Mineração de Dados, Aprendizado de Máquina, linguística, Processamento de Linguagem Natural e análise textual (Silva, 2016). É um campo de estudo que ganhou destaque com o crescimento exponencial da Internet e do conteúdo gerado por seus usuários, especialmente em redes sociais. Nessas plataformas, as pessoas expressam opiniões de forma coloquial, frequentemente utilizando artifícios gráficos, abreviações e *emojis* para tornar seus diálogos mais concisos. Esse cenário trouxe um valor significativo para a análise de sentimentos, dado o impacto dessas opiniões no comportamento individual e coletivo.

As redes sociais, além de facilitar a difusão de informações, desempenham um papel central na sociedade moderna, sendo objeto de interesse de pesquisadores, empresas, jornalistas e governos, uma vez que grande parte da população se concentra nesse ambiente virtual. Elas criam um cenário interativo e dinâmico no qual indivíduos, influenciadores e organizações promovem discussões que podem levar à mobilização em torno de tópicos comuns. Essas plataformas são

um reflexo da revolução digital, permitindo a expressão e a disseminação de emoções e opiniões de maneira quase instantânea em uma proporção jamais antes vista.

Opiniões têm um peso significativo nas decisões das pessoas, sejam elas simples, como escolher um filme, ou mais complexas, como decidir em qual empresa investir. As organizações, por sua vez, utilizam as opiniões de seus consumidores para guiar estratégias de mercado e aperfeiçoar produtos e serviços. Com o avanço das mídias sociais, tornou-se possível acessar uma vasta quantidade de opiniões e sentimentos, mas o volume gigantesco de dados requer ferramentas automatizadas para processar e interpretar esse conteúdo de maneira eficaz.

A rede social X, por exemplo, é uma das plataformas mais populares e uma das principais fontes para estudos de análise de sentimentos. Caracterizado inicialmente por um limite de 140 caracteres por postagem, posteriormente ampliado para 280, o X permite a divulgação de mensagens curtas, frequentemente em tempo real. Com mais de 500 milhões de usuários ativos mensais, a plataforma é uma rica fonte de dados para análise de opiniões e tendências sociais (Backlinko, 2024).

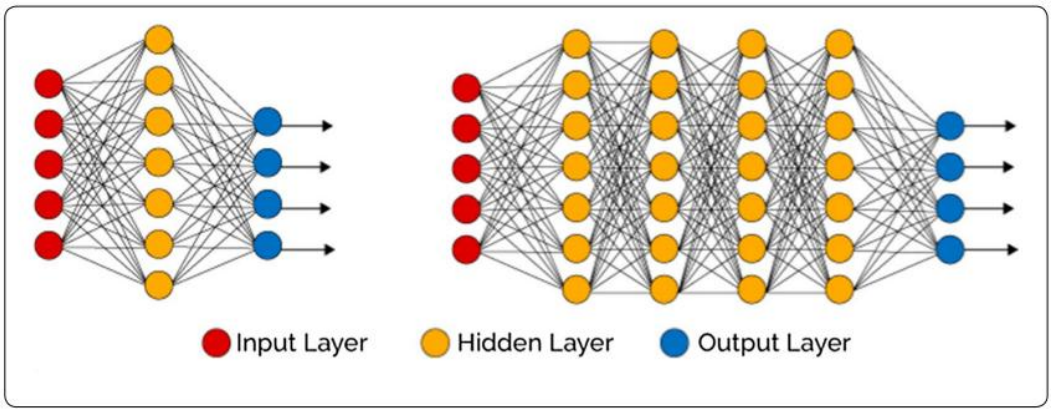
De acordo com Liu (2012), a mineração de opiniões opera sobre textos de diferentes formatos e tamanhos, incluindo páginas web, comentários, *posts* e *tweets*. Cada opinião consiste, no mínimo, em dois elementos centrais: o alvo e o sentimento associado. O alvo pode ser um produto, pessoa, organização, marca ou evento, enquanto o sentimento reflete a atitude ou emoção do autor em relação a esse alvo. A polaridade do sentimento varia em uma escala, indicando se a avaliação é positiva, neutra ou negativa (Tsytsarau; Palpanas, 2012).

Esse campo de pesquisa, portanto, conecta o comportamento humano com a tecnologia, permitindo entender como as opiniões se formam, se disseminam e influenciam decisões individuais e sociais em larga escala.

2.4 Redes Neurais

Redes Neurais são um tipo de modelo de AM inspirado na estrutura e funcionamento do cérebro humano em especial na forma como os neurônios biológicos processam e transmitem informações (Luger; Stubblefield, 1997). Esses modelos são compostos por unidades de processamento interconectadas, chamadas de neurônios artificiais, que são organizadas em camadas. Na Figura 1 é mostrado um exemplo esquemático de uma rede neural com três camadas: de entrada (*input layer*), oculta (*hidden layer*) e de saída (*output labels*) e um exemplo de rede neural profunda com mais camadas ocultas.

Figura 1 - Rede Neural Simples e Rede Neural Profunda.



Fonte: Vázquez, 2017.

Cada neurônio artificial recebe uma ou mais entradas numéricas, as quais são processadas por meio de uma função de ativação e combinadas de acordo com pesos ajustáveis. Esses pesos sinápticos são aprendidos automaticamente durante o treinamento do modelo por meio de algoritmos como o *backpropagation* e a descida do gradiente, permitindo que a rede ajuste seus parâmetros para minimizar o erro nas predições (Haykin, 2009).

As Redes Neurais podem ser classificadas em diferentes tipos, dependendo da arquitetura e do propósito. Modelos mais simples, como *Perceptron* de Camada Única, são utilizados para tarefas de classificação linear. Já as Redes Neurais Profundas (*Deep Neural Networks* – DNNs) possuem múltiplas camadas ocultas e são capazes de aprender representações mais complexas dos dados. São tipicamente formadas por múltiplas camadas de neurônios artificiais.

