Comparação de Ferramentas para Análise de Sentimentos Aplicada no Contexto Educacional

Benjamin G. Moreira[1], **Luiz C. Camargo**[2], **Ricardo J. Pfitscher**[1], **Tatiana R. Garcia**[1]

[1]Universidade Federal de Santa Catarina (UFSC)
Joinville – SC – Brasil

[2]Centro Universitário Católica de Santa Catarina
Joinville – SC – Brasil

benjamin.moreira@ufsc.br, lzcamargo@outlook.com

ricardo.pfitscher@ufsc.br, tatiana.garcia@ufsc.br

Abstract. *This article is part of a project to mitigate dropout in programming courses in Brazilian higher education, using sentiment analysis combined with psycho-pedagogical methods and Natural Language Processing (NLP) techniques. The present research compares the effectiveness of three automated sentiment extraction tools - two based on Large Language Models (LLMs) and a lexical analyzer - using a database of 540 student responses.*

Resumo. *Este artigo faz parte de um projeto voltado para combater a evasão em disciplinas de programação no ensino superior brasileiro, utilizando análise de sentimentos combinada com métodos psicopedagógicos e técnicas de Processamento de Linguagem Natural (PLN). A presente pesquisa compara a eficácia de três ferramentas automatizadas de extração de sentimentos, duas baseadas em Large Language Models (LLMs) e um analisador léxico, usando uma base de 540 respostas de estudantes.*

1. Introdução

A análise de sentimentos é o estudo computacional das opiniões, atitudes e emoções das pessoas em relação a uma entidade. A entidade pode representar indivíduos, eventos ou tópicos. Normalmente a análise de sentimentos está associada à uma opinião sobre uma entidade, expresso em uma revisão sobre a mesma [Medhat et al. 2014]. No contexto estudantil, a análise de sentimentos tem sido vista com potencial de aplicação em diversos contextos, seja para capturar a satisfação ou o *feedback* dos estudantes sobre um determinado curso [Rani and Kumar 2017, Neumann and Linzmayer 2021], seja para compreender as dificuldades dos estudantes [Atiq and Loui 2022], ou até para mitigar potenciais casos de desistência ou evasão [Bóbó et al. 2022].

Este trabalho é parte de um projeto de pesquisa com a finalidade de investigar a utilização do sentimento de estudantes com vistas ao uso de ferramentas automatizadas para apoio à permanência estudantil. A pesquisa utiliza a análise de sentimentos coletados de forma ativa dos estudantes, combinando métodos psicopedagógicos e técnicas automatizadas baseadas em Processamento de Linguagem Natural (PLN).

Há diferentes opções na literatura para extração dos sentimentos através de PLN. As formas mais consolidadas consistem na aplicação de analisadores léxicos, modelos

supervisionados de classificação treinados com bases públicas de textos de sentimentos (normalmente com bases de avaliações sobre produtos e filmes, como é o caso do IMDb) [Wankhade et al. 2022]. Por outro lado, o crescente interesse nas ferramentas de inteligência artificial generativas, que utilizam de Large Language Models (LLMs) para construção de conhecimento [Chang et al. 2024], tem instigado pesquisadores de diversas áreas a aplicar essas ferramentas na extração de sentimentos [Mughal et al. 2024].

2. Trabalhos relacionados

A análise de sentimentos é realizada em diversos domínios de aplicação. Em [Lazarini et al. 2023, Seno et al. 2023], o objetivo é entender melhor o sentimento público em debates políticos. No contexto educacional, em [Atiq and Loui 2022], é apresentado um estudo sobre as emoções de estudantes durante a realização de atividades de programação, com observações qualitativas. Em [Bóbó et al. 2022], a análise de sentimentos é utilizada para prever o risco de evasão com dados coletados de textos presentes em um ambiente virtual de aprendizagem. Ainda no domínio educacional, um mapeamento sistemático é apresentado em [Coto et al. 2022].

Outro trabalho próximo a esse artigo é apresentado em [Mughal et al. 2024], em que uma extensiva análise é realizada para comparar ferramentas de aprendizado profundo e LLMs quanto ao desempenho na análise de sentimentos baseada em aspectos (ABSA, do inglês aspect-based sentiment analysis). Os resultados da avaliação de bases de dados públicas de domínios específicos (opiniões sobre hotéis, restaurantes e livros) mostram que o modelo PaLM apresentou os melhores resultados na maioria dos casos, sendo inferior ao GPT3.5 em uma base de dados projetada para conter ao menos dois sentimentos em uma mesma frase.

3. Metodologia

Os dados utilizados neste trabalho foram obtidos a partir da coleta ativa de sentimentos de alunos. Detalhes do processo são apresentados em [Pfitscher et al. 2023]. A coleta foi realizada por meio de um questionário respondido de forma anônima. Duas perguntas foram realizadas: (1) Como você está se sentindo em relação à disciplina?; e (2) Como você está se sentindo em relação à universidade?

