Uso de aprendizado de maquina em profundidade na detecção de mensagens perigosas ou ofensivas em comunidades ou redes socias

Autor: Rômulo Júnio Vieira Rocha, Alessandro D'Angelo
Orientador: Rodrigo César Pedrosa Silva

19 de novembro de 2023

1 Problema de pesquisa

Qual é a eficácia de técnicas de aprendizado de maquina na detecção de mensagens perigosas ou ofensivas em comunidades ou redes sociais?

Com o rápido crescimento das comunidades online e o aumento da interação nas redes sociais, a detecção de mensagens perigosas ou ofensivas tornou-se uma questão crítica,a disseminação dessas mensagens pode levar a danos emocionais, ameaçar a segurança dos usuários e criar um ambiente hostil nas plataformas digitais. tal como mostrado por California et al. (2017). O alcance global das redes sociais permite que conteúdos ofensivos se espalhem rapidamente, afetando um grande número de pessoas em questão de minutos, ampliando os danos potenciais. Além disso, a disseminação de conteúdo prejudicial pode fomentar discursos de ódio e comportamentos negativos, criando uma espiral de violência e comportamento tóxico nas comunidades online. A incapacidade das técnicas tradicionais de detecção, como filtros baseados em palavras-chave, em lidar com a complexidade linguística e contextual das mensagens,revela as limitações em termos de precisão (TRAJANO, 2023), destacando-se a necessidade de investigar abordagens mais avançadas para combater esse problema. Portanto, é essencial investigar a eficácia de abordagens avançadas nesse contexto de acordo com Alshehri, Nagoudi e Abdul-Mageed (2020). Este estudo tem como objetivo avaliar a eficácia de técnicas avançadas, como processamento de linguagem natural, aprendizado de máquina e análise de sentimentos, na detecção de mensagens perigosas ou ofensivas em comunidades e redes sociais. Além disso, pretende-se examinar fatores que possam influenciar a detecção, como a diversidade linguística e a evolução das formas de comunicação online.

A pesquisa irá investigar métodos de pré-processamento, técnicas de classificação,modelagem de dados adequadas e uso de redes neurais para identificar com precisão conteúdos perigosos ou ofensivos. Além disso, serão analisadas a aplicabilidade e as limitações dessas abordagens em diferentes plataformas de redes sociais e comunidades online, considerando suas características específicas.

Espera-se que os resultados dessa pesquisa forneçam insights valiosos sobre a eficácia e os desafios enfrentados pelas abordagens na detecção de mensagens perigosas ou ofensivas em ambientes online. Esses insights podem contribuir para o desenvolvimento de técnicas mais sofisticadas e eficientes na proteção dos usuários contra conteúdos prejudiciais, promovendo uma experiência mais segura e saudável nas redes sociais e comunidades online.

2 Revisão Bibliográfica

O crescente volume de dados e a disseminação de discursos de ódio nas plataformas da Internet têm impulsionado esforços da comunidade científica para desenvolver abordagens automáticas de detecção. Essas abordagens vão desde técnicas tradicionais de processamento de linguagem natural até modelos baseados em aprendizado profundo, como redes neurais convolucionais e recorrentes. No entanto, desafios como a evolução constante dos discursos de ódio e a presença de vieses nos modelos ainda precisam ser superados. Pesquisas futuras exploram o uso de modelos de redes neurais e aprendizado de maquina, modelos de linguagem pré-treinados e abordagens multilíngues para aprimorar a detecção e promover um ambiente online mais seguro.

Dentre as diversas abordagens para a identificação de mensagens de ódio em redes sociais, a pesquisa conduzida por Coutinho e Malheiros (2020) se concentra na análise de sentimentos como uma estratégia eficaz para a detecção de mensagens homofóbicas no Twitter, especialmente no contexto da língua portuguesa. Os resultados obtidos demonstraram uma acurácia de 0,6148, uma precisão de 0,6667, uma sensibilidade de 0,6216 e uma medida-F (f-measure) de 0,6433. A abordagem de análise de sentimentos adotada pelos pesquisadores se mostrou promissora na identificação dessas mensagens prejudiciais. Seis algoritmos de aprendizagem de máquina foram utilizados para teste: Regressão Logística, Naive Bayes, Árvores de Decisão, Florestas Aleatórias e SVM. Apesar de terem alcançado valores máximos de precisão e sensibilidade de 0,91 e 0,90, respectivamente, quando observados apenas os discursos de ódio, esses valores caem para 0,44 e 0,61. Ao utilizar técnicas de aprendizado de máquina e treinamento com um conjunto de dados previamente disponível, foi possível obter resultados satisfatórios. Esses valores de acurácia, precisão, sensibilidade e medida-F indicam a eficácia do método proposto para a detecção de mensagens homofóbicas em um contexto específico, neste caso, a língua portuguesa no ambiente do Twitter.

