Uso de aprendizado de máquina em profundidade na detecção de mensagens perigosas ou ofensivas em comunidades ou redes sociais

Alessandro D'Angelo[1], **Rômulo Júnio Vieira Rocha**[2]

[1]Universidade Federal de Ouro Preto (UFOP)
Campus Morro do Cruzeiro – 35402-160 – Ouro Preto – MG – Brasil

[2]Departamento de Computação – UFOP
Minas Gerais, BR.

{alessandro.angelo@aluno.ufop.edu.br,romulo.rocha@aluno.ufop.edu.br

Abstract. *This study investigates the effectiveness of machine learning techniques in detecting dangerous or offensive messages in online communities and social networks. Given the escalating impact of such messages on emotional well-being, user safety, and the overall online environment, traditional detection methods like keyword-based filters prove insufficient. The research focuses on advanced approaches, including natural language processing, machine learning, and sentiment analysis, aiming to evaluate their accuracy in identifying harmful content. Factors such as linguistic diversity, evolving communication forms, and platform-specific characteristics will be analyzed. The findings are expected to offer valuable insights into the challenges and efficacy of these approaches, guiding the development of more sophisticated techniques for safeguarding users and promoting a safer online experience.*

Resumo. *Este estudo investiga a eficácia de técnicas de aprendizado de máquina na detecção de mensagens perigosas ou ofensivas em comunidades online e redes sociais. Dada a crescente influência dessas mensagens no bem-estar emocional, na segurança do usuário e no ambiente online como um todo, métodos tradicionais de detecção, como filtros baseados em palavras-chave, mostram-se insuficientes. A pesquisa concentra-se em abordagens avançadas, incluindo processamento de linguagem natural, aprendizado de máquina e análise de sentimentos, visando avaliar sua precisão na identificação de conteúdo prejudicial. Serão analisados fatores como diversidade linguística, formas de comunicação em constante evolução e características específicas de plataformas. Os resultados esperados oferecerão insights valiosos sobre os desafios e a eficácia dessas abordagens, orientando o desenvolvimento de técnicas mais sofisticadas para proteger os usuários e promover uma experiência online mais segura.*

1. Introdução

Com o crescimento exponencial das comunidades online e o aumento constante da interação nas redes sociais, a detecção de mensagens perigosas ou ofensivas emergiu como uma questão crítica. A propagação dessas mensagens potencialmente prejudiciais pode resultar em danos emocionais, ameaçar a segurança dos usuários e gerar um ambiente hostil nas plataformas digitais, conforme evidenciado por California et al. [2]. A natureza global das redes sociais amplifica a rapidez com que o conteúdo ofensivo se espalha, impactando um número significativo de pessoas em questão de minutos e aumentando os danos potenciais.

Além disso, a disseminação de conteúdo prejudicial contribui para a promoção de discursos de ódio e comportamentos negativos, criando uma espiral de violência e toxicidade nas comunidades online. A limitação das técnicas tradicionais de detecção, como filtros baseados em palavras-chave, em lidar com a complexidade linguística e contextual das mensagens ressalta suas limitações em termos de precisão [8]. Isso destaca a necessidade premente de explorar abordagens mais avançadas para abordar esse problema, e é nesse cenário que o aprendizado de máquina surge como uma ferramenta poderosa.

O aprendizado de máquina representa uma forma avançada de abordar a detecção de mensagens perigosas ou ofensivas, possibilitando a análise automática e inteligente de grandes volumes de dados textuais. Ao contrário das técnicas convencionais, que muitas vezes se mostram limitadas diante da sofisticação das expressões linguísticas, os modelos de aprendizado de máquina são capazes de aprender padrões complexos e nuances, adaptando-se dinamicamente às evoluções na linguagem online.

Portanto, torna-se essencial investigar a eficácia de abordagens mais sofisticadas, como o aprendizado de máquina, nesse contexto, conforme proposto por Alshehri, Nagoudi, and Abdul-Mageed [1]. Diante desse cenário, surge a inevitável indagação: qual é a eficácia das técnicas de aprendizado de máquina na detecção de mensagens perigosas ou ofensivas em comunidades ou redes sociais?

