Análise de sentimentos: uma avaliação do impacto de respostas ambíguas na análise de sentimentos

Benjamin Grando Moreira
Universidade Federal de Santa Catarina (UFSC)
Joinville, Santa Catarina, Brazil
benjamin.moreira@ufsc.br

Ricardo José Pfitscher
Universidade Federal de Santa Catarina (UFSC)
Joinville, Santa Catarina, Brazil
ricardo.pfitscher@ufsc.br

Luiz Carlos Camargo
Centro Universitário Católica de Santa Catarina
Joinville, Santa Catarina, Brazil
luiz.camargo@catolicasc.edu.br

Tatiana Renata Garcia
Universidade Federal de Santa Catarina (UFSC)
Joinville, Santa Catarina, Brazil
tatiana.garcia@ufsc.br

Luiz Gustavo Cordeiro
Universidade Federal de Santa Catarina (UFSC)
Joinville, Santa Catarina, Brazil
luiz.cordeiro@ufsc.br

ABSTRACT

Sentiment analysis, which involves the computational study of people's opinions and emotions, has been applied in the student context to understand satisfaction and difficulties and mitigate school dropout. This article is part of a project aimed at mitigating dropout in programming courses in Brazilian higher education, utilizing sentiment analysis in conjunction 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. This study focuses on distinguishing ambiguous sentiment with a phrase-to-phrase refinement approach. Results show that the approach improves ambiguous classification from 6% to 84% in the best case.

KEYWORDS

Sentiment Analysis, Natural Language Processing, Large Language Models

1 INTRODUÇÃO

A análise de sentimentos é o estudo computacional das opiniões, atitudes e emoções das pessoas em relação a uma entidade, que pode representar indivíduos, eventos ou tópicos, e normalmente está associada a uma opinião expressa em uma revisão sobre a mesma [1]. No contexto estudantil, a análise de sentimentos tem sido vista como um potencial de aplicação em diversos contextos, seja para capturar a satisfação ou o *feedback* dos estudantes sobre um determinado curso [2, 3], seja para compreender as dificuldades dos estudantes [4], ou até para mitigar potenciais casos de desistência ou evasão [5].

Este trabalho é parte de um projeto de pesquisa com a finalidade de investigar o uso de informações relacionadas ao sentimento de estudantes, visando o uso de ferramentas automatizadas para apoio à permanência estudantil. A pesquisa utiliza a análise de sentimentos coletados ativamente dos estudantes, combinando métodos psicopedagógicos e técnicas automatizadas baseadas em Processamento de Linguagem Natural (PLN). O objetivo central é identificar sentimentos negativos que possam indicar potenciais desistências, permitindo a adoção de intervenções para melhorar a permanência dos alunos.

Há diferentes opções na literatura para a extração de sentimentos por meio de PLN. As formas mais consolidadas consistem na aplicação de analisadores léxicos e de modelos supervisionados de classificação treinados com bases públicas de textos anotados de sentimentos (por exemplo, com bases de avaliações sobre produtos e filmes, como é o caso do IMDb) [6]. Por outro lado, o crescente interesse nas ferramentas de inteligência artificial generativas, que utilizam Large Language Models (LLMs) para a construção de conhecimento [7], tem instigado pesquisadores de diversas áreas a aplicar essas ferramentas na extração de sentimentos [8].

O presente artigo contribui para a área de análise de sentimentos ao comparar ferramentas automatizadas de extração de sentimentos. Para tanto, três ferramentas foram comparadas: duas LLMs e um analisador léxico. Essas ferramentas foram testadas em uma base de 540 respostas coletadas de estudantes de turmas introdutórias de programação. Os resultados das ferramentas foram comparados com uma classificação manual realizada com apoio de uma equipe de psicopedagogas.

Em trabalho anterior [9], observou-se que a maioria dos erros de classificação ocorreu quando as respostas dos estudantes continham tanto sentimentos negativos quanto positivos, ou seja, ambíguos. Assim, o presente trabalho propõe um método baseado em uma análise frase a frase dos sentimentos, com refinamento de sentimentos ambíguos. Os resultados mostram que a abordagem refinada aumentou sensivelmente a taxa de acerto dos sentimentos ambíguos em todas as ferramentas; no melhor caso, obteve-se um aumento de 6% para 84% com o analisador léxico.