Ademais, uma outra maneira utilizada para identificar algum tipo de linguajar ofensivo foi o treinamento de uma IA com auxílio de bag of words, que consiste em criar um vocabulário a partir de um conjunto de documentos, atribuir a cada palavra uma posição no vocabulário e representar cada documento como um vetor de frequência de palavras(CULOTTA; SORENSEN, 2004) e técnicas de aprendizado de máquina. Nessa abordagem, as palavras-chave são selecionadas com base em seu potencial ofensivo ou discriminatório. O modelo é treinado em um conjunto de dados contendo exemplos rotulados de mensagens ofensivas e não ofensivas.

A combinação dos métodos de aprendizado de máquina e palavras-chave resultou em resultados promissores na detecção de linguagem ofensiva. Em uma pesquisa conduzida por Paiva, Silva e Moura (2019), utilizando essa abordagem, foi alcançada uma acurácia de 0.81 ao utilizar o baf of words, em comparação com 0.73 sem o uso dessa técnica. Esses resultados demonstram a eficácia da combinação dessas abordagens para identificar e classificar com precisão mensagens ofensivas em textos e contribuem para o desenvolvimento de sistemas mais robustos de detecção de linguagem ofensiva em ambientes online.

Paralelamente, outra abordagem apresentada no estudo de modelos distribucionais para detecção de discursos de ódio em português, realizado por Silva (2021), envolveu o uso de SVM (Support Vector Machine) com N-Gram. O N-Gram é um modelo de língua que considera a ordem sequencial de N palavras ou caracteres. Os resultados dessa abordagem demonstraram uma precisão significativa na identificação de discursos de ódio, alcançando um valor de 0.8046.Essa abordagem combina o uso de SVM, uma técnica de aprendizado de máquina amplamente utilizada na classificação de textos, com o N-Gram para capturar a relação sequencial das palavras em um texto.

Os resultados obtidos em pesquisas recentes demonstram a efetividade dessas abordagens na identificação de mensagens ofensivas e homofóbicas. Os estudos de Coutinho e Malheiros (2020) e Paiva, Silva e Moura (2019) mostraram acurácias significativas na detecção de linguagem ofensiva, utilizando técnicas como análise de sentimentos e o uso de SVM com N-Gram.No entanto, ainda há desafios a serem superados, como a evolução dos discursos de ódio e a presença de vieses nos modelos. Pesquisas futuras estão explorando o uso de modelos de linguagem pré-treinados e abordagens multilíngues para aprimorar a detecção e promover um ambiente online mais seguro.

Em suma, as abordagens mencionadas na revisão bibliográfica apresentam resultados positivos na detecção de discursos de ódio e linguagem ofensiva em comunidades ou redes sociais. A contínua investigação e desenvolvimento nessa área são fundamentais para melhorar a capacidade de identificar e combater esses tipos de comportamentos prejudiciais, contribuindo para a construção de ambientes online mais inclusivos e respeitosos.

Entretanto, existem problemas com essas abordagens é a constante evolução dos discursos de ódio e da linguagem ofensiva. À medida que novas formas de expressão e novos termos surgem, os algoritmos e modelos existentes podem não ser capazes de detectar adequadamente essas novas manifestações. Isso pode levar a falsos negativos, onde mensagens prejudiciais passam despercebidas, ou a falsos positivos, onde mensagens inofensivas são erroneamente identificadas como perigosas ou ofensivas.Além disso, a presença de vieses nos modelos de detecção também é uma preocupação. Os modelos de aprendizado de máquina são treinados em conjuntos de dados existentes, que podem refletir vieses e preconceitos presentes na sociedade.

Outro desafio está relacionado à diversidade linguística e cultural das comunidades e redes sociais. As abordagens desenvolvidas em um determinado contexto linguístico podem não ser diretamente aplicáveis a outros idiomas ou culturas.O desenvolvimento de abordagens mais robustas e inclusivas requer a consideração dessas limitações e a busca por soluções que levem em conta a evolução dos discursos de ódio, a mitigação de vieses nos modelos, a diversidade linguística e cultural, bem como a disponibilidade de conjuntos de dados representativos.