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 máquina, 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 and Malheiros [3] 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[4] 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, and Moura [6], 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 [7], 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 and Malheiros [3] e Paiva, Silva, and Moura [6] 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.[5]

Os estudos revisados evidenciam progressos significativos na identificação de discursos de ódio e linguagem ofensiva em plataformas online, notadamente em contextos específicos, como o Twitter em língua portuguesa. As abordagens abrangem desde técnicas tradicionais, como análise de sentimentos, até modelos avançados de aprendizado de máquina, como Regressão Logística, Naive Bayes, Árvores de Decisão, Florestas Aleatórias, SVM, e até mesmo o uso de modelos baseados em redes neurais como BERT. Contudo, desafios persistentes, como a evolução constante dos discursos de ódio, vieses nos modelos e a diversidade linguística e cultural, são claramente identificados. A necessidade de desenvolver abordagens mais robustas, capazes de lidar com a diversidade e a constante evolução desses fenômenos, é um tema recorrente. A utilização de modelos de linguagem pré-treinados e abordagens multilíngues é apontada como uma possível direção para melhorar a eficácia na detecção. Além disso, a consideração cuidadosa de vieses nos conjuntos de dados e a adaptação das abordagens para diferentes contextos linguísticos e culturais emergem como pontos cruciais para o aprimoramento desses sistemas. A pesquisa futura, portanto, deve focar no desenvolvimento de metodologias mais

inclusivas e na superação desses desafios identificados, integrando avanços da inteligência artificial com uma compreensão mais profunda das dinâmicas linguísticas e sociais online.

3. Metodologia

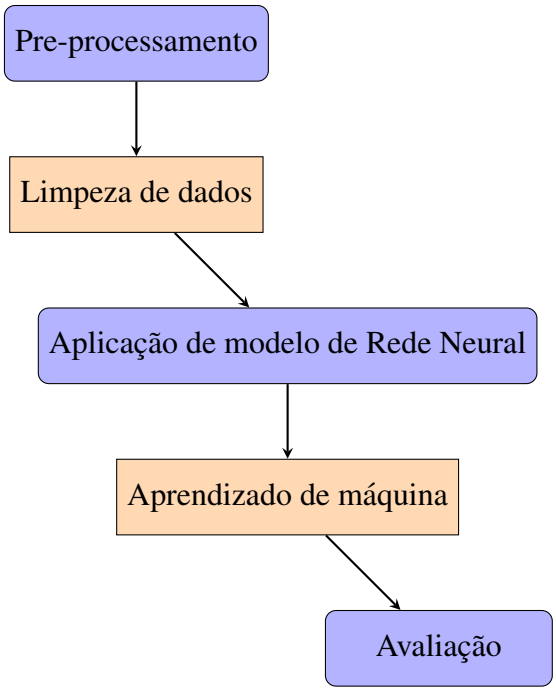
3.1. Abordagem Proposta

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. Todo o desenvolvimento do projeto esta disponivel em :

$$"https://github.com/Romz16/Detec-o_MensagensOfensivas.git"$$

3.1.1. Metodologia Adotada

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



3.1.2. Pré-processamento

Na etapa de pré-processamento, as mensagens passam por várias etapas para prepará-las para análise posterior. Inicialmente, realiza-se a remoção de pontuação, emojis e links, visando eliminar caracteres especiais e símbolos que não contribuem para a classificação das mensagens. Em seguida, ocorre a tokenização, na qual as mensagens são divididas em palavras ou tokens individuais. Posteriormente, é efetuada a remoção de stopwords, que são palavras comuns na língua portuguesa sem impacto significativo na classificação. Por fim, aplica-se o stemming para reduzir as palavras às suas raízes, normalizando-as. Essas etapas de pré-processamento têm como objetivo reduzir a dimensionalidade do texto e padronizar o conteúdo para análise posterior.

3.1.3. 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.

3.1.4. Aprendizado de maquina

Na metodologia do aprendizado de máquina adotamos uma estratégia abrangente para análise de dados textuais. Inicialmente, conduzimos uma etapa detalhada de limpeza, desenvolvendo funções personalizadas para remover ruídos, pontuações e stop words, além de incorporar técnicas como a utilização de Bag-of-Words (BoW) e N-grams para capturar informações relevantes na representação do texto. Como ponto de partida, implementamos um classificador baseline com o algoritmo Naive Bayes, proporcionando insights fundamentais sobre a natureza dos dados.

