15º Congresso de Inovação, Ciência e Tecnologia do IFSP - 2024

Analisando a eficácia de modelos de linguagem em larga escala na detecção de aspectos em avaliações de restaurantes

Fabiano M. Rocha Junior[1], Luiz H. N. Silva[2], Eloize R. M. Seno[3]

[1]Graduando em Bach. em Eng. de Software, Bolsista PIBIC, IFSP São Carlos, fabiano.j@aluno.ifsp.edu.br
[2]Graduando em Bach. em Eng. de Software, Bolsista PIBIFSP, IFSP São Carlos, nascimento.henrique1@aluno.ifsp.edu.br
[3]Orientadora, IFSP São Carlos, eloize@ifsp.edu.br

Área de conhecimento (Tabela CNPq): 1.03.03.00-6 Metodologias e Técnicas da Computação.

RESUMO: A identificação de aspectos é uma fase essencial da Análise de Sentimentos Baseada em Aspectos, cujo objetivo é detectar os alvos de opinião em avaliações de produtos ou serviços publicados nas mídias sociais. Embora exista alguns trabalhos dedicados à detecção de aspectos em textos opinativos em português, os modelos de linguagem em larga escala (LLMs) têm sido pouco explorados. Dado esse contexto, esta pesquisa investigou a eficácia de LLMs ao identificar aspectos explícitos e implícitos em avaliações de restaurantes e comparou os resultados obtidos com experimentos preliminares que avaliaram o uso de LLMs ao lidar com a identificação de aspectos explícitos e implícitos de forma independente. Estudos mostraram um ganho de até 30 pontos percentuais em termos de F-*score* na detecção de aspectos explícitos ao abordar a identificação dos dois tipos de aspectos em uma mesma tarefa. Os experimentos também mostraram uma dificuldade maior dos modelos em lidar com os aspectos implícitos do que com os aspectos explícitos.

PALAVRAS-CHAVE: análise de sentimentos baseada em aspectos; detecção de aspectos; LLMs; revisões de restaurantes.

Analyzing the effectiveness of large-scale language models in aspect detection in restaurant reviews

ABSTRACT: Aspect detection is an essential phase in Aspect-Based Sentiment Analysis, which aims detecting the opinion target aspects in product or service reviews published on social media. Although there are some studies dedicated to aspect detection in opinionated texts in Portuguese, large language models (LLMs) have been little explored. Given this context, this research investigated the effectiveness of LLMs in identifying explicit and implicit aspects in restaurant reviews and compared the results with preliminary experiments that assessed the use of LLMs in independently handling the identification of explicit and implicit aspects. Studies showed an improvement of up to 30 percentage points in F-score for detecting explicit aspects when addressing the detection of both types of aspects in a single task. The experiments also showed that the models have a greater difficulty dealing with the implicit aspects than with the explicit ones.

KEYWORDS: aspect-based sentiment analysis; aspect detection; LLM; restaurant reviews.

INTRODUÇÃO

A análise de sentimentos (AS) é uma subárea do processamento de linguagem natural (PLN) que visa analisar opiniões expressas em textos, como avaliações (comentários) sobre produtos e serviços publicados *online*. Dentro desse campo, a análise de sentimentos baseada em aspectos (ASBA) é um tipo de AS mais refinada que busca analisar o sentimento ou opinião relacionados a aspectos (atributos) específicos de uma entidade, produto ou serviço. Por exemplo, na avaliação "O risoto estava muito saboroso, mas chegou frio." sobre um restaurante tem-se o aspecto explícito "risoto" com sentimento positivo e o aspecto implícito "temperatura" com sentimento negativo. Esse tipo de análise é particularmente útil em contextos onde a opinião sobre diferentes aspectos de uma entidade pode variar significativamente. Numa avaliação de um restaurante, por exemplo, um cliente pode estar satisfeito com a comida, mas insatisfeito com o atendimento. A ASBA proporciona uma visão mais profunda e precisa das percepções públicas, permitindo um melhor aproveitamento das informações contidas em suas avaliações pelas partes interessadas.

Um dos maiores desafios da ASBA consiste em identificar os aspectos alvos de opinião em uma avaliação. Na literatura há abordagens de detecção de aspectos baseadas em padrões morfossintáticos (CATHARIN; FELTRIM, 2018), modelo discursivo (CATHARIN; FELTRIM, 2018), regras de linguagem (SAIAS; MOURãO; OLIVEIRA, 2018; MACHADO et al., 2021), léxicos (COSTA; PARDO, 2020), em aprendizado de máquina tradicional (BALAGE, 2017; VARGAS; PARDO, 2018; MACHADO; PARDO, 2022) e em modelos de linguagem pré-treinados (LOPES; CORREA; FREITAS, 2021; ASSI et al., 2022).

Em um experimento preliminar investigamos o uso de vários modelos de linguagem em larga escala (no inglês, *Large Language Models - LLM*) na detecção de aspectos em avaliações de restaurantes (SILVA et al., 2024). As análises foram realizadas considerando a detecção de aspectos explícitos