O restante deste artigo está organizado da seguinte forma. Na Seção 2, é realizada uma fundamentação teórica e discutidos os trabalhos relacionados. Já na Seção 3, é apresentada a metodologia utilizada. Os resultados são apresentados e discutidos na Seção 4. Por fim, as conclusões deste esforço de pesquisa são apresentadas na Seção 5.

2 FUNDAMENTAÇÃO TEÓRICA

A análise de sentimentos é uma área fundamental do Processamento de Linguagem Natural (PLN), utilizada para extrair opiniões e emoções de textos. Tradicionalmente, essa tarefa era realizada

por meio de técnicas de análise léxica, que utilizam dicionários léxicos predefinidos para atribuir polaridades específicas (positiva, negativa, neutra) às palavras de um texto. Apesar de serem computacionalmente eficientes e facilmente interpretáveis, essas abordagens enfrentam limitações em capturar nuances de contexto e significado, especialmente em textos que usam gírias, sarcasmo ou expressões idiomáticas [10].

Com o advento dos Large Language Models (LLMs), o cenário de análise de sentimentos ganhou uma nova perspectiva. Os LLMs foram desenvolvidos com base em redes neurais profundas treinadas em enormes quantidades de dados, permitindo-lhes entender e gerar linguagem de forma coerente e contextualizada. Esses modelos superam as limitações da análise léxica, pois capturam nuances complexas e contextos mais amplos das expressões linguísticas. Estudos demonstram que os LLMs apresentam um desempenho superior, especialmente em contextos desafiadores, como na identificação de sarcasmo e ironia [10].

Entretanto, o uso de LLMs não é isento de desafios. Eles exigem recursos computacionais significativos para treinamento e inferência, o que pode ser um obstáculo para algumas aplicações [11]. Além disso, os LLMs são frequentemente criticados por sua natureza de "caixa-preta", dificultando a compreensão de como certas conclusões são alcançadas, o que é uma preocupação em aplicações que requerem interpretabilidade clara [12].

A fundamentação teórica deste artigo aborda aspectos relacionados com os LLMs e a análise léxica, com foco nos recursos empregados neste trabalho. Também são apresentados trabalhos relacionados à proposta.

2.1 Large Language Models

De maneira simplificada, LLMs são modelos desenvolvidos em duas etapas principais: a) uma etapa de pré-treinamento na qual o modelo é treinado em grande escala usando tarefas simples como previsão da próxima palavra ou legendagem; b) uma etapa de pós-treinamento em que o modelo é ajustado para seguir instruções, alinhar-se com as preferências humanas e melhorar capacidades específicas (por exemplo, programação e raciocínio) [13].

LLMs são considerados um tipo de modelo de aprendizagem de máquina projetado para o PLN. Emergem como sistemas de Inteligência Artificial capazes de processar e gerar texto com coerência e de generalizar para múltiplas tarefas de PLN [14, 15]. Há uma lista considerável de LLMs, desde GPT-1 (Generative Pre-trained Transformer), lançado em 2018, até Llama4, lançado em abril de 2025. Para este trabalho, dois LLMs de código aberto foram utilizados, o Llama e o Gemma.

Llama (Large Language Model Meta AI) é uma família de LLMs lançada em fevereiro de 2023 pela empresa Meta [16]. Dessa família, o modelo Llama 3, utilizado neste trabalho, foi desenvolvido em duas etapas principais: pré-treinamento e pós-treinamento. Na primeira etapa, um modelo com 405B (bilhões) de parâmetros em 15,6T de tokens em um contexto de 8K tokens é utilizado. Esta fase padrão de pré-treinamento é contínua, aumentando a janela de contexto suportada para 128K tokens. Na segunda etapa, foram adicionadas melhorias, incluindo capacidades de programação e raciocínio. Além disso, codificadores para imagens e para fala foram treinados separadamente e incorporados ao modelo em 8 idiomas diferentes, incluindo o português [13].

De uma família open-source de pequenos modelos de linguagem (small language models), Gemma é desenvolvida e mantida pela Google DeepMind [17]. Baseada nos modelos Gemini [18], Gemma foi lançada em fevereiro de 2024 em dois tamanhos: um modelo de 7 bilhões de parâmetros para implantação e desenvolvimento em GPU e TPU e outro de 2 bilhões de parâmetros para CPU e aplicações em dispositivos. Tais modelos foram treinados com 8192 tokens, utilizando um modelo de arquitetura baseado em *transformer decoders* [19].