Para aprimorar a robustez do modelo, incorporamos técnicas como Dropout durante o treinamento das redes neurais recorrentes (RNN), visando evitar overfitting. Além disso, exploramos o uso de TF-IDF (Term Frequency-Inverse Document Frequency) para ponderação de termos, destacando a importância relativa das palavras nos documentos.

Nessa linha, também incluímos a utilização de unidades recorrentes de portas (GRU), uma variação de redes neurais recorrentes, que são especialmente eficazes na modelagem de sequências temporais, como texto. A inclusão de GRU oferece uma abordagem mais sofisticada para capturar dependências de longo prazo em dados sequenciais, permitindo uma melhor compreensão do contexto e, consequentemente, melhorando o desempenho do modelo na tarefa de classificação de textos.

Essa abordagem metodológica diversificada reflete nossa busca por uma análise abrangente e eficaz do processamento de linguagem natural, incorporando técnicas inovadoras para lidar com a complexidade inerente aos dados textuais.

3.2. 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.

3.2.1. Descrição dos Materiais

Para a realização dos experimentos, foi utilizada uma base de dados do keaggle, uma biblioteca de processamento de linguagem natural, um ambiente de desenvolvimento inte-

grado (IDE) e um servidor para execução dos experimentos.Utilizamos PyTorch LSTM, que oferece uma abordagem robusta para processamento de linguagem natural, o Nave Bayers que é reconhecido por sua eficiência em tarefas de classificação de texto e GRUs, por sua vez, destacam-se na modelagem de sequências temporais, como texto, capturando dependências de longo prazo de maneira eficaz. Enquanto o IDE adotado foi o Jupyter Notebook.

3.2.2. Conjunto de Dados Utilizados

Os dados empregados nos modelos para avaliação foram adquiridos através do Kaggle, acessíveis pelo link: "https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification/data". Este conjunto consiste em 47.000 tweets em língua inglesa, previamente classificados em diversas categorias, tais como Idade, Etnia, Gênero, Religião, outros tipos de ciberbullying e não ciberbullying. Entretanto, para esta pesquisa, realizou-se uma adaptação nas categorias, simplificando-as para dois grupos mais abrangentes: ciberbullying e não ciberbullying. Esta modificação visa fornecer uma análise mais focalizada nos aspectos fundamentais do fenômeno, facilitando a compreensão e interpretação dos resultados obtidos durante as análises realizadas nos modelos.

3.2.3. 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 método) e calculando a média das métricas de desempenho obtidas.

3.2.4. 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.

3.2.5. 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.

4. Resultados

4.1. Introdução

O presente estudo visa analisar e aprimorar algoritmos de detecção de bullying em mensagens utilizando o mesmo conjunto de dados para diferentes arquiteturas de redes neurais, como Bi-LSTM (Long Short-Term Memory), e modelos baseados em Naive Bayes e GRU (Gated Recurrent Unit). O desempenho dos modelos foi avaliado por meio de métricas como precisão, recall e F1-score, sendo essencial entender o impacto de ajustes como batch normalization, learning rate variável e variações na arquitetura dos modelos.

4.2. Resultados obtidos

Durante o desenvolvimento do projeto, foram criados alguns ”modelo 2”, que se referem a um modelo orginial, ao inves de pegar o modelo disponivel no projeto de referência e tentar adapta-lo.

O modelo original de Naive Bayes utiliza uma combinação de CountVectorizer e TfidfTransformer para processar dados textuais, seguido por um classificador Naive Bayes multinomial. Em contraste, o Naive Bayes Model 2 adota uma abordagem mais eficiente, empregando um pipeline com TfidfVectorizer e MultinomialNB, integrados em um único objeto. A introdução de parâmetros ajustáveis, otimizados por meio do GridSearchCV, possibilita uma exploração sistemática para encontrar os melhores hiperparâmetros. Isso resulta em um modelo mais flexível e adaptável, contribuindo potencialmente para melhorias no desempenho preditivo do classificador Naive Bayes.