3 Fundamentos

A detecção de mensagens perigosas ou ofensivas em comunidades ou redes sociais é um tema relevante na era digital, considerando o aumento da disseminação de conteúdo prejudicial. A utilização de algoritmos e técnicas de processamento de linguagem natural e aprendizado de máquina oferece uma abordagem promissora para lidar com esse desafio (BORIOLA; PAETZOLD, 2021). Esses algoritmos permitem identificar padrões, analisar o contexto e classificar automaticamente as mensagens, auxiliando na proteção dos usuários e na promoção de ambientes mais seguros e saudáveis online.

3.1 Linguagem Ofensiva

A linguagem ofensiva se refere a um tipo de comunicação que envolve insultos, ameaças, ofensas, palavras de baixo calão e obscenidade. É um comportamento que pode causar danos emocionais e impactar negativamente as interações nas comunidades e redes sociais (BORIOLA; PAETZOLD, 2021).

3.2 Processamento de Linguagem Natural (PLN)

Algoritmos e técnicas de Processamento de Linguagem Natural referem-se a métodos computacionais e abordagens utilizados para analisar, compreender e processar textos e linguagem humana. O PLN combina conhecimentos de linguística, inteligência artificial e ciência da computação para permitir que computadores compreendam e interajam com a linguagem humana de forma significativa (BRAGA, 2008).

3.3 Aprendizado de Máquina

O aprendizado de máquina é uma disciplina da inteligência artificial que busca desenvolver técnicas computacionais para que os sistemas adquiram conhecimento de forma automática. Esses sistemas são capazes de tomar decisões com base em experiências anteriores, utilizando algoritmos e métodos estatísticos. Ao analisar e processar dados, eles identificam padrões e aprendem com eles ao longo do tempo (BRUNIALTI et al., 2015).

3.4 Mineração de Texto

A mineração de texto desempenha um papel fundamental na detecção de linguagem ofensiva em comunidades e redes sociais. Por meio de algoritmos e técnicas de processamento de linguagem natural, a mineração de texto permite analisar e extrair informações relevantes dos textos, identificando padrões, contextos e sentimentos subjacentes às mensagens (SILVA; PAPA; COSTA, 2016). Essa abordagem possibilita a detecção automatizada de palavras, frases ou expressões ofensivas, contribuindo para a criação de ambientes online mais seguros e saudáveis. Ao utilizar métodos de mineração de texto, é possível identificar e classificar eficientemente conteúdos inapropriados, auxiliando na proteção dos usuários e no combate ao discurso de ódio, cyberbullying e outras formas de comportamento prejudicial.

3.5 Detecção com Processamento de Linguagem Natural e Aprendizado de Máquina

Com a aplicação de algoritmos e técnicas de Processamento de Linguagem Natural e aprendizado de máquina na detecção de mensagens perigosas ou ofensivas em comunidades ou redes sociais, através da análise automatizada de textos, é possível identificar padrões, contextos e sentimentos subjacentes às mensagens, permitindo uma detecção mais precisa e eficiente.

3.6 Abordagem Robusta com PLN e Aprendizado de Máquina

A combinação de algoritmos de Processamento de Linguagem Natural (PLN) e aprendizado de máquina, como PyTorch LSTM e BERT, proporciona uma abordagem ainda mais robusta e eficaz para lidar com a detecção de mensagens perigosas ou ofensivas em escala. O uso de PyTorch LSTM permite modelar dependências temporais em sequências de texto, capturando nuances e padrões complexos nas mensagens. Por outro lado, a incorporação de BERT, um modelo pré-treinado altamente avançado baseado em transformers, possibilita uma compreensão mais profunda das relações semânticas em contextos textuais complexos.

Essas técnicas avançadas oferecem a capacidade de identificar e filtrar conteúdos inapropriados em tempo real, contribuindo de maneira significativa para a criação de ambientes online mais seguros e saudáveis. A combinação desses algoritmos permite uma análise mais abrangente, considerando não apenas a estrutura sequencial das mensagens, mas também as relações

semânticas e contextuais, resultando em uma detecção mais precisa e eficiente de mensagens prejudiciais em comunidades e redes sociais. Essa abordagem integrada promove a promoção de um ambiente digital mais seguro e a construção de comunidades online mais saudáveis.