2.2 Analisador léxico

A análise léxica é o processo de converter uma sequência de caracteres em uma sequência de tokens com um significado atribuído e identificado. Esse processo visa manipular o léxico, o qual é composto por palavras que armazenam seus significados e categorias lexicais (semânticas ou sintáticas). No caso de uma língua natural, essas categorias incluem substantivos, verbos, adjetivos, pontuações, etc [20, 21].

Há um conjunto considerável de ferramentas de análise léxica, utilizadas para diferentes propósitos: desenvolvimento de interpretadores e compiladores, validação de dados, segurança, linguagens de domínio específico (domain-specific languages - DSLs), entre outros. Neste trabalho, a ferramenta LeIA (Léxico para Inferência Adaptada) é um fork do léxico e da ferramenta para análise de sentimentos VADER (Valence Aware Dictionary and sEntiment Reasoner), adaptados para textos em português, já que a LeIA é uma cópia independente do projeto VADER e também uma contribuição desse projeto [22, 23].

2.3 Trabalhos relacionados

A análise de sentimentos é realizada em diversos domínios de aplicação. Em [24, 25], o objetivo é compreender melhor o sentimento público em debates políticos. No contexto educacional, em [4], apresenta-se um estudo sobre as emoções de estudantes durante a realização de atividades de programação, com observações qualitativas.

Para [26], foi utilizada a arquitetura SASys, que utiliza a análise de sentimentos para identificar o perfil emocional dos alunos e detectar quais deles têm risco de evasão. Em um trabalho mais recente dos mesmos autores, [5], 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 [27].

Um trabalho bastante próximo a este artigo é apresentado em [8], no qual uma extensa 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. Apesar de alguns resultados semelhantes, como a dificuldade das ferramentas em extrair sentimentos de respostas

com múltiplos sentimentos, o presente artigo se diferencia do artigo de [8] por ser direcionado ao contexto estudantil e por propor o refinamento da classificação de sentimentos ambíguos.

3 METODOLOGIA

Os dados utilizados neste trabalho foram obtidos por meio da coleta ativa de sentimentos de alunos de turmas introdutórias de programação de computadores. A coleta ativa de sentimentos foi realizada por meio de um questionário com perguntas abertas, de modo que os estudantes elaborassem suas respostas livremente, formando frases. A coleta foi realizada de forma anônima, visando reduzir a inibição dos estudantes para expressar críticas e insatisfações.

Os questionários foram aplicados antes da primeira avaliação e após uma ou mais avaliações das disciplinas e compreenderam os primeiros e segundos semestres de 2023. A coleta foi realizada por meio de um formulário online, com duas perguntas: (1) Como você está se sentindo em relação à disciplina? e (2) Como você está se sentindo em relação à universidade?

A coleta de dados com os participantes humanos foi realizada de forma anônima. Este procedimento está em conformidade com as diretrizes éticas brasileiras. Especificamente, a Resolução CNS nº 510/2016, em seu Art. 1º (parágrafo único, inciso I), isenta de registro e avaliação pelo Sistema CEP/CONEP as pesquisas que utilizam 'participantes não identificados' [28], como é o caso deste estudo. Após a coleta, cada resposta recebeu uma classificação psicopedagógica, realizada manualmente 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 [29], que é uma análise qualitativa do discurso manifesto da comunicação. Em caso de divergência na classificação, foi utilizada a observação de outra pessoa pesquisadora, sem formação pedagógica.

As respostas sobre o sentimento em relação à universidade e à disciplina foram consolidadas para a análise. 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;
- *Neutro:* respostas sem identificação de informação polarizada ou respostas que, a depender do contexto, poderiam ser vistas como positivas ou negativas.

A classificação como *Ambos* mostrou-se necessária, uma vez que algumas respostas são compostas por diversas frases, podendo indicar tanto aspectos positivos quanto negativos dos itens avaliados. Cabe uma observação: as 540 respostas analisadas são compostas por 771 frases, sendo que 135 delas têm pelo menos duas frases. As classificações como *Neutro* foram posteriormente removidas neste trabalho.

Na Tabela 1, é apresentada a distribuição dos 540 sentimentos classificados manualmente e sua estratificação com sentimentos escritos em uma única frase ou com múltiplas frases.