2.4.1 Redes Neurais Recorrentes

As Redes Neurais Recorrentes (RNNs) representam uma classe especializada de redes neurais artificiais projetadas para processar dados sequenciais, onde a ordem e a dependência temporal dos elementos são cruciais. Diferentemente das redes neurais *feedforward* tradicionais, as RNNs possuem conexões de feedback que permitem que a informação persista e seja utilizada em etapas futuras da sequência. Essa característica confere às RNNs uma espécie de "memória" interna, possibilitando que a saída de um neurônio em um determinado instante de tempo influencie a entrada do mesmo neurônio ou de outros neurônios em instantes subsequentes (Lipton, 2023).

Essa arquitetura as torna particularmente adequadas para tarefas que envolvem séries temporais, como reconhecimento de fala, tradução automática, geração de texto e, notavelmente, o PLN, onde a compreensão do contexto de uma palavra depende das palavras que a precedem. No entanto, RNNs básicas podem enfrentar desafios como o problema do gradiente evanescente (*vanishing gradient*) ou explosivo (*exploding gradient*) em sequências muito longas, o que limita sua capacidade de aprender dependências de longo prazo.

Para mitigar essas questões, variações como as LSTMs foram desenvolvidas, incorporando mecanismos que controlam o fluxo de informações, permitindo que a rede retenha ou esqueça informações seletivamente ao longo do tempo (Hochreiter; Schmidhuber, 1997). Dessa forma, as LSTMs surgem como uma evolução direta das RNNs, projetadas especificamente para superar essas limitações e lidar de maneira mais eficaz com dependências de longo prazo.

2.4.2 *Long Short-Term Memory*

As *Long Short-Term Memory*, propostas por Hochreiter e Schmidhuber (1997), representam uma evolução das Redes Neurais Recorrentes, desenvolvidas com o objetivo de superar limitações relacionadas ao problema do gradiente evanescente e explosivo em sequências longas. Esse tipo de rede introduz um mecanismo de portões — de entrada, esquecimento e saída — que regula de forma seletiva o fluxo de informações ao longo do tempo, possibilitando que a rede retenha dados relevantes por intervalos mais longos e descarte informações irrelevantes.

A principal vantagem das LSTMs está na sua capacidade de capturar dependências de longo prazo em dados sequenciais, aspecto fundamental para tarefas de PLN, como análise de sentimentos, tradução automática e detecção de *cyberbullying*. Por meio da memória controlada pelos portões, o modelo consegue identificar relações complexas entre palavras distantes em um texto, melhorando a interpretação do contexto global (Olah, 2015).

2.4.3 Mecanismos de Atenção

O mecanismo de *Attention* surgiu como uma solução para uma limitação central das arquiteturas recorrentes, incluindo as LSTMs: a dificuldade de capturar dependências de longo prazo em sequências muito extensas. Proposto inicialmente por Bahdanau, Cho e Bengio (2014), o *Attention* introduz a capacidade de o modelo “focar” seletivamente em diferentes partes da sequência de entrada, atribuindo pesos diferenciados a cada elemento conforme sua relevância para a tarefa.

Em vez de depender apenas do estado oculto final da rede recorrente, o *Attention* permite que o modelo acesse diretamente todos os estados ocultos da sequência, calculando uma combinação ponderada que enfatiza as informações mais importantes (Vaswani *et al.*, 2017). Esse mecanismo trouxe avanços significativos em tarefas de tradução automática, análise de sentimentos e compreensão de linguagem natural, pois possibilita que o modelo compreenda relações complexas entre palavras distantes.

Além de melhorar o desempenho, também contribui para a interpretabilidade dos modelos, permitindo visualizar quais partes do texto influenciaram mais a decisão. Essa característica tornou-se a base para arquiteturas mais avançadas, como os *Transformers*, que dispensam estruturas recorrentes e alcançaram o estado da arte em diversas tarefas de PLN.

2.5 Acurácia, Precisão, Recall e F1-Score

Na avaliação de modelos de classificação, é fundamental considerar métricas que permitam mensurar não apenas a taxa de acertos globais, mas também a qualidade das previsões em termos de classes positivas e negativas. A acurácia representa a proporção de classificações corretas em relação ao total de instâncias analisadas. Embora seja amplamente utilizada, pode ser insuficiente em contextos de classes desbalanceadas, pois um modelo pode apresentar alta acurácia mesmo ignorando a classe minoritária (Sokolova; Lapalme, 2009).

Já a precisão mede a proporção de instâncias corretamente classificadas como positivas em relação ao total de instâncias preditas como positivas, indicando o grau de exatidão das previsões. O *recall* (ou sensibilidade) quantifica a capacidade do modelo de identificar corretamente as instâncias positivas entre todas as que realmente pertencem a essa classe. Já o F1-Score corresponde à média harmônica entre precisão e *recall*, equilibrando os dois indicadores e fornecendo uma visão mais robusta do desempenho do modelo, especialmente em cenários de desbalanceamento.

Essas métricas são comumente utilizadas, pois permitem avaliar a eficácia dos modelos na identificação de conteúdo ofensivo, equilibrando erros de falso positivo e falso negativo.

3 Materiais e Métodos

Nesta seção, serão apresentados os materiais, as ferramentas e os métodos utilizados para o desenvolvimento da pesquisa. Será descrito o conjunto de dados adotado, o ambiente computacional, as bibliotecas empregadas, bem como as técnicas de pré-processamento, preparação dos dados e modelagem, que foram aplicadas para alcançar os objetivos propostos.