Cada resposta recebeu uma classificação psicopedagógica, realizada de modo manual por duas psicopedagogas, que empiricamente buscaram distinguir termos de sentimentos dos demais em cada frase das respostas dos estudantes. Para esta classificação, primeiramente as avaliadoras definiram as categorias e estabeleceram as listas de itens sobre os aspectos positivos e negativos na experiência com a disciplina/universidade utilizando a análise de conteúdo [Bardin 1977], que é uma análise qualitativa do discurso manifesto da comunicação. Foi estabelecido um conjunto de categorias de classificação para orientar esse processo:

- *Positivo:* respostas exclusivamente positivas;
- *Negativo:* respostas exclusivamente negativas;
- *Ambos:* respostas com menções positivas e negativas;

A classificação como *Ambos* mostrou-se necessária uma vez que diversas respostas são formadas por diversas frases, podendo indicar tanto aspectos positivos quanto negativos dos itens avaliados.

Para uma classificação automática dos sentimentos, foram comparados três modelos, dois deles baseados em LLMs (ambas as LLMs são opções open source) e outro com análise baseada em Léxico. Segue descrição dos modelos:

- Llama[1] (versão Meta Llama 3 Instruct I 8B): O Llama 3 é um LLM desenvolvido pela empresa Meta. A versão utilizada é o modelo com 8 bilhões de parâmetros (o menor disponível).
- Gemma[2] (versão Gemma 2 9B IT): O Gemma 2 é o LLM de código aberto do Google. A versão utilizada é o modelo com 9 bilhões de parâmetros.
- LeIA[3] (Léxico para Inferência Adaptada): é uma adaptação para o português do léxico e ferramenta para análise de sentimentos VADER (Valence Aware Dictionary and sEntiment Reasoner).

Para as LLMs, foi utilizado como *prompt* do sistema a seguinte entrada: "Analise o sentimento da frase e responda apenas com POSITIVO, NEGATIVO ou AMBOS". Também foi fornecida a temperatura de respostas com o valor zero[4].

Para utilização do LeIA, foi definido que o parâmetro *compound* superior à 0,05 foi considerado *Positivo*, um *compound* inferior à -0,05 foi considerado *Negativo*, e o intervalo utilizado para definir o sentimento como *Ambos*. Essa faixa da classificação como *Ambos* é considerada para uma pontuação neutra, na qual existe um equilíbrio entre termos positivos e negativos, mas foi considerada nesse trabalho como uma possibilidade para qualificar a existência das duas polaridades no mesmo texto.

4. Resultados

Nesta seção, são apresentados os resultados dos modelos aplicados em comparação com as classificações psicopedagógicas realizadas.

Os modelos obtiveram acurácia aproximada de 75%, 77% e 62% para o Llama, Gemma e LeIA, respectivamente. Embora o Gemma tenha obtido melhor acurácia, uma análise a partir das matrizes de confusão (conforme a Figura 1), mostra que o Gemma diminuiu o acerto para sentimentos positivos e negativos. Por outro lado, o Llama teve melhor desempenho em classificar os sentimentos positivos e negativos, mas praticamente não conseguiu classificar os sentimentos definidos como *Ambos*, classificando esses, em sua maioria, como negativos. Já o modelo LeIA teve uma taxa de acertos na classificação de sentimentos positivos e negativos similar ao Gemma, mas a classificação para a categoria *ambos*, mostrou-se bastante ruim.

Aprimorando a análise e, uma vez que as classes estão desbalanceadas, é considerado o resultado da métrica F1-score, que apresenta uma medida mais robusta do desempenho do modelo, levando em consideração tanto a precisão quanto o revocação nas classificações. Os valores percentuais apresentados nas discussões sobre o F1-score são apresentados de forma aproximada.

[1]Site do projeto do Llama: https://llama.meta.com

[2]Site do projeto do Gemma: https://ai.google.dev/gemma

[3]Site do projeto do LeIA: https://github.com/rafjaa/LeIA

[4]Utiliza-se uma temperatura baixa para que a LLM forneça respostas mais precisas, o que é útil quando se deseja consistência e precisão.