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

- Llama: versão Meta Llama 3 Instruct I 8B - modelo com 8 bilhões de parâmetros (o menor disponível).
- Gemma: versão Gemma 2 9B IT - modelo com 9 bilhões de parâmetros.
- LeIA[1] (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*). Embora tenha foco na análise de sentimentos em textos de mídias sociais, essa ferramenta é funcional para textos de outros domínios.

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. Para isso, textos classificados manualmente como neutros foram removidos da análise (eram apenas 6 no total).

Para as LLMs, o seguinte *prompt* foi utilizado: "Analise o sentimento da frase e responda apenas com POSITIVO, NEGATIVO ou AMBOS". Também foi fornecida a temperatura de respostas com o valor zero[2].

Em relação aos retornos obtidos pelas LLMs, um ajuste manual foi aplicado. Mesmo utilizando a temperatura zero, algumas respostas foram diferentes das solicitadas. Por exemplo, algumas respostas foram "NEUTRA", "NEUTRAL", "*POSITIVO*" e "**ANÁLISE DE SENTIMENTO:** NEGATIVO". Para algumas, também foi gerada uma explicação da classificação do sentimento. O modelo Llama não conseguiu realizar a classificação de dois textos, ambos eram respostas com apenas um "Ok", sendo atribuída a classificação como *Ambos*. Os principais problemas desse tipo foram encontrados nas respostas do Llama.

3.1 Análises realizadas

A primeira etapa de análise consistiu na avaliação dos 540 sentimentos com os modelos indicados anteriormente. Como se vê nos resultados, a identificação da classe *Ambos* mostrou-se desafiadora para todos os modelos utilizados. Uma segunda etapa de análises foi realizada, na qual os sentimentos avaliados manualmente ou por alguma das outras técnicas e que indicavam o sentimento como *Ambos*, foram removidos. Um total de 284 registros resultou dessa filtragem. Na segunda análise, verificou-se o grau de sucesso dos modelos em identificar as duas polaridades de sentimento.

Como terceira análise, os sentimentos foram divididos em frases. O critério foi a utilização do ponto como separador. A segunda etapa de análise mostrou um desempenho aceitável dos modelos na classificação de sentimentos positivos e negativos; portanto, a hipótese é que uma boa classificação de sentimentos rotulados como *Ambos* poderia ser obtida a partir da análise separada das frases.

[1] Site do projeto do LeIA: https://github.com/rafjaa/LeIA
[2] Utiliza-se uma temperatura baixa para que a LLM forneça respostas mais determinísticas, o que é útil quando se deseja consistência e precisão.

Tabela 1: Distribuição das respostas nas classes definidas para o trabalho

Estratificação	Positivo	Negativo	Ambos
Total de sentimentos	54,3%	24,0%	21,7%
Sentimentos em frase únicas	66,4%	20,7%	12,8%
Sentimentos escritos em duas ou mais frases	41,0%	31,4%	27,6%

4 RESULTADOS

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

4.1 Primeira etapa de análise

Os resultados dessa análise foram apresentados anteriormente em [9], sendo aqui apresentada uma síntese.

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

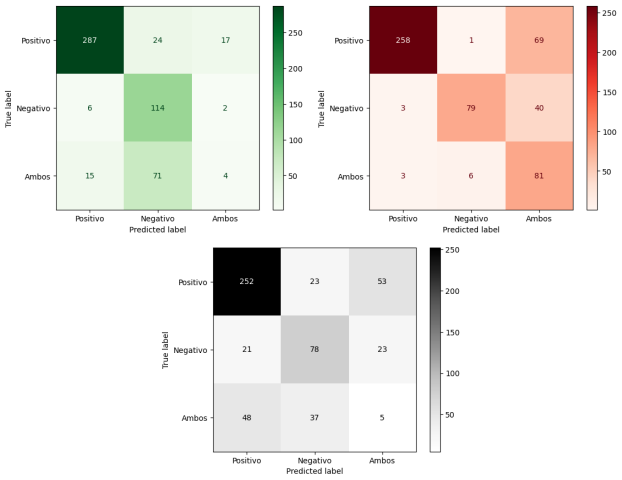

Figura 1: Matrizes de confusão da classificação dos sentimentos nos modelos avaliados

Utilizando a métrica de F1-score devido ao desbalanceamento das classes, os 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 7%, Gemma 58% e LeIA apenas 6%. O Llama teve desempenho similar na classificação de *Ambos* tanto nos sentimentos negativos quanto nos sentimentos positivos, enquanto o Gemma apresentou uma redução significativa. Quanto ao LeIA, sua classificação para a categoria *Ambos* é insignificante. Na avaliação da classe *Ambos*, fica evidente a dificuldade dos modelos em classificar os sentimentos nessa categoria, o que é objeto de melhoria.