3.1 *Dataset* utilizado

Com o crescente uso das redes sociais, a maneira como as pessoas se comunicam e interagem *online* foi significativamente alterada. As plataformas digitais se tornaram uma extensão do cotidiano, oferecendo novas formas de expressão, conexão e compartilhamento de informações. Foi possível testemunhar como as redes sociais têm o poder de conectar pessoas, mas proporcionalmente o de criar um terreno fértil para o abuso e a agressão anônima, o que pode ter consequências profundas na saúde mental e emocional de usuários (Hinduja; Patchin, 2019).

O *cyberbullying* é um dos mais alarmantes. Esse fenômeno caracteriza-se pelo uso da tecnologia para assediar, humilhar ou intimidar indivíduos, muitas vezes de forma anônima, o que dificulta a identificação dos agressores. A falta de interação presencial, combinada com o anonimato garantido pelas plataformas digitais, criou um ambiente propício para que discursos de ódio se espalhassem mais facilmente, afetando, especialmente, adolescentes e jovens adultos, que são as principais vítimas desse tipo de violência virtual.

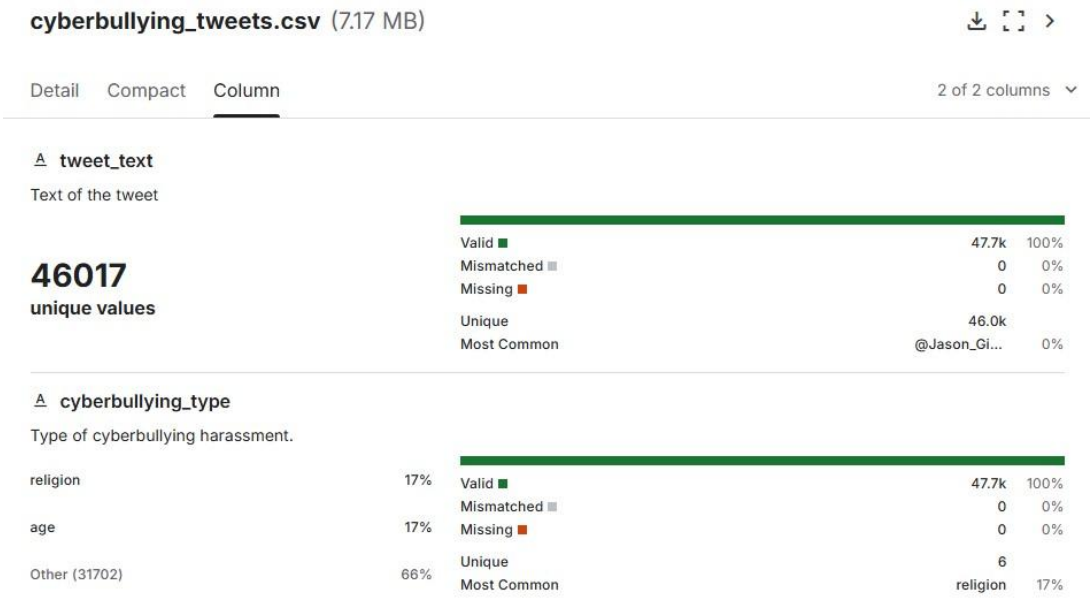
Ademais, o impacto do *cyberbullying* vai além das agressões imediatas, com efeitos duradouros sobre a autoestima, o bem-estar psicológico e até mesmo a motivação para engajamento nas redes sociais. A luta contra o *cyberbullying* exige não só ações individuais, mas também políticas de conscientização, apoio psicológico

e a implementação de tecnologias capazes de detectar e combater esses comportamentos prejudiciais. O entendimento do problema e a implementação de soluções eficazes são passos essenciais para garantir um ambiente digital mais seguro e saudável para todos (Hinduja; Patchin, 2019).

Para compreender e combater esse problema, foi utilizado o *dataset* "*Cyberbullying Classification*", disponível no Kaggle, que contém mais de 47.000 publicações do X, classificados de acordo com diferentes categorias de *cyberbullying*: idade, etnia, gênero, religião, outros tipos de *cyberbullying* e não-*cyberbullying*. Originalmente, a base de dados foi disponibilizada em arquivo CSV contendo apenas os textos das publicações (*tweet_text*) e o tipo de *cyberbullying* (*cyberbullying_type*), como está representado na Figura 2.

De acordo com o criador do conjunto de dados, foram balanceados para conter aproximadamente 8.000 *tweets* em cada classe, garantindo uma distribuição equitativa entre as categorias. Esse conjunto de *posts* permitirá a análise da linguagem e dos padrões sentimentais associados a cada tipo de discurso ofensivo. Vale ressaltar que os *tweets* da rede X foram coletados na língua inglesa e têm sua origem pré-processada e balanceada no artigo de Wang e Lu (2020).

Figura 2 – Algumas informações sobre o *dataset*.



Fonte: Kaggle, 2020

3.2 Google Colab e bibliotecas Python

O *Google Colab* é plataforma de computação em nuvem, disponibilizada pela Google, que oferece um ambiente de desenvolvimento gratuito para a execução de *notebooks Jupyter* diretamente no navegador. Ademais, facilita o armazenamento e compartilhamento de arquivos e modelos treinados, além de suportar bibliotecas populares de ciência de dados e *machine learning*, como Pandas, NumPy, Matplotlib, Scikit-learn e PyTorch, sem a necessidade de instalação manual, já que os pacotes vêm pré-configurados no ambiente (Pedregosa *et al.*, 2011).

Ele permite a colaboração em tempo real e oferece acesso gratuito a GPUs (Unidades de Processamento Gráfico) e TPUs (Unidades de Processamento Tensorial). Esses recursos de hardware especializados são essenciais para a computação paralela e são altamente eficazes em tarefas que envolvem grandes volumes de dados, como o treinamento de modelos de aprendizado profundo (Google, 2024).