4 Metodologia

4.1 Abordagem Proposta

Neste capítulo, será apresentada a abordagem proposta para a detecção de mensagens perigosas em comunidades ou redes sociais. Serão descritos em detalhes a metodologia adotada, os recursos utilizados e as técnicas aplicadas na implementação do sistema.

4.1.1 Metodologia Adotada

A metodologia adotada para implementar a abordagem proposta segue um processo em etapas, visando a detecção eficiente de mensagens perigosas.

4.1.2 Pré-processamento

Na etapa de pré-processamento, as mensagens passam por uma série de etapas para prepará-las para a análise posterior. Inicialmente, é realizada a remoção de pontuação,emojis e links, com o intuito de eliminar caracteres especiais e símbolos que não contribuem para a classificação das mensagens. Em seguida, ocorre a tokenização, em que as mensagens são divididas em palavras ou tokens individuais. Posteriormente, é feita a remoção de stopwords, que são palavras comuns na língua portuguesa que não têm impacto significativo na classificação. Por fim, é aplicado o stemming, que reduz as palavras às suas raízes, a fim de normalizá-las. Essas etapas de pré-processamento visam reduzir a dimensionalidade do texto e padronizar o conteúdo para análise posterior.

4.1.3 Extração de Características

Na etapa de extração de características, são selecionados os atributos relevantes dos textos que serão utilizados para a classificação das mensagens. Diferentes técnicas são aplicadas para capturar informações semânticas, estruturais e estilísticas presentes nos textos. Entre as abordagens utilizadas estão o modelo bag-of-words, que representa as mensagens como vetores de frequência de palavras, e a utilização de n-grams, que considera sequências de palavras adjacentes. Além disso, são exploradas técnicas de word embeddings, que representam as palavras em um espaço vetorial de alta dimensão, capturando relações semânticas entre elas. Também são consideradas características estilísticas, como tamanho médio das palavras e quantidade de letras maiúsculas. Essas características enriquecem a representação das mensagens, permitindo uma classificação mais precisa.

4.1.4 Aprendizado de maquina

Ajuste Fino (Fine-Tuning):Ao inicio do desenvolvimento do aprendizado de maquina, com a finalidade de encontra o melhor ajuste para o desenvolvimento do detector de mensagens ofensivas, sera iniciado o treinamento com os parâmetros pré-treinados dos modelos escolhidos(principalmente para BERT) e ajustados finamente para a tarefa, de modo a evitar

overfitting ou problemas com generalizações. Ademais, serão utilizas técnicas de validação cruzada para avaliar o desempenho do modelo em diferentes conjuntos de dados e garantir essa generalização.

O monitoramento das métricas de desempenho (precisão, recall, F1-score) será realizado durante o treinamento para ajustar os hiperparâmetros conforme necessário.

4.2 Método Experimental

Neste capítulo, será apresentado o método experimental adotado para validar a abordagem proposta. Serão descritos em detalhes os materiais utilizados, os procedimentos de coleta de dados, os experimentos realizados e as métricas de avaliação utilizadas.

4.2.1 Descrição dos Materiais

Para a realização dos experimentos, foi utilizada a seguinte base de dados do keaggle, uma biblioteca de processamento de linguagem natural, um ambiente de desenvolvimento integrado (IDE) e um servidor para execução dos experimentos.Utilizamos tanto o PyTorch LSTM quanto o BERT, destacando a capacidade dessas ferramentas na análise de textos, como biblioteca de processamento de linguagem natural, enquanto o IDE adotado foi o Jupyter Notebook. O conjunto de dados foi obtido a partir do link :"https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification/data", contando no total 47000 tweets para treinamento e teste.

4.2.2 Procedimentos Experimentais

Os procedimentos experimentais foram conduzidos com o objetivo de avaliar a eficácia da abordagem proposta na detecção de mensagens perigosas. Para isso, o conjunto de dados foi dividido em conjunto de treinamento e conjunto de teste, seguindo uma proporção de 70 por cento para treinamento e 30 por cento para teste. Foi utilizada a técnica de validação cruzada, realizando múltiplas execuções do experimento(5 para cada metodo) e calculando a média das métricas de desempenho obtidas.

4.2.3 Avaliação e Métricas

A avaliação do desempenho da estratégia proposta foi conduzida através da utilização de métricas de avaliação em classificação, que incluem acurácia, precisão, recall e F1-score. Essas métricas desempenham papéis fundamentais na análise de como o modelo está se saindo em termos de classificação de mensagens perigosas e não perigosas.

