Investigando o Uso de Análise de Sentimentos para Identificar Comportamentos de Bullying em Grupo de WhatsApp Escolar

Mannoella Renata L. Pereira, Lucas P. Alves, Thereza Patrícia P. Padilha

Curso de Licenciatura em Ciência da Computação (LCC)
Universidade Federal da Paraíba (UFPB) Rio Tinto - PB - Brasil

{mannoella.lima, lucas.pessoa, thereza}@dcx.ufpb.br

Abstract. *Due to the increase in digital communication, cyberbullying has become a critical issue in schools, with various negative impacts as it can affect the well-being and academic performance of victims. Thus, this paper presents ongoing research on the use of sentiment analysis techniques to identify signs of bullying behavior in WhatsApp groups within the school environment. A sentiment classification model is being trained using logistic regression and advanced models such as BERT to identify sentiments and potential bullying behaviors. First results are presented.*

Resumo. *Devido ao aumento da comunicação digital, o bullying virtual (e-bullying) tornou-se um problema crítico nas escolas, pois pode afetar o bem-estar e o desempenho acadêmico das vítimas. Assim, este artigo apresenta uma pesquisa em andamento sobre o uso da técnica de análise de sentimentos para identificar indícios de comportamentos de bullying em grupos de WhatsApp no ambiente escolar. Um modelo de classificação de sentimentos está sendo treinado utilizando regressão logística e modelos avançados como BERT para identificar sentimentos e potenciais comportamentos de bullying. Resultados preliminares são apresentados.*

1. Introdução

Bullying é um comportamento agressivo, repetitivo, que envolve um desequilíbrio de poder entre o agressor e a vítima e pode ocorrer em diferentes formas, como física, verbal, social e virtual, impactando significativamente a vida das vítimas [Olweus, 1993]. Atualmente, com o advento das plataformas digitais para comunicação entre pessoas, o bullying de forma virtual (e-bullying) se tornou uma grande preocupação devido ao seu impacto negativo na saúde mental e no bem-estar das vítimas. As plataformas digitais, como redes sociais (Facebook, Twitter e Instagram) e aplicativos de mensagens instantâneas (WhatsApp e Telegram), facilitam a disseminação rápida e ampla de mensagens, potencializando tanto os efeitos positivos quanto negativos da comunicação online.

O WhatsApp, por exemplo, é um dos aplicativos de mensagens instantâneas mais populares, oferecendo uma maneira conveniente e eficiente de comunicação para bilhões de usuários. Encontra-se disponível em dispositivos móveis e desktops, permitindo não somente o envio de mensagens de texto, mas também fotos, vídeos e documentos, além de chamadas de voz e vídeo, de forma individual ou para um grupo [Church e de Oliveira, 2013].

DOI: 10.5753/sbie.2024.244966

No contexto escolar (incluindo universitário), o WhatsApp tem sido um recurso tecnológico amplamente utilizado para facilitar a comunicação entre gestores, professores e alunos. A criação de grupos específicos para cada turma é uma prática comum, permitindo que os alunos discutam questões relacionadas à escola e aumentem a interação entre si. No entanto, há preocupações significativas sobre a ocorrência de comportamentos inadequados, como o bullying, que podem afetar o bem-estar dos alunos vítimas e, assim, impactar negativamente o desempenho escolar. Essas preocupações destacam a importância de monitorar as mensagens e abordar esses comportamentos para garantir um ambiente seguro e positivo para todos os participantes. Mesmo que um professor faça parte do grupo, analisar uma grande quantidade de mensagens e estar sempre disponível e atento a esse ponto, é uma tarefa bastante desgastante.

Além disso, as interações via mensagens de texto frequentemente contêm nuances e contextos que são difíceis de interpretar sem uma análise detalhada. A privacidade e a informalidade dessas plataformas tornam a detecção de comportamentos abusivos mais complexa. Neste cenário, a Análise de Sentimentos (AS), uma subárea do processamento de linguagem natural (PLN), surge como uma alternativa valiosa. AS envolve a identificação e a extração de informações subjetivas em grandes quantidades de textos, aplicando técnicas de PLN e aprendizado de máquina [Liu, 2012]. Logo, é possível analisar automaticamente o tom e o conteúdo emocional das mensagens, facilitando a detecção precoce de comportamentos de bullying. Diante do contexto, este artigo apresenta o andamento de uma pesquisa que visa aplicar essas técnicas de análise de sentimentos em mensagens de grupo de WhatsApp usado na disciplina de "Lógica Aplicada à Computação", do 2° período do curso de Licenciatura em Ciência da Computação da UFPB, com a finalidade de identificar possíveis comportamentos de bullying.