Em conjunto com o Google Colab, este projeto também se beneficiou do uso de várias bibliotecas Python, cada uma contribuindo com funções e recursos específicos para o desenvolvimento e implementação da pesquisa.

3.3 Pré-processamento dos dados

A análise de sentimentos, como um método computacional para identificar e categorizar opiniões subjetivas, depende fortemente da qualidade dos dados de entrada. No entanto, os dados textuais brutos são ruidosos, contendo uma mistura de elementos linguísticos e não linguísticos que podem obscurecer as características associadas ao sentimento. Esta etapa do estudo aborda esse desafio por meio de um *pipeline* de pré-processamento projetado para transformar o texto em um formato estruturado adequado para aprendizado de máquina.

As técnicas de limpeza dos dados modificam a base original com o objetivo de padronizar o texto e eliminar elementos como palavras irrelevantes, pontuações e termos sem valor semântico para a classificação (Silva, 2016). Foram definidos três estágios hierárquicos para eliminar diferentes tipos de ruído no texto. Os resultados obtidos estão apresentados na Tabela 1.

Tabela 1 – Resultado das etapas de pré-processamento

Texto original	Sentimento	Texto limpo
RT @smoovfellow The only thing a woman should help a man build is a relationship #tbh #nosexist	gender	rt thing woman help man build relationship honest nosexist
are u referring to islamic terrorism? Ur definition and understanding of islam is what the media has taught u which frankly makes ur arguments worthless and not worthy of engaging with	religion	referring islamic terrorism definition understanding islam medium taught frankly make argument worthless worthy engaging
This is it this is the moment the perfect description of a girl I knew in high school no one bullied her but she was generally a weirdo who was very very determined to become Korean in any and every way possible even with the weird cringey shit she	age	moment perfect description girl knew high school one bullied generally weirdo determined become korean every way possible even weird cringey shit crazy obsessed annoying

was crazy obsessed and annoying	

Fonte: A autora.

O primeiro estágio, chamado de Normalização Estrutural, envolveu a remoção de *emojis*, a expansão de contrações e a filtragem de textos não escritos em inglês, visando padronizar o texto para a análise lexical.

No segundo estágio, a Padronização Lexical, URLs (*Uniform Resource Locator*), menções e pontuações foram removidas com expressões regulares, enquanto *stopwords* foram eliminadas utilizando o NLTK (*Natural Language Toolki*). As *hashtags* foram tratadas conforme sua posição no texto e a lematização ajudou a reduzir a variação das palavras.

Por fim, o terceiro estágio, chamado de Preservação Semântica, cuidou do tratamento de negações, remoção de textos duplicados e eliminação de frases muito curtas, garantindo um mínimo de contexto para a análise. O *pipeline* ainda gera textos delimitados por espaços para manter compatibilidade com modelos de Aprendizado de Máquina tradicionais, equilibrando precisão e flexibilidade.

3.4 Preparação para modelagem e representação vetorial dos textos

Após o pré-processamento e a limpeza dos dados, deu-se início à etapa de preparação para a modelagem. Inicialmente, os rótulos da coluna *sentiment*, originalmente categóricos, foram convertidos em valores numéricos por meio da codificação ordinal, etapa necessária para a aplicação dos algoritmos de Aprendizado de Máquina. Posteriormente, realizou-se a separação entre as variáveis independentes (X), correspondentes aos textos pré-processados, e a variável alvo (y), representada pelos sentimentos codificados.

A divisão dos dados em conjuntos de treinamento (80%) e teste (20%) foi efetuada utilizando estratificação, a fim de preservar a distribuição das classes. A análise do conjunto de treinamento pode revelar um desbalanceamento entre as categorias da variável alvo, o que pode comprometer o desempenho dos modelos. Para mitigar esse problema, a técnica de *oversampling* é utilizada, aumentando-se artificialmente a quantidade de exemplos das classes minoritárias, de modo a equilibrar o conjunto de dados e favorecer a capacidade preditiva dos algoritmos.

Para viabilizar a aplicação do modelo, é necessário converter os textos em representações numéricas interpretáveis pelo algoritmo. Para isso, utilizou-se a técnica *Bag of Words*, que gera uma matriz vetorial na qual cada dimensão corresponde a uma palavra do vocabulário e seu valor indica a frequência dessa palavra em um documento. Em seguida, a transformação TF-IDF ajustou os valores da matriz considerando a frequência dos termos em todos os documentos, atribuindo menor peso às palavras comuns e maior relevância às palavras mais raras e informativas.

3.5 Implementação do modelo LSTM

Nessa etapa, foi implementado um modelo de Rede Neural Recorrente do tipo LSTM aprimorado com um mecanismo de atenção, desenvolvido em PyTorch. Essa arquitetura é eficaz para tarefas de PLN devido à sua capacidade de capturar dependências de longo alcance em sequências e focar em informações relevantes. As LSTMs são capazes de "lembrar" informações importantes ao longo de uma frase longa, o que é fundamental para entender o contexto de um *tweet*. Para aprimorar ainda mais essa capacidade, incorporamos um mecanismo de atenção ao modelo (Olaniyan, 2023).

A priori, foi necessário um processo de preparação. Os *tweets* foram divididos em suas palavras constituintes, um processo conhecido como tokenização, no qual cada palavra recebe um código numérico único. Em seguida, essas sequências de códigos foram padronizadas para um mesmo tamanho, preenchendo com zeros ou cortando o excesso, garantindo que todas as entradas para o modelo tivessem um formato uniforme.

Um passo crucial foi a transformação dessas palavras em representações numéricas que capturassem seus significados e relações. Para isso, utilizamos *Word Embeddings* gerados pelo modelo Word2Vec. Nessa etapa, cada palavra é convertida em um ponto em um espaço multidimensional, onde palavras com significados semelhantes ficam próximas. Isso permite que o modelo compreenda o contexto e a semântica das palavras, o que é vital para identificar nuances em mensagens de *cyberbullying*.