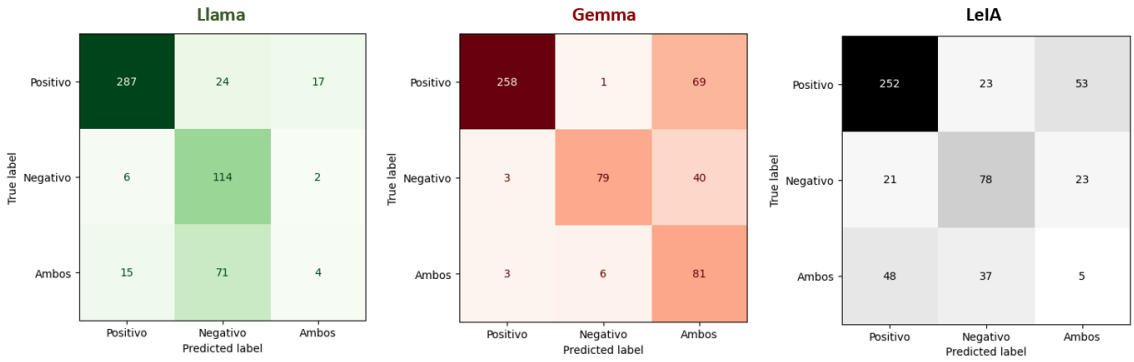

Figura 1. Matrizes de confusão da classificação dos sentimentos

Os três modelos obtiveram F1-score considerado bom na classificação de sentimentos positivos, sendo que a Llama obteve acerto de 90%, enquanto o Gemma obteve 87% e o LeIA obteve 77%. A qualidade da classificação realizada pelos modelos diminui quanto aos sentimentos negativos, sendo que o Llama obteve 69%, Gemma alcançou 76% e LeIA obteve 60%. Observa-se que, embora o Llama tenha obtido uma melhor classificação dos sentimentos positivos, para o sentimento negativo o F1-score é reduzido significativamente (21% do Llama, contra 11% do Gemma).

Para classificação como *Ambos*, o Llama obteve 71%, Gemma 58% e LeIA apenas 6%. O Llama teve desempenho similar na classificação de *Ambos* quanto nos sentimentos negativos, enquanto o Gemma teve uma redução significativa. Quanto ao LeIA, sua classificação para a categoria *Ambos* é insignificante. Esses valores representam aspectos percebíveis a partir das matrizes de confusão, onde existe um "espalhamento" das classificações de sentimentos negativos e ambos.

5. Considerações finais

Neste artigo, dados coletados ativamente de estudantes de programação foram utilizados na extração e análise de sentimento estudantil. Sob a perspectiva da métrica F1-score, os resultados obtidos com as ferramentas são considerados bons com a classificação de sentimentos "positivos". Para os sentimentos "negativos", a qualidade da classificação não é mantida e há um declínio de desempenho das três ferramentas. O desempenho segue em declínio para as ferramentas Llama e LeIA na classificação "Ambos", mas não para a ferramenta Gemma, que manteve o nível desempenho para esse tipo de classificação próximo do obtido com a classificação "negativos".

Uma vez que a classificação como *Ambos* possui menor acerto, uma alternativa é classificar as respostas por frases, indicando quantos sentimentos positivos e negativos estão presentes na resposta. Como a classificação de sentimentos como positivos e negativos apresentou boa taxa de acertos, o resultado da classificação pode ser mais confiável.

Para a aplicação pretendida, que envolve a permanência estudantil, é considerado como mais importante a identificação dos sentimentos negativos. Dessa forma, a escolha do Llama é a mais indicada, uma vez que essa identifica melhor os sentimentos negativos e positivos, bem como as classificações incorretas dos *Ambos* serem, em sua maioria, classificadas como negativos (entende-se que os aspectos negativos podem ter potencial em impactar na relação do estudante com a disciplina).

Referências

Atiq, Z. and Loui, M. C. (2022). A qualitative study of emotions experienced by first-year engineering students during programming tasks. *ACM Transactions on Computing Education (TOCE)*, 22(3):1–26.

Bardin, L. (1977). *Análise de conteúdo*. Lisboa: edições 70.

Bóbó, M. L., Campos, F., Stroele, V., David, J. M. N., Braga, R., and Torrent, T. T. (2022). Using sentiment analysis to identify student emotional state to avoid dropout in e-learning. *International Journal of Distance Education Technologies (IJDET)*, 20(1):1–24.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Coto, M., Mora, S., Grass, B., and Murillo-Morera, J. (2022). Emotions and programming learning: systematic mapping. *Computer Science Education*, 32(1):30–65.

Lazarini, L., Anno, F. S. I., Seno, E. R. M., and Caseli, H. M. (2023). Abordagens baseadas em léxicos para a classificação de sentimentos orientada aos alvos de opinião em comentários do domínio político. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 375–380. SBC.

Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

Mughal, N., Mujtaba, G., Shaikh, S., Kumar, A., and Daudpota, S. M. (2024). Comparative analysis of deep natural networks and large language models for aspect-based sentiment analysis. *IEEE Access*, 12:60943–60959.

Neumann, M. and Linzmayer, R. (2021). Capturing student feedback and emotions in large computing courses: A sentiment analysis approach. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, pages 541–547.

Pfitscher, R., Camargo, L., Moreira, B., Wang, C., Zedral, R., and Garcia, T. (2023). Análise de sentimentos em turmas de programação com vistas ao apoio à permanência estudantil. In *Anais do XXXIV Simpósio Brasileiro de Informática na Educação*, pages 1329–1340, Porto Alegre, RS, Brasil. SBC.

Rani, S. and Kumar, P. (2017). A sentiment analysis system to improve teaching and learning. *Computer*, 50(5):36–43.

Seno, E. R. M., Anno, F. S. I., Lazarini, L., and Caseli, H. M. (2023). Classificação de polaridade orientada aos alvos de opinião em comentários sobre debate político em português. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 84–93. SBC.

Wankhade, M., Rao, A. C. S., and Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.