2. Análise de Sentimentos

A análise de sentimentos, também conhecida como mineração de opinião, é uma abordagem clássica da área de PLN que visa identificar as opiniões, sentimentos, emoções e atitudes expressas em textos [Pang e Lee, 2008]. Para isso, AS detecta, extrai e classifica as informações, determinando se o texto apresentado expressa uma opinião/sentimento positivo, negativo ou neutro sobre um assunto. A identificação do tipo de sentimento em um texto é complexa, pois pode incluir sarcasmo, ironia e nuances emocionais que podem distorcer a interpretação das emoções expressas.

Historicamente, a AS era um tema pouco explorado até o ano de 2000 devido à escassez de textos disponíveis em formato digital. No entanto, com a explosão da Internet e, sobretudo, das redes sociais, houve uma proliferação de dados textuais acessíveis e em diferentes formatos e que podem ser utilizados como recurso para mineração de textos. A partir de então, a análise de sentimentos se expandiu e continua a ser uma área de intenso desenvolvimento e pesquisa em diversos setores [Liu, 2012]. Por exemplo, empresas utilizam para monitorar a satisfação do cliente e a reputação da marca, enquanto instituições de pesquisa aplicam esses métodos para estudar fenômenos sociais e comportamentais. AS tornou-se uma ferramenta essencial para a tomada de decisões estratégicas e informadas em um mundo cada vez mais digital e orientado por dados. Como ainda há muito a ser explorado, a análise de sentimentos continua sendo um potencial para novas descobertas e avanços tecnológicos, incluindo áreas sensíveis como a detecção de bullying em ambientes escolares digitais.

Do ponto de vista técnico, a análise de sentimentos utiliza diversas abordagens. Uma das principais abordagens é o uso de dicionários léxicos, que associam palavras a polaridades de sentimentos (positivo, negativo ou neutro), permitindo a atribuição de pontuações de sentimento a cada palavra e, consequentemente, o cálculo de uma pontuação geral do texto. Outra abordagem comum é o aprendizado de máquina supervisionado, em que algoritmos são treinados em conjuntos de dados rotulados para classificar novos textos automaticamente. Técnicas mais avançadas, como redes neurais convolucionais (*Convolutional Neural Networks - CNNs*) e modelos de linguagem pré-treinados como BERT, capturam relações semânticas complexas e contextuais entre palavras, proporcionando uma compreensão mais profunda e precisa das emoções expressas nos textos [Devlin *et al.*, 2019].

3. Trabalhos Relacionados

A detecção de bullying em ambientes digitais tem sido alvo de estudos por diversos pesquisadores. A seguir são apresentados quatro trabalhos recentes envolvendo a área de análise de sentimentos e bullying.

Paul e Saha (2020) introduziram uma nova aplicação do BERT para a identificação de cyberbullying, utilizando técnicas avançadas de PLN, incluindo o modelo BERT e ajuste fino, além de análise de sentimentos. O modelo de classificação baseado em BERT alcançou excelentes resultados em corpora renomadas, tais como Formspring, Twitter e Wikipédia. Os experimentos demonstraram melhorias em relação às abordagens existentes de detecção de cyberbullying, superando modelos de redes neurais profundas baseados em slot ou baseados em atenção na detecção de comportamentos abusivos em plataformas de mídia social.

Já o estudo de Tapia *et al.* (2018) concentrou-se na detecção de cyberbullying na rede social Twitter utilizando técnicas de mineração de dados para identificar padrões comportamentais indicativos de comportamentos agressivos e intimidadores. A pesquisa focou especificamente em páginas com termos em espanhol, adaptando as análises para considerar nuances linguísticas e culturais que possam influenciar a detecção de cyberbullying. As plataformas Mr. Tweet e Sentimento140 foram usadas nessas etapas otimizando assim o trabalho de detecção.