O mecanismo de atenção permite que a rede neural "preste mais atenção" às palavras mais relevantes em uma frase ao tomar uma decisão de classificação. Por exemplo, em um tweet, algumas palavras podem ser mais indicativas de *cyberbullying* do que outras. A atenção ajuda o modelo a focar nessas palavras-chave, atribuindo-lhes maior peso, o que melhora significativamente a precisão da classificação (Olaniyan, 2023). O modelo foi construído com camadas que processam as representações das palavras, uma camada LSTM para analisar a sequência e a camada de atenção para ponderar a importância das palavras, culminando em uma camada final que decide a categoria de *cyberbullying*.

Para garantir que o modelo aprendesse de forma eficaz e fosse capaz de generalizar para novos *tweets*, o conjunto de dados foi dividido em partes para treinamento, validação e teste. O conjunto de treinamento foi usado para ensinar o modelo, enquanto o conjunto de validação foi utilizado para ajustar seus parâmetros e monitorar seu desempenho em dados não vistos durante o aprendizado. Isso é crucial para evitar o *overfitting*, um problema onde o modelo se torna muito bom em classificar os dados que já viu, mas falha em novos dados.

Para lidar com o desbalanceamento natural das categorias de *cyberbullying* (alguns tipos são mais frequentes que outros), aplicamos uma técnica de balanceamento de classes no conjunto de treinamento. Isso assegura que o modelo não ignore as categorias menos comuns. O treinamento foi otimizado com um processo de parada antecipada, que interrompe o aprendizado se o desempenho do modelo no conjunto de validação parar de melhorar, salvando a melhor versão do modelo. Essa metodologia garante que o modelo seja robusto e eficaz na identificação de *cyberbullying* em tempo real.

4 Resultados e discussão

Após a limpeza dos textos, foram removidos os *tweets* duplicados da coluna *text_clean*. A análise revelou que a classe *other_cyberbullying* apresentou uma redução significativa no número de amostras após a etapa de limpeza, passando de 7.823 para 4.902 *tweets* (-37,3%), assim como mostra a Tabela 2.

Tabela 2 – Comparação resultados pré-processamento

Sentimento	Antes da Limpeza	Depois da Limpeza	Redução	% Redução
Religião	7.998	7.916	-82	-1,0%
Idade	7.992	7.814	-178	-2,2%
Etnia	7.961	7.423	-538	-6,8%
Gênero	7.973	7.283	-690	-8,7%
Não *Cyberbullying*	7.945	6.064	-1.881	-23,7%
Outro *Cyberbullying*	7.823	4.902	-2.921	-37,3%

Fonte: A autora.

Esse decréscimo tornou a classe desproporcionalmente menor em comparação com as demais categorias, como “religião”, que manteve 99% de suas amostras, e o desequilíbrio pode comprometer o desempenho do modelo, favorecendo as classes majoritárias. Nesse ínterim, por se tratar de uma categoria genérica que engloba casos mal definidos, sua inclusão pode introduzir ruído no treinamento, dificultando a identificação de padrões consistentes. Diante disso, a remoção dessa classe se apresenta como a melhor estratégia.

Com as classes todas balanceadas, foi dado início ao treinamento do modelo de Rede Neural. O desempenho do modelo foi avaliado com base em métricas amplamente utilizadas em tarefas de classificação multiclasse, incluindo acurácia, precisão (*precision*), recuperação (*recall*) e F1-*score*.

É importante destacar que, previamente foi feito um treinamento e teste dessa mesma base de dados com o modelo classificador Naive Bayes, e obteve-se os resultados apresentados na Figura 3. As médias gerais (macro e ponderada) ficaram entre 0,86 e 0,88, evidenciando uma consistência e eficácia do modelo na tarefa de classificação multiclasse. Entretanto, foi notado que havia a possibilidade de melhorias.

Figura 3 – Relatório de classificação conjunto de teste Naive Bayes.

	precision	recall	f1-score	support
religion	0.87	0.97	0.92	1568
age	0.89	0.95	0.92	1552
ethnicity	0.93	0.93	0.93	1469
gender	0.91	0.84	0.87	1445
not bullying	0.74	0.63	0.68	1215
accuracy			0.88	7249
macro avg	0.87	0.86	0.86	7249
weighted avg	0.87	0.88	0.87	7249

Fonte: A autora.

Após essa etapa, foi realizado um treinamento e teste, com a mesma base de dados, utilizando o modelo de Rede Neural LSTM com mecanismos de atenção, sendo o relatório de classificação obtido no conjunto de teste apresentado na Figura 4.

Este modelo evidenciou um desempenho geral consistente, alcançando uma acurácia de 93% no conjunto de dados de teste. As métricas de F1-*score* — que combinam precisão e *recall* para oferecer uma visão equilibrada da performance — foram especialmente elevadas para as classes *'age'* (0,98), *'ethnicity'* (0,98) e *'religion'* (0,95), evidenciando a elevada eficácia do modelo na detecção desses tipos específicos de *cyberbullying*. A classe *'gender'* também apresentou um bom desempenho, com um F1-*score* de 0,91.

Figura 4 – Relatório de classificação conjunto teste LSTM com *Attention*.