Considerando que muitos dos sentimentos foram expressos por múltiplas frases (135 de 540, ou 25% dos sentimentos), um dos possíveis motivos para a dificuldade da classificação pode estar relacionado à quantidade de sentimentos expressos, conforme já discutido e apresentado na Tabela 1.

A Figura 2 apresenta as matrizes de confusão e o resultado da classificação dos sentimentos expressos em apenas uma frase. Existe uma manutenção na acurácia dos modelos (pouca variação), sendo que o Llama obtém 77,5%, Gemma 76,5% e LeIA 62,9%. Mesmo assim, a classificação para *Ambos* ainda é ruim, sendo obtido F1-score de 5,7%, 51,3% e 4,6% para Llama, Gemma e LeIA, respectivamente.

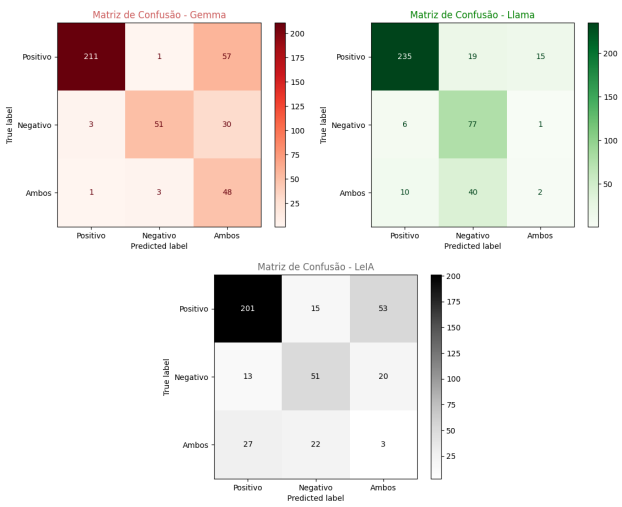

Figura 2: Matrizes de confusão da classificação sentimentos escrito em apenas uma frase

4.2 Segunda etapa de análise

Uma vez que a classificação das respostas com sentimento *Ambos* apresentou baixa taxa de acertos em todos os modelos avaliados, os dados foram reanalisados considerando apenas os sentimentos

positivos e negativos e suas respectivas classificações pelos modelos. As matrizes de confusão dessa análise são mostradas na Figura 3[3].

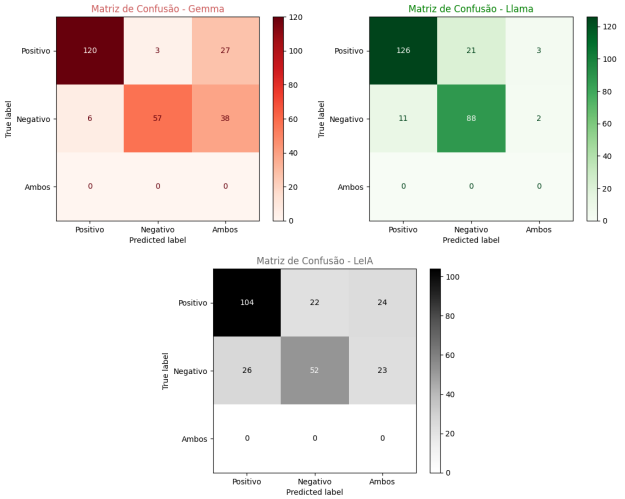

Figura 3: Matrizes de confusão da classificação apenas dos sentimentos classificados manualmente como positivos e negativos

Ao computar o F1-score para os resultados da Figura 3, Llama obteve acurácia de 85,2% (F1-score de 87,8% para Positivo e 83,8% para Negativo), enquanto Gemma obteve 70,5% (F1-score de 86,9% para Positivo e 70,8% para Negativo) e a LeIA 62,1% (F1-score de 74,3% para Positivo e 59,4% para Negativo). Essa segunda análise mostra uma melhora na acurácia das classificações do Llama, uma manutenção na classificação obtida com o LeIA e piora do Gemma, mas todas com resultados superiores à classificação como *Ambos* obtidas na análise anterior. Por esse motivo, tem-se a realização da terceira análise, que considera as frases individualmente.