O estudo de Urtig e Castro (2018) investigou a criação e implementação de um modelo de mineração de texto para detectar cyberbullying em redes sociais. Focado em adolescentes, o trabalho identificou padrões de comportamento agressivo em plataformas como Twitter, Facebook, Tumblr e Blogger. Utilizando uma abordagem sistemática, os padrões identificados foram compilados em um checklist, que pode ser usado por escolas para monitorar comportamentos de crimes virtuais. O estudo enfatizou a importância da intervenção preventiva e do papel das escolas e pais na identificação e mitigação de comportamentos de bullying.

Outro estudo relevante foi conduzido por Pfitscher *et al.* (2023), que investigaram o uso da análise de sentimentos para apoiar a permanência estudantil em turmas de programação. Utilizando uma abordagem de coleta ativa de sentimentos e PLN, os autores identificaram padrões emocionais nos alunos com uma precisão de 68%. Os resultados sugeriram que essa técnica pode ser aplicada para prever evasão escolar e implementar intervenções pedagógicas mais eficazes, o que complementa as abordagens

de detecção de bullying ao possibilitar um monitoramento mais detalhado dos sentimentos dos estudantes em contextos educacionais.

Além disso, Silva *et al.* (2018) analisaram técnicas de pré-processamento, conjuntos de atributos e classificadores aplicados à tarefa de detecção de traços de bullying em textos em português do Brasil, retirados de redes sociais. Diversos classificadores e configurações de atributos foram estudados e comparados para determinar a abordagem mais eficaz. Os resultados indicaram que a combinação de unigramas e bigramas com o classificador SVM (*Support Vector Machine*) utilizando kernel RBF (*Radial Basis Function*) produziu os melhores resultados em termos de precisão e eficácia. Os autores destacaram a importância do uso de grandes conjuntos de dados de treinamento e da configuração adequada dos atributos para melhorar a detecção de bullying e a identificação dos papéis dos autores nos episódios relatados.

A pesquisa apresentada neste artigo se diferencia dos trabalhos anteriores ao se concentrar especificamente na análise de mensagens de texto no WhatsApp, plataforma amplamente utilizada em ambientes escolares. O WhatsApp apresenta características linguísticas únicas, como o uso frequente de gírias, abreviações e emojis, que diferem das usadas em redes sociais públicas abordadas nos trabalhos anteriores. Essa especificidade permite uma análise mais aprofundada das nuances linguísticas específicas dessa plataforma. Além disso, como o foco é exclusivo no grupo do WhatsApp do ambiente escolar, busca-se compreender melhor os padrões de comunicação e interações atuais entre os alunos, tornando a análise desse ambiente fundamental para uma detecção mais precisa e contextualizada do bullying.

4. Metodologia

A metodologia deste trabalho seguiu as etapas do processo de descoberta de conhecimento de dados de forma estruturada, iniciando pela obtenção dos dados, seguida pela seleção das técnicas apropriadas para o problema em questão. A primeira etapa consistiu na coleta de dados de mensagens de grupos de WhatsApp no ambiente escolar. Em seguida, a etapa de pré-processamento, consistiu na preparação dos dados para remover ruídos textuais e vetorizados utilizando a técnica TF-IDF (*Term Frequency-Inverse Document Frequency*). A etapa de mineração de dados teve como meta a aplicação de técnicas de classificação, como regressão logística e modelos avançados como BERT, para identificar sentimentos e comportamentos de bullying. Por fim, na etapa de validação, os resultados foram observados para validar a eficácia do modelo gerado, garantindo a relevância e a precisão das previsões realizadas, conforme passos do processo ilustrados na Figura 1.

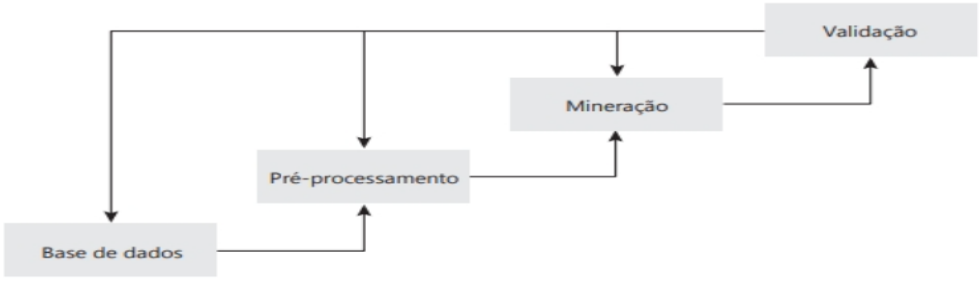

Figura 1. Descoberta de Conhecimento em Bases de Dados (Silva & Ferrari, 2016).