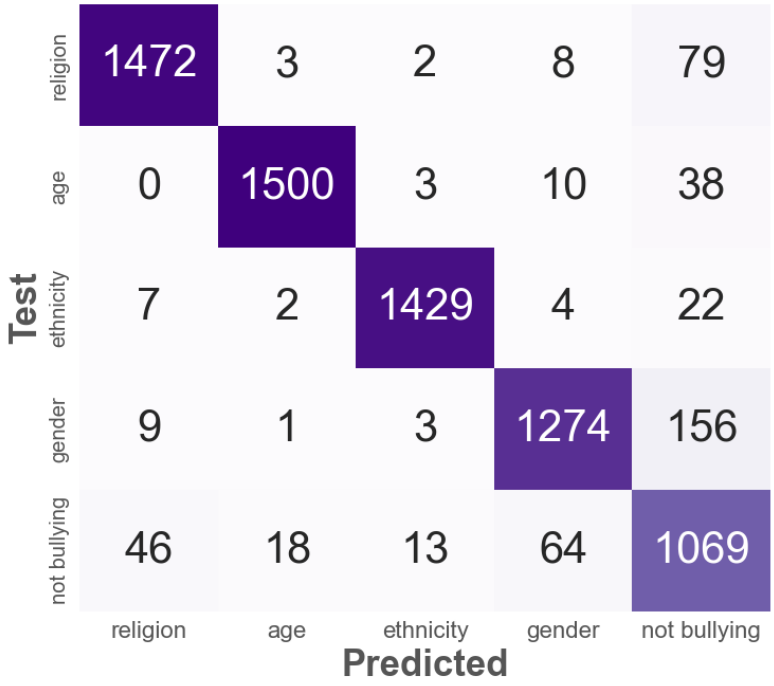
	precision	recall	f1-score	support
religion	0.96	0.94	0.95	1564
age	0.98	0.97	0.98	1551
ethnicity	0.99	0.98	0.98	1464
gender	0.94	0.88	0.91	1443
not bullying	0.78	0.88	0.83	1210
accuracy			0.93	7232
macro avg	0.93	0.93	0.93	7232
weighted avg	0.94	0.93	0.93	7232

Fonte: A autora.

Por outro lado, a classe *'not bullying'* obteve um F1-*score* inferior, de 0,83, em comparação com as categorias ofensivas. Embora esse resultado ainda represente uma boa performance, a precisão de 0,78 sugere que o modelo teve maior propensão a confundir mensagens não ofensivas com conteúdo de *cyberbullying*, ou vice-versa. Ainda assim, o *recall* de 0,88 indica que a maioria das instâncias corretamente rotuladas como não ofensivas foi identificada, provando uma boa sensibilidade do modelo a essa categoria.

A análise da Matriz de Confusão, exposta na Figura 5, reforça as conclusões extraídas do relatório de classificação.

Figura 5 – Matriz de Confusão.



Fonte: A autora.

Observa-se que a diagonal principal — que representam as classificações corretas — apresenta valores expressivos para todas as classes, especialmente para *'age'*, *'ethnicity'* e *'religion'*, indicando uma taxa de acerto elevada. A classe *'ethnicity'*, por exemplo, mostra um desempenho quase perfeito, com número mínimo de erros de classificação. No caso de *'not bullying'*, apesar do desempenho satisfatório, a matriz revela a natureza dos erros cometidos.

Algumas instâncias dessa classe foram incorretamente rotuladas como categorias ofensivas, o que pode indicar que o modelo encontra maior dificuldade em distinguir entre mensagens realmente neutras e aquelas que contêm nuances ou ambiguidades que se aproximam do discurso ofensivo. Essa observação está alinhada com o F1-*score* mais baixo observado para essa classe, e destaca a complexidade da tarefa de identificar corretamente conteúdos que tangem entre o aceitável e o ofensivo nas redes sociais.

4.1 Implicações para Sistemas de Detecção de *Cyberbullying*

A acurácia geral mostra que aplicação de Redes Neurais LSTM com mecanismos de atenção mostrou um avanço significativo na detecção de *cyberbullying*. A capacidade de capturar dependências de longo alcance em sequências textuais, combinada com a habilidade de compreensão mais focada nas nuances dos discursos online, pode ser uma ferramenta eficaz para a detecção automatizada em plataformas de redes sociais. No entanto, os resultados também apontam para a necessidade de abordagens mais refinadas em cenários de aplicação real.

Essas observações têm implicações cruciais para o desenvolvimento de sistemas de moderação de conteúdo. Um sistema baseado neste modelo, embora eficiente para formas explícitas de *cyberbullying*, ainda exigiria um refinamento para minimizar falsos positivos na classificação de conteúdo não ofensivo. Para otimizar a eficácia e a aceitação de tais sistemas, pode-se pensar em uma abordagem híbrida, onde a Inteligência Artificial complementa e aprimora a supervisão humana:

1. *Tweets* classificados com alta confiança como *cyberbullying* relacionado a etnia, religião ou idade poderiam ser automaticamente sinalizados ou para remoção ou moderação imediata. A alta acurácia e F1-scores nessas categorias justificam um nível maior de automação, reduzindo a carga de trabalho para moderadores humanos e permitindo uma resposta rápida a incidentes graves.

2. Casos menos claros, especialmente aqueles relacionados a gênero ou classificados como não-*bullying* com baixa confiança, poderiam ser encaminhados para revisão humana. Isso inclui instâncias onde a ambiguidade linguística é alta ou onde o contexto social é crucial para uma interpretação correta. A intervenção humana é vital para resolver essas incertezas e evitar a censura indevida.

3. O *feedback* dos usuários humanos poderia ser incorporado para refinar continuamente o modelo por meio de técnicas de aprendizado ativo, no qual o modelo aprende com seus erros e se adapta a novas nuances do *cyberbullying*. A incorporação de dados rotulados por humanos, pode melhorar a capacidade do modelo de distinguir entre conteúdo ofensivo e não ofensivo, e de identificar novas formas de assédio que emergem nas plataformas.