4.3 Terceira etapa de análise

É importante considerar que a classificação das psicopedagogas analisou os sentimentos considerando as respostas das múltiplas frases e não frase-a-frase. Nesse sentido, a terceira forma de análise considerou apenas os sentimentos rotulados como *Ambos* pelas psicopedagogas e somente os sentimentos compostos de múltiplas frases. Nesse sentido, um total de 115 frases (14,9% do total), de 38 sentimentos (7% do total) estão presentes nessa forma de análise.

Adotou-se os seguintes critérios para a classificação das frases como um sentimento *Ambos*: (C1) se o sentimento possui apenas frases rotuladas como *Ambos*; (C2) se o sentimento é escrito com frases Positivas e Negativas (ocorrência dos dois tipos, independentemente da quantidade); e (C3) a ocorrência de alguma das polaridades e *Ambos*, desde que a quantidade *Ambos* supere a quantidade da polaridade. Caso esses critérios não sejam atendidos, a polaridade é respeitada. A Tabela 2 ilustra sentimentos cuja composição de suas frases determinam um dos critérios apresentados.

[3]As matrizes de confusão podem parecer estranhas pelas últimas linhas estarem zeradas, mas isso é esperado, uma vez que os registros de sentimentos como *Ambos* foram removidos (classificação das psicopedagogas), embora as LLMs indiquem essa classificação quando aplicadas.

Na Figura 4 é apresentado o resultado da classificação dos sentimentos de múltiplas frases rotuladas pelas psicopedagogas como *Ambos* e utilizando os critérios elaborados neste trabalho. A partir dessa estratégia, os seguintes F1-score para essa classificação são obtidos: 84,8% para o Gemma, 68,4% para o Llama e 84,8% para o LeIA.

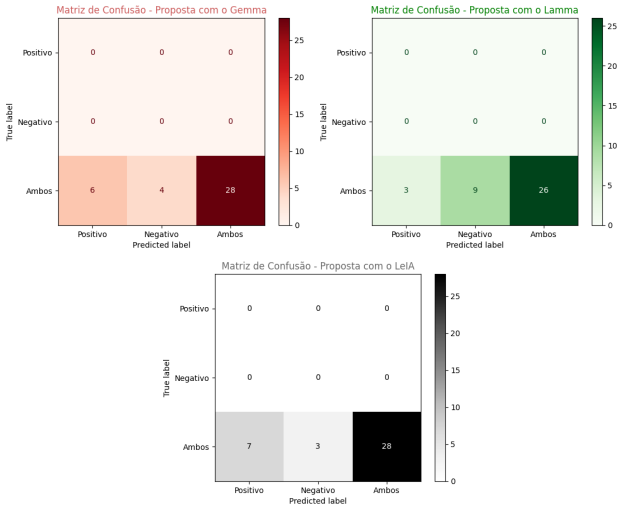

Figura 4: Matrizes de confusão da classificação dos sentimentos com múltiplas frases e rotulados inicialmente como *Ambos*

Com os resultados da análise, fica evidente a melhoria da classificação como *Ambos* em todos os modelos, destacando-se uma melhora significativa. Na Tabela 3, é apresentada, de forma simplificada, a comparação dos resultados com base na métrica F1-score.

4.4 Discussões

Apesar do estudo realizado neste trabalho apresentar resultados que se mostram promissores, existem alguns limites do estudo a serem discutidos, principalmente associados à subjetividade. A análise de sentimentos empírica realizada pelas profissionais da área psicopedagógica depende da sua observação sobre o significado da frase. Ainda, a classificação desses sentimentos em categorias pode implicar viés, uma vez que alguém poderia considerar que o sentimento *ansioso* pode significar tanto muita insatisfação quanto satisfação (a depender do contexto da frase). A mesma subjetividade está presente na classificação por parte das técnicas baseadas em PLN.