O modelo original de LSTM utiliza uma arquitetura básica de LSTM para classificação de sentimentos, empregando uma camada de atenção para calcular um vetor de contexto ponderado e uma camada totalmente conectada para classificar o vetor de contexto em classes. Por outro lado, o modelo 2 aprimora a arquitetura adicionando dropout na camada LSTM para regularização e uma camada de normalização de lote (Batch Normalization) antes da camada totalmente conectada. Além disso, o modelo 2 incorpora uma inicialização de pesos Xavier Uniforme para a camada totalmente conectada e utiliza o otimizador Adam com uma taxa de aprendizado inicial de 0.001, bem como um agendador de taxa de aprendizado (scheduler) para ajustar dinamicamente a

taxa de aprendizado durante o treinamento. Essas melhorias visam aumentar a estabilidade do treinamento e melhorar o desempenho do modelo na tarefa de classificação de sentimentos

Os resultados obtidos nas métricas de avaliação para os diferentes modelos de Bi-LSTM, como mostrado em 1 revelam algumas tendências interessantes. Inicialmente, observamos que a introdução de Batch Normalization, por si só, não gerou melhorias significativas nos resultados em comparação com o modelo original. Embora tenha ocorrido uma ligeira redução na precisão para a classe ”Não Bullying”, houve um aumento correspondente no recall para a classe ”Bullying”. Por outro lado, a inclusão do Learning Rate Variável parece ter proporcionado um leve aumento na precisão para a classe ”Não Bullying”, embora tenha sido acompanhada por uma diminuição no recall para ambas as classes. Curiosamente, o Modelo 2, que incorpora tanto Batch Normalization quanto Learning Rate Variável, demonstra resultados semelhantes aos demais modelos em termos de precisão, recall e F1-Score, indicando que essas técnicas podem não ter impacto significativo nas métricas de avaliação para este conjunto de dados específico. No geral, todos os modelos apresentam uma acurácia consistente de 0.87, sugerindo que eles são igualmente eficazes na classificação de textos relacionados a bullying. legenda:

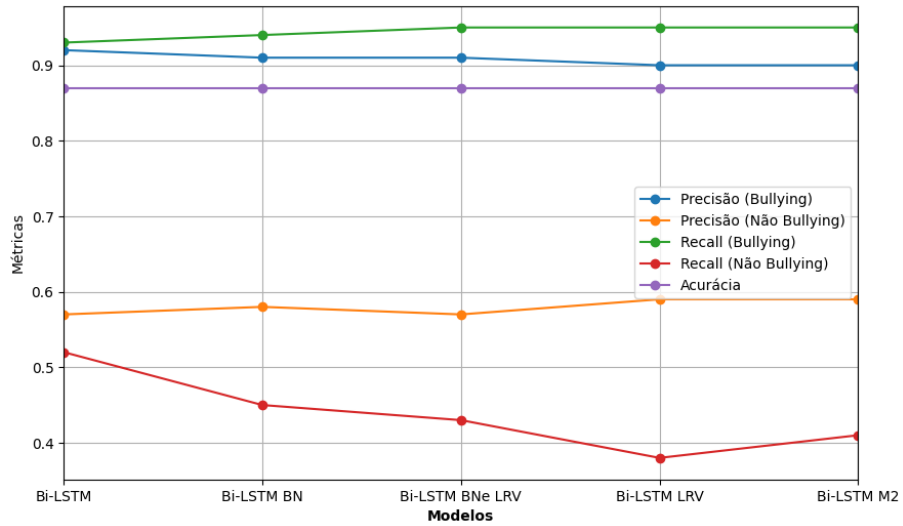


Figure 1. Figura 1- Resultados do Modelo LSTM

BN =batch normalization; LRV : Learning rate variavel. Os resultados obtidos para os modelos Naive Bayes apresentados em 2 demonstram algumas diferenças significativas entre o modelo original e o Modelo 2. No modelo original, observamos uma precisão de 0.93 para a classe ”Bullying” e 0.50 para a classe ”Não Bullying”, com um recall de 0.89 para ”Bullying” e 0.63 para ”Não Bullying”. O Modelo 2, por sua vez, mostra uma precisão ligeiramente maior para ambas as classes, com 0.92 para ”Bullying” e 0.58 para ”Não Bullying”. No entanto, há uma pequena diminuição no recall para ambas as classes, com 0.93 para ”Bullying” e 0.54 para ”Não Bullying”. No geral, o Modelo 2 alcança uma acurácia de 0.87, indicando uma melhoria marginal em relação ao modelo original. Embora haja uma discrepância notável na precisão entre as classes.