4.2 Limitações do Modelo e da Abordagem

Os resultados obtidos devem ser interpretados considerando algumas limitações inerentes ao modelo e à abordagem adotada. Em primeiro lugar, embora redes neurais recorrentes do tipo LSTM sejam capazes de capturar dependências temporais e sequenciais entre palavras, ainda podem apresentar dificuldades em lidar com contextos muito longos, onde a informação relevante se encontra distante da palavra analisada. Mesmo com a utilização do mecanismo de atenção, que melhora a capacidade de focar em partes relevantes da sequência, existe a possibilidade de que nuances sutis de significado, como sarcasmo ou ironia, não sejam corretamente identificadas.

Além disso, a representação textual utilizada depende de *embeddings* previamente treinados, que podem carregar vieses ou limitações do corpus em que foram baseados. Isso impacta diretamente na capacidade do modelo de generalizar

para novos contextos, principalmente quando o vocabulário ou o estilo de escrita diverge daquele encontrado no conjunto de treinamento.

Deve-se considerar que, apesar de o modelo ter apresentado bom desempenho em algumas classes, ainda há evidências de maior dificuldade na distinção entre mensagens que não configuram *cyberbullying*. Isso pode estar relacionado à subjetividade dessa categoria, à presença de linguagem ambígua ou à proximidade léxica com mensagens de ataque.

Por fim, mensagens ofensivas em redes sociais frequentemente utilizam gírias, referências culturais, ironia e construções não convencionais, que representam um desafio mesmo para arquiteturas avançadas. Portanto, ainda que o modelo baseado em LSTM com *Attention* apresente ganhos expressivos em relação a abordagens probabilísticas, ele não elimina integralmente as limitações impostas pela complexidade da linguagem natural.

4.3 Procedimentos Futuros e Possíveis Melhorias

Apesar dos resultados satisfatórios obtidos, há espaço para avanços que podem contribuir para o aprimoramento da abordagem. Uma possibilidade é a utilização de arquiteturas mais modernas de *transformers*, como BERT ou RoBERTa, que têm mostrado desempenho superior em tarefas de classificação de texto por capturarem relações contextuais de maneira mais robusta do que modelos recorrentes. Outra melhoria diz respeito ao pré-processamento dos dados. Estratégias mais sofisticadas, como a normalização de gírias, o reconhecimento de *emojis* e a detecção de sarcasmo, poderiam enriquecer a representação textual e reduzir ambiguidades. Do mesmo modo, a incorporação de *embeddings* contextuais, em substituição aos *embeddings* estáticos, poderia tornar o modelo mais sensível às variações de significado conforme o uso da palavra em diferentes contextos. Ademais, o treinamento em bases de dados mais amplas e diversas poderia aumentar a capacidade de generalização do modelo, reduzindo vieses e melhorando a identificação de diferentes formas de *cyberbullying*. Por fim, futuros trabalhos podem explorar modelos multimodais, que integrem não apenas o texto, mas também informações contextuais de imagens, metadados de usuários ou redes de interação social, uma vez que o fenômeno do *cyberbullying* frequentemente se manifesta em múltiplas dimensões na internet.

5 Conclusão

Esta pesquisa teve como objetivo principal desenvolver e avaliar um modelo de classificação de *cyberbullying* em *tweets*, utilizando técnicas de Processamento de Linguagem Natural e Aprendizado de Máquina. Ao longo do trabalho, foram exploradas diversas etapas, desde o pré-processamento dos dados textuais e a representação dos atributos, até a implementação e a avaliação de um modelo de Rede Neural Recorrente do tipo LSTM aprimorado com um mecanismo de atenção. Os resultados obtidos mostraram a viabilidade da detecção automática de *bullying*, evidenciando a capacidade do modelo em identificar diferentes categorias de comportamento ofensivo online.

As principais contribuições deste estudo incluem a aplicação prática de conceitos teóricos de *Machine Learning* em um problema socialmente relevante, a

análise detalhada do desempenho do modelo LSTM em um *dataset* real de *tweets*, e a identificação de desafios específicos relacionados à ambiguidade e evolução da linguagem. A análise de métricas como precisão, *recall* e F1-*score*, juntamente com a interpretação da matriz de confusão, forneceu uma série de entendimentos valiosos sobre os pontos fortes e as limitações do modelo.

Outro aspecto foi a oportunidade de refletir sobre como fenômenos sociais podem ser traduzidos em problemas computacionais e tratados a partir de técnicas matemáticas e estatísticas. O *cyberbullying*, sendo uma manifestação de violência digital, é uma questão que ultrapassa o campo tecnológico e atinge dimensões sociais, psicológicas e educacionais. Ao propor soluções computacionais para esse tipo de problema, este trabalho também reforça a necessidade de diálogo interdisciplinar, em que a tecnologia atua como aliada em estratégias de prevenção, conscientização e mitigação dos efeitos do discurso de ódio e da violência online.

Do ponto de vista acadêmico, este estudo contribui para o avanço da pesquisa em PLN aplicada à língua inglesa e ao contexto das redes sociais, apresentando resultados que podem servir como base comparativa para trabalhos futuros que explorem outros modelos ou técnicas de representação textual. Além disso, a utilização de dados reais provenientes de interações sociais online fortalece a relevância prática da pesquisa, aproximando-a das condições encontradas em aplicações reais de monitoramento e moderação de conteúdo.

Em síntese, este trabalho alcançou seus objetivos ao mostrar, por meio do modelo de Redes Neurais, o potencial da aplicação de técnicas de PLN e AM na detecção de discurso de ódio em redes sociais. A pesquisa evidencia a importância de avanços contínuos nesse campo, não apenas no âmbito tecnológico, mas também pelo impacto social que tais ferramentas podem exercer na construção de ambientes digitais mais saudáveis, inclusivos e seguros. Assim, mais do que apresentar resultados numéricos, é reforçada a ideia de que a ciência de dados pode ser uma ferramenta de transformação social, contribuindo para um uso mais ético e responsável das tecnologias digitais.