Embora não tenha sido avaliada formalmente, destaca-se a questão de desempenho observada ao executar os modelos de classificação. O tempo de processamento do LeIA é muito menor do que o dos outros modelos. O tempo de processamento de todas as respostas analisadas com o LeIA é inferior ao de uma das respostas analisadas com a Llama (LLM que processou mais rapidamente do que o Gemma). A escolha de um modelo deve atender à assertividade, mas também aos recursos computacionais disponíveis. Para uma aplicação que precise classificar em tempo real, o uso do LeIA é

Tabela 2: Tabela exemplificativa para a classificação dos sentimentos

Sentimento	Polaridade	Critério
O professor é bom, mas a disciplina é difícil	Ambos	C1
Não gosto da disciplina, mas meus amigos gostam	Ambos	
O professor é bom	Positiva	C2
A disciplina é difícil	Negativa	
O professor é bom	Positiva	C3
Não gosto da disciplina, mas meus amigos gostam	Ambos	
Tenho dificuldade com o conteúdo	Negativa	C3
O professor é bom, mas a disciplina é difícil	Ambos	
Não gosto da disciplina, mas meus amigos gostam	Ambos	

Tabela 3: Comparação dos resultados da classificação dos sentimentos rotulados como *Ambos*

	Llama	Gemma	LeIA
Inicial	7%	58%	6%
Proposto	68%	84%	84%

uma opção com acerto satisfatório tanto em positivos quanto em negativos.

5 DISPONIBILIZAÇÃO DOS DADOS

Os arquivos com os sentimentos coletados e classificações em cada um dos modelos estão disponíveis em https://doi.org/10.5281/zenodo.15521409.

6 CONCLUSÕES

A metodologia envolveu a coleta de 540 respostas abertas de estudantes, que foram classificadas manualmente por psicopedagogas em categorias de sentimento: Positivo, Negativo, Ambos (para respostas com polaridades mistas) e Neutro (posteriormente removido da análise). As ferramentas automatizadas foram então aplicadas a esses dados.

Inicialmente, os resultados da classificação geral revelaram uma dificuldade considerável das ferramentas para identificar corretamente sentimentos ambíguos. Enquanto o Llama obteve uma acurácia de 75% e o Gemma de 77% (com o LeIA em 62%), a classificação da categoria "Ambos" foi particularmente baixa: 7% para Llama, 58% para Gemma e apenas 6% para LeIA. Essa lacuna motivou a proposição de uma abordagem de refinamento frase a frase.

A principal contribuição deste trabalho reside na implementação e avaliação de um método baseado na análise individualizada das frases que compõem as respostas dos estudantes. Para os sentimentos inicialmente rotulados como "Ambos" pelas psicopedagogas, foi aplicada uma estratégia que considera a ocorrência de frases positivas e negativas em uma mesma resposta. Essa abordagem de refinamento demonstrou uma melhoria substancial na classificação de sentimentos ambíguos. Os resultados da análise refinada são notáveis: a taxa de acerto na classificação "Ambos" aumentou dramaticamente. O Llama, que inicialmente apresentava 7%, saltou para 68%; o Gemma, de 58%, alcançou 84%; e o LeIA, que partiu de 6%, também atingiu 84%. Essa melhoria significativa valida a eficácia da abordagem frase a frase para lidar com a complexidade dos sentimentos ambíguos em textos educacionais.

Em termos de discussões, o estudo ressalta a subjetividade inerente à análise de sentimentos, tanto manual quanto automatizada, e a importância de considerar os recursos computacionais. Embora as LLMs ofereçam maior capacidade de compreensão contextual, o analisador léxico LeIA demonstrou um desempenho computacional superior, sendo uma opção viável para aplicações que exigem processamento em tempo real, mesmo com uma acurácia ligeiramente inferior em alguns cenários.

Em conclusão, este artigo demonstra que a aplicação de uma abordagem de refinamento frase a frase é fundamental para aprimorar a precisão da análise de sentimentos em contextos educacionais, especialmente ao lidar com a complexidade das respostas ambíguas dos estudantes. Os achados fornecem insights valiosos para o desenvolvimento de ferramentas mais robustas e eficazes no apoio à permanência estudantil, ao mesmo tempo em que destacam a necessidade de equilibrar a acurácia com a eficiência computacional na escolha das ferramentas de PLN.