Os resultados sobre o modelo GRU, apresentados em 3 revelam variações significativas no desempenho de acordo com as técnicas empregadas. No modelo GRU padrão,

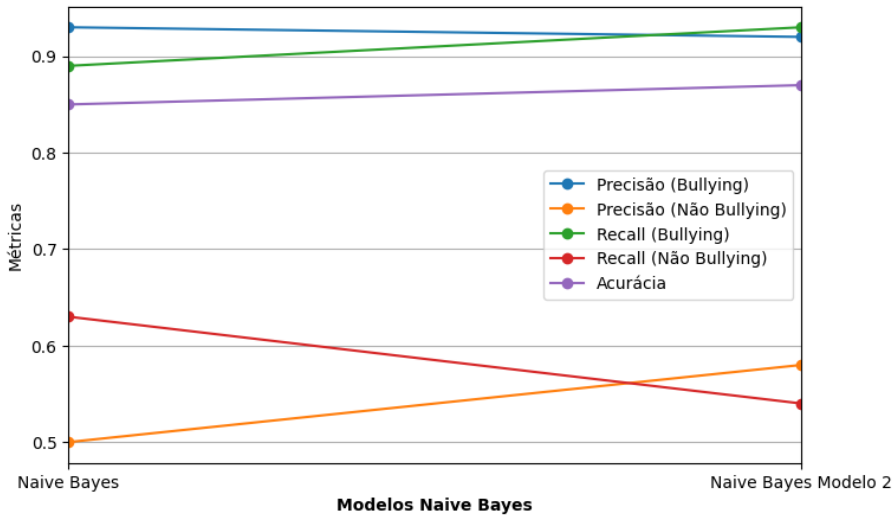


Figure 2. Figura 2- Resultados do Modelo Naive Bayes

observamos uma precisão de 0.91 para a classe 0 e 0.55 para a classe 1, com um recall de 0.93 para a classe 0 e 0.49 para a classe 1. Isso resulta em um f1-score de 0.92 para a classe 0 e 0.52 para a classe 1, com uma acurácia geral de 0.87. Introduzindo dropout no modelo, observamos uma leve diminuição na precisão para ambas as classes, com um f1-score de 0.92 para a classe 0 e 0.51 para a classe 1. Por outro lado, a inclusão de atenção e o otimizador Adam resultaram em uma precisão de 0.92 para a classe 0 e 0.53 para a classe 1, com um f1-score de 0.92 para a classe 0 e 0.55 para a classe 1. No entanto, essas melhorias na precisão foram acompanhadas por uma diminuição na acurácia geral do modelo para 0.86. Esses resultados destacam a importância de explorar e ajustar diferentes técnicas para otimizar o desempenho dos modelos de GRU na tarefa de classificação de textos relacionados a bullying.

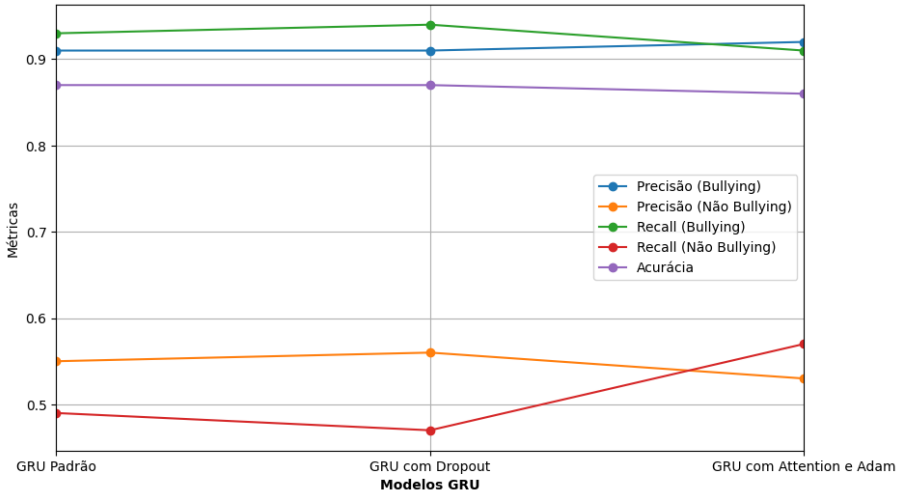


Figure 3. Figura 3- Resultados do Modelo GRU

4.3. Discurssão

Os resultados evidenciam que a adição de técnicas como batch normalization e ajuste no learning rate impactam positivamente a performance do Bi-LSTM. No entanto, é necessário cautela para evitar overfitting. O Naive Bayes demonstrou boa performance, especialmente no Modelo 2. A GRU apresentou resultados competitivos, sendo importante considerar a aplicação de técnicas como dropout para melhorar o desempenho. A escolha do modelo dependerá do contexto de aplicação e da importância dada a diferentes métricas de avaliação.