A acurácia é uma medida que avalia a proporção de previsões corretas em relação ao total de previsões realizadas pelo modelo. Essa métrica oferece uma visão geral da precisão global do sistema em todas as classes.A precisão é uma métrica que foca na proporção de verdadeiros positivos (amostras corretamente classificadas como positivas) em relação a todas as amostras classificadas como positivas pelo modelo. Isso indica a capacidade do sistema de evitar classificações incorretas como positivas.O recall, também conhecido como sensibilidade ou taxa de verdadeiros positivos, mede a proporção de verdadeiros positivos em relação a todas as amostras que realmente são positivas. Isso avalia a capacidade do modelo de identificar corretamente todas as instâncias positivas, evitando falsos negativos.O F1-score é uma métrica que considera tanto a precisão quanto o recall, calculando a média harmônica entre essas duas métricas. Ele é especialmente útil quando o equilíbrio entre evitar falsos positivos e falsos negativos é importante.

Além disso, foram utilizadas curvas de aprendizado e matriz de confusão para uma análise mais aprofundada dos resultados obtidos.

4.2.4 Considerações Éticas

Durante a realização dos experimentos, foram consideradas as questões éticas relacionadas à utilização de dados sensíveis e ao impacto potencial na privacidade dos usuários. Todas as diretrizes éticas estabelecidas pela instituição foram seguidas, garantindo a confidencialidade e anonimato dos dados coletados. Também foram adotadas medidas para proteger a identidade dos usuários e garantir que as mensagens fossem utilizadas apenas para fins de pesquisa.

Referências

ALSHEHRI, Ali; NAGOUDI, El Moatez Billah; ABDUL-MAGEED, Muhammad. Understanding and detecting dangerous speech in social media. **arXiv.org**, mai. 2020. Disponível em: <https://arxiv.org/abs/2005.06608>.

BORIOLA, Marcos Aurélio Hermógenes; PAETZOLD, Gustavo Henrique. Detectando linguagem ofensiva em tweets utilizando modelos Transformer. Guarapuava/PR, 2021.

BRAGA, Daniela. Algoritmos de processamento da linguagem natural para sistemas de conversao texto-fala em portugués. portugués. Coruña, España, 2008.

BRUNIALTI, Lucas et al. Aprendizado de Máquina em Sistemas de Recomendação Baseados em Conteúdo Textual: Uma Revisão Sistemática. Goiânia, p. 203–210, 2015.

CALIFORNIA, Donghyeon Won University of et al. Protest activity detection and perceived violence estimation from social media images: Proceedings of the 25th ACM International Conference on Multimedia. **ACM Conferences**, out. 2017. Disponível em: <https://dl.acm.org/doi/abs/10.1145/3123266.3123282?casa_token=w_PYml8OKL0AAAAA%3ApVcDW_xfCP7ukrxtYxos-J9yD9lIWYbn4KwZu-2mUH9QWVf3-UbwNeF6Ie_COQ7nY__isnCFSdU_sg>.

COUTINHO, Vinicius Matheus de Medeiros Silva; MALHEIROS, Yuri. Detecção de Mensagens Homofóbicas em Português no Twitter usando Análise de Sentimentos. Cuiabá, p. 1–12, 2020.

CULOTTA, Aron; SORENSEN, Jeffrey. Dependency tree kernels for relation extraction. In: PROCEEDINGS of the 42nd Annual Meeting on Association for Computational Linguistics (ACL). [S.l.: s.n.], 2004.

PAIVA, Peter Dias; SILVA, Vanecy Matias da; MOURA, Raimundo Santos. Detecção automática de discurso de ódio em comentários online. In: SOCIEDADE BRASILEIRA DE COMPUTAÇÃO. ESCOLA Regional de Computação Aplicada à Saúde (ERCAS), 7th. Teresina: [s.n.], 2019. P. 157–162.

SILVA, Adriano dos Santos Rodrigues da. **Estudo de modelos distribucionais para detecção de discurso de ódio em português**. 2021. Tese (Doutorado) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo. Master's Dissertation. DOI: 10.11606/D.100.2021.tde-28012022-074813. Disponível em: <https://doi.org/10.11606/D.100.2021.tde-28012022-074813>.

SILVA, Luis Alexandre da; PAPA, João Paulo; COSTA, Kelton Augusto Pontara da. Unsupervised learning features for malicious content detection, 2016.

TRAJANO, Douglas de Oliveira. Detecção de linguagem tóxica aplicada a textos em português, 2023.