REFERÊNCIAS

[1] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113, 2014.
[2] Sujata Rani and Parteek Kumar. A sentiment analysis system to improve teaching and learning. *Computer*, 50(5):36–43, 2017. doi: 10.1109/MC.2017.133.
[3] Marion Neumann and Robin Linzmayer. 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, 2021.
[4] Zahra Atiq and Michael C Loui. A qualitative study of emotions experienced by first-year engineering students during programming tasks. *ACM Transactions on Computing Education (TOCE)*, 22(3):1–26, 2022.
[5] Míria LDR Bóbó, Fernanda Campos, Victor Stroele, José Maria N David, Regina Braga, and Tiago Timponi Torrent. 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, 2022.
[6] Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780, 2022.
[7] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024.
[8] Nimra Mughal, Ghulam Mujtaba, Sarang Shaikh, Aveenash Kumar, and Sher Muhammad Daudpota. Comparative analysis of deep natural networks and large language models for aspect-based sentiment analysis. *IEEE Access*, 12: 60943–60959, 2024. doi: 10.1109/ACCESS.2024.3386969.
[9] Benjamin Moreira, Luiz Camargo, Ricardo Pfitscher, and Tatiana Garcia. Comparação de ferramentas para análise de sentimentos aplicada no contexto educacional. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 474–478, Porto Alegre, RS, Brasil, 2024. SBC. doi:

10.5753/stil.2024.245409. URL https://sol.sbc.org.br/index.php/stil/article/view/31165.

[10] John Astera. Comparing lexical analysis and large language models in sentiment analysis, 2023. URL https://www.example.com/astera2023sentiment. Accessed: 2023-05-19.

[11] OpenAI. Gpt-4 technical report, 2023. URL https://openai.com/research/gpt-4. Accessed: 2023-05-19.

[12] Alice Gemini. Interpretability challenges in large language models, 2023. URL https://www.example.com/gemini2023interpretability. Accessed: 2023-05-19.

[13] Aaron Grattafiori and Llama Team. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

[14] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models, 2024. URL https://arxiv.org/abs/2307.06435.

[15] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1 (8):9, 2019.

[16] Meta. Llama 3.1, 2024. URL https://ai.facebook.com/llama.

[17] Team Google DeepMind. Gemma docs. url=https://ai.google.dev/gemma/docs, 2025. [Online, accessado em Maio de 2025].

[18] Rohan Anil and Gemini Team. Gemini: A family of highly capable multimodal models, 2024. URL https://arxiv.org/abs/2312.11805.

[19] Thomas Mesnard and Gemma Team. Gemma: Open models based on gemini research and technology, 2024. URL https://arxiv.org/abs/2403.08295.

[20] Yao Fu, Hao Zhou, Jiaze Chen, and Lei Li. Rethinking text attribute transfer: A lexical analysis, 2019. URL https://arxiv.org/abs/1909.12335.

[21] Pai T. Vaikunta, Devi A. Jayanthila, and Aithal P. S. A systematic literature review of lexical analyzer implementation techniques in compiler design. *International Journal of Applied Engineering and Management Letters (IJAEML)*, 4(2):285–301, 2020.

[22] Rafael J. A. Almeida. Leia - léxico para inferência adaptada. https://github.com/rafjaa/LeIA, 2018.

[23] CJ Hutto Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf*, 2014.

[24] Lucas Lazarini, Fábio S Igarashi Anno, Eloize R Marques Seno, and Helena M Caseli. 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, 2023.

[25] Eloize R Marques Seno, Fábio S Igarashi Anno, Lucas Lazarini, and Helena M Caseli. 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, 2023.

[26] Míria Bobó, Fernanda Campos, Victor Stroele, José David, and Regina Braga. Identificação do perfil emocional do aluno através de análise de sentimento: Combatendo a evasão escolar. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 30, page 1431, 2019.

[27] Mayela Coto, Sonia Mora, Beatriz Grass, and Juan Murillo-Morera. Emotions and programming learning: systematic mapping. *Computer Science Education*, 32(1):30–65, 2022.

[28] Conselho Nacional de Saúde CNS. ResoluÇÃo nº 510„ 2016. URL https://www.gov.br/conselho-nacional-de-saude/pt-br/atos-normativos/resolucoes/2016/resolucao-no-510.pdf/view.

[29] Lawrence Bardin. *Análise de conteúdo*. Lisboa: edições 70, 1977.