5. Resultados Iniciais e Discussão

Para a construção de um modelo de AS eficaz para a detecção de comportamentos de bullying, inicialmente, foram coletadas 1.576 mensagens de um grupo Whatsapp, referente ao período de março a agosto de 2024. Importante ressaltar que um termo de privacidade e proteção dos dados seguindo as diretrizes éticas e regulatórias foi assinado pelos autores do projeto. Em seguida, as mensagens foram exportadas para um arquivo .CSV e carregado como um objeto do tipo DataFrame da biblioteca Pandas para pré-processamento:

- limpeza: remoção de *stopwords*, links, emojis e caracteres especiais;
- tokenização: divisão de texto em unidades menores;
- normalização: lemmatização e conversão para minúsculas utilizando NLTK e *SpaCy.*

O conjunto de dados original continha 45.635 caracteres. Durante o processo de limpeza dos dados, foram removidos 8.707 caracteres (*stopwords,* URLs e 125 emojis), o que resultou em 36.928 caracteres. A biblioteca *TextBlob* foi também utilizada para uma análise léxica adicional. A técnica TF-IDF foi empregada para a vetorização das mensagens, transformando os textos em representações numéricas adequadas para análise. Para a análise de sentimentos, foram aplicados métodos de classificação e regressão, incluindo Regressão Logística, Máquinas de Vetores de Suporte (SVM) e *Random Forest*. As bibliotecas Pandas e *Seaborn* foram utilizadas para a manipulação, organização e visualização dos dados, respectivamente. O modelo VADER, ajustado para o Português do Brasil, foi utilizado para fornecer scores de polaridade precisos. As mensagens foram classificadas em três categorias: negativo (4), neutro (1.510) e positivo (63), conforme mostradas na Figura 2. Quanto à acurácia, o melhor modelo gerado obteve um desempenho impressionante, com uma taxa de acerto de 93%. Esse resultado evidencia a eficácia do modelo em classificar corretamente as mensagens, mostrando sua precisão na identificação dos diferentes sentimentos presentes no conjunto de dados.

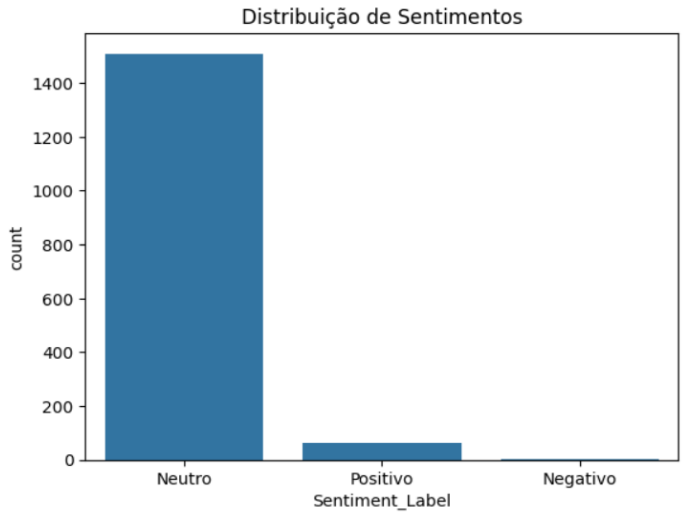

Figura 2. Distribuição de Classificação dos Sentimentos.

É importante refletir sobre os métodos usados e resultados obtidos. A etapa de pré-processamento, que incluiu a remoção de emojis, *stopwords* e URLs, foi fundamental para assegurar a qualidade dos dados e a precisão da análise de sentimentos. A remoção de 125 emojis, por exemplo, teve um impacto significativo na clareza e consistência das análises subsequentes. A combinação da técnica TF-IDF para vetorização com métodos de classificação e regressão, como Regressão Logística, Máquinas de Vetores de Suporte (SVM) e *Random Forest*, bem como com modelos avançados como BERT, revelou-se uma abordagem robusta para a identificação de padrões e sentimentos nas mensagens. Essas técnicas forneceram uma visão mais precisa e detalhada dos sentimentos expressos, demonstrando a eficácia das metodologias aplicadas. Esses resultados iniciais foram cruciais para entender o impacto das técnicas de processamento de texto e análise de sentimentos no contexto do bullying em ambientes virtuais. As observações aqui apresentadas estabelecem uma base sólida para a realização de análises mais aprofundadas e a implementação de melhorias contínuas nos modelos e métodos utilizados. A integração de análises mais detalhadas e a consideração de diversas abordagens analíticas serão essenciais para avançar no estudo e obter conclusões mais robustas.