5. Conclusão

Ao concluir esta análise, destacamos a importância de considerar o contexto específico da aplicação ao decidir pela implementação de alterações nos métodos. A escolha entre simplicidade e complexidade do modelo deve ser guiada pelas características da tarefa em questão, com uma atenção especial para o equilíbrio entre ganhos de desempenho e a robustez do modelo.Recomenda-se uma abordagem cautelosa ao explorar ajustes mais avançados nos algoritmos, considerando métricas como overfitting e underfitting. O processo de aprimoramento deve ser iterativo, com avaliações constantes para garantir que as modificações aplicadas contribuam positivamente para a detecção eficaz de bullying em mensagens online.

Com base nos resultados apresentados, a recomendação se inclina para o modelo Naive Bayes Modelo 2 como a escolha mais promissora para a detecção de bullying em mensagens online. Este modelo evidenciou melhorias consistentes na precisão e recall em relação ao modelo original de Naive Bayes, enquanto manteve uma acurácia geral estável. A refinada abordagem adotada, incorporando um pipeline mais eficiente e a otimização dos hiperparâmetros por meio do GridSearchCV, destaca-se como uma estratégia eficaz para alcançar um equilíbrio satisfatório entre desempenho e robustez do modelo. Embora os modelos Bi-LSTM e GRU tenham demonstrado competitividade, suas variações de desempenho em resposta às técnicas empregadas ressaltam a necessidade de uma investigação mais profunda. Em resumo, o Naive Bayes Modelo 2 emerge

como uma opção sólida e eficiente, alinhada aos objetivos de detecção de bullying online, atingindo um patamar satisfatório de desempenho e adaptabilidade.

References

[1] Ali Alshehri, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. "Understanding and detecting dangerous speech in social media". In: *arXiv.org* (May 2020). URL: https://arxiv.org/abs/2005.06608.

[2] Donghyeon Won University of California et al. *Protest activity detection and perceived violence estimation from social media images: Proceedings of the 25th ACM International Conference on Multimedia*. Oct. 2017. URL: https://dl.acm.org/doi/abs/10.1145/3123266.3123282?casa_token=w_PYml8OKL0AAAAA%3ApVcDW_xfCP7ukrxtYxos-J9yD9lIWYbn4KwZu-2mUH9QWVf3-UbwNeF6Ie_COQ7nY__isnCFSdU_sg.

[3] Vinicius Matheus de Medeiros Silva Coutinho and Yuri Malheiros. "Detecção de Mensagens Homofóbicas em Português no Twitter usando Análise de Sentimentos". In: (2020), pp. 1–12.

[4] Aron Culotta and Jeffrey Sorensen. "Dependency tree kernels for relation extraction". In: *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL)*. 2004.

[5] Anderson Almeida Firmino. "Uma abordagem para detecção de discurso de ódio utilizando aprendizado de máquina baseado em cruzamento de idiomas". Tese (Doutorado em Ciência da Computação) – Centro de Engenharia Elétrica e Informática. PhD thesis. Campina Grande: Universidade Federal de Campina Grande, 2022, p. 97.

[6] Peter Dias Paiva, Vanecy Matias da Silva, and Raimundo Santos Moura. "Detecção automática de discurso de ódio em comentários online". In: *Escola Regional de Computação Aplicada à Saúde (ERCAS), 7th*. Sociedade Brasileira de Computação. Teresina, 2019, pp. 157–162.

[7] Adriano dos Santos Rodrigues da Silva. "Estudo de modelos distribucionais para detecção de discurso de ódio em português". Master's Dissertation. PhD thesis. Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, 2021. DOI: 10.11606/D.100.2021.tde-28012022-074813. URL: https://doi.org/10.11606/D.100.2021.tde-28012022-074813.

[8] Douglas de Oliveira Trajano. "Detecção de linguagem tóxica aplicada a textos em português". In: (2023).