Referências

ALPAYDIN, E. Introduction to Machine Learning. 2nd. ed. The MIT Press, 2010. ISBN 026201243X.

ALY, Mohamed. Survey on multiclass classification methods. 2005.

ASLAM, S. Twitter by the Numbers: Stats, Demographics & Fun Facts. 2020. Omnicore. Disponível em: https://www.omnicoreagency.com/twitter-statistics/. Acesso em: 22 mar. 2025.

BACKLINKO. Twitter Users. 2024. Disponível em: https://backlinko.com/twitter-users. Acesso em: 03 fev. 2025.

BAHDANAU, D.; CHO, K.; BENGIO, Y. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv preprint arXiv:1409.0473, 2014. Disponível em: https://arxiv.org/abs/1409.0473. Acesso em: 05 jul. 2025.

BIRD, Steven; KLEIN, Ewan; LOPER, Edward. Natural Language Processing with Python. 1st ed. Sebastopol: O'Reilly Media, 2009. Disponível em: https://www.nltk.org/book/. Acesso em: 11 jul. 2025.

BISHOP, C. M. Pattern Recognition and Machine Learning. New York, NY, USA: Springer Science+Business Media, 2006.

GOOGLE. Ask a Techspert: What's the difference between a CPU, GPU and TPU? 30 out. 2024. Disponível em: https://blog.google/technology/ai/difference-cpu-gpu-tpu-trillium/. Acesso em: 30 abr. 2025.

HAYKIN, S. Neural Networks and Learning Machines. Pearson, 2009.

HINDUJA, S.; PATCHIN, J. W. Cyberbullying: An Update and Synthesis of the Research. In: Cyberbullying Research Center, 2019.

HOCHREITER, S.; SCHMIDHUBER, J. Long Short-Term Memory. Neural Computation, v. 9, n. 8, p. 1735–1780, 1997. Disponível em: https://doi.org/10.1162/neco.1997.9.8.1735. Acesso em: 15 ago. 2025.

INDURKHYA, Nitin; DAMERAU, Fred J. (Eds.). Handbook of Natural Language Processing. 2. ed. Boca Raton: CRC Press, 2010.

JURAFSKY, D.; MARTIN, J. H. Speech and Language Processing. 3. ed. Pearson, 2019.

LIU, B. Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers. 2012.

LIPTON, Z. C. Recurrent Neural Networks (RNNs): Architectures, Training Tricks, and Applications. 2023. Disponível em: https://www.ncbi.nlm.nih.gov/books/NBK597502/. Acesso em: 20 ago. 2025.

LUGER, G. F.; STUBBLEFIELD, W. A. Artificial Intelligence: Structures and Strategies for Complex Problem Solving. 3rd. ed. USA: Addison-Wesley Longman Publishing Co., Inc., 1997. ISBN 0805311963.

MANNING, C. D.; SCHÜTZE, H. Foundations of Statistical Natural Language Processing. Cambridge: MIT Press. 1999.

MCCARTHY, J. Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. 1956.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient Estimation of Word Representations in Vector Space. 2013. Disponível em: https://arxiv.org/abs/1301.3781. Acesso em: 20 ago. 2025.

OLAH, C. Understanding LSTM Networks. 2015. Disponível em: https://colah.github.io/posts/2015-08-Understanding-LSTMs/. Acesso em: 07 abr. 2025.

OLANIYAN, D. et al. Utilizing an Attention-Based LSTM Model for Detecting Sarcasm and Irony in Social Media Text. 2023. Disponível em: https://www.mdpi.com/2073-431X/12/11/231. Acesso em: 20 ago. 2025.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, v. 12, p. 2825–2830, nov. 2011.

PROVOST, Foster; FAWCETT, Tom. Data Science for Business: What You Need to Know About Data Mining and Data-analytic Thinking. O'Reilly Media, Inc., 1st edição, 2013. ISBN 1449361323, 9781449361327.

RIBEIRO, L. B. Análise de sentimento em comentários sobre aplicativos para dispositivos móveis: estudo do impacto do pré-processamento. 83 f. Monografia (Bacharelado em Ciência da Computação) — Universidade de Brasília, Brasília, DF, 2015.

RUSSELL, S.; NORVIG, P. Artificial Intelligence: A Modern Approach. 3rd. ed. USA: Prentice Hall Press, 2009. ISBN 0136042597.

SHALEV-SHWARTZ, S.; BEN-DAVID, S. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press. 2014.

SILVA, N. F. F. da. Análise de sentimentos em textos curtos provenientes de redes sociais. 112 p. Tese (Doutorado) — Universidade de São Paulo, São Carlos, SP, Brasil, 2016.

SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. Information Processing & Management, v. 45, n. 4, p. 427–437, 2009.

STATISTA. Global digital population as of February 2025. 2025. Disponível em: https://www.statista.com/statistics/617136/digital-population-worldwide/. Acesso em: 12 jul. 2025.

TSYTSARAU, M.; PALPANAS, T. Survey on mining subjective data on the web. Data Mining and Knowledge Discovery, 24(3), 478–514. Kluwer Academic Publishers. 2012.
VASWANI, A. et al. Attention is All You Need. In: Advances in Neural Information Processing Systems (NeurIPS). 2017. Disponível em: https://arxiv.org/abs/1706.03762. Acesso em: 14 jun. 2025.

VÁZQUEZ, F. Deep Learning made easy with Deep Cognition. 2017. Disponível em: https://becominghuman.ai/deep-learningmade-easy-with-deep-cognition-403fbe445351. Acesso em: 17 abr. 2025.

WANG, J.; FU, K.; LU, C. T. SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection. In: Proceedings of the 2020 IEEE International Conference on Big Data (IEEE BigData 2020), December 10-13, 2020.