6. Conclusão

A identificação antecipada de comportamentos de bullying no espaço escolar pode prevenir diversos problemas como evasão escolar, falta de interesse pelos estudos, dificuldades de concentração e socialização. Assim, a técnica de análise de sentimentos pode ser uma aliada nesse acompanhamento, pois permite monitorar as interações dos alunos de forma contínua e eficaz. Este artigo apresentou um trabalho envolvendo a combinação de técnicas de pré-processamento de texto, vetorização utilizando TF-IDF e a aplicação de algoritmos de aprendizado de máquina como regressão logística e BERT, para a criação de um modelo. Os resultados iniciais indicaram uma alta precisão na classificação dos sentimentos expressos nas mensagens, evidenciando o potencial dessa abordagem para auxiliar educadores e gestores escolares na criação de um ambiente mais seguro e acolhedor.

Além disso, a metodologia desenvolvida pode ser adaptada e expandida para outras plataformas e contextos, oferecendo um caminho promissor para a integração de tecnologias de inteligência artificial no combate ao bullying escolar. Entretanto, é crucial que a implementação desses sistemas seja acompanhada de medidas rigorosas de proteção à privacidade dos alunos e de políticas claras de intervenção e suporte às vítimas. Portanto, acredita-se que a aplicação da análise de sentimentos ao monitoramento de interações escolares digitais representa uma inovação significativa, com potencial para transformar a maneira como as instituições de ensino lidam com o bullying, promovendo um ambiente mais positivo e seguro para todos os estudantes.

Referências Bibliográficas

Church, K., & De Oliveira, R. (2013). What's up with WhatsApp? Comparing mobile instant messaging behaviors with traditional SMS, *15th International Conference on Human-Computer Interaction with Mobile Devices and Services, p. 352-361*.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding, *Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1. Minneapolis, p. 4171-4186.

Liu, B. (2012). Sentiment analysis and opinion mining, Morgan & Claypool Publishers.

Olweus, D. (1993). Bullying at school: What we know and what we can do, Oxford: Blackwell.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval, v. 2, n. 1-2, p. 1-135*.

Paul, S., & Saha, S. (2020). CyberBERT: BERT para identificação de cyberbullying, *Multimedia Systems, v. 28, p. 1897–1904*.

Pfitscher, R. J., Camargo, L. C., Moreira, B. G., Wang, C., Zedral, R., & Garcia, T. R. (2023). Análise de sentimentos em turmas de programação com vistas ao apoio à permanência estudantil, *Simpósio Brasileiro de Informática na Educação (SBIE), p. 1329-1340*. doi: https://doi.org/10.5753/sbie.2023.234753.

Silva, G. M., Silva, N. F. F., & Dias, M. de S. (2018). Detecção de bullying: como identificar automaticamente essa prática em redes sociais?, *Revista de Sistemas de Informação da FSMA, Campos dos Goytacazes, v. 21, p. 11-19*.

Silva, L. N. de C., & Ferrari, D. G. (2016). Introdução à mineração de dados. Conceitos básicos, algoritmos e aplicações. 2. ed.

Tapia, F., Aguinaga, C., & Luje, R. (2018). Detection of behavior patterns through social networks like Twitter, using data mining techniques as a method to detect cyberbullying, *7th International Conference on Software Process Improvement (CIMPS), p. 111-118*.

Urtig, L. A. N., & Castro, M. A. N. (2018). Análise de sentimentos e suas aplicações na educação: uma revisão de literatura, *Anais do Simpósio Brasileiro de Informática na Educação, v. 29, n. 1, p. 1002-1011.* doi: http://milanesa.ime.usp.br/rbie/index.php/sbie/article/view/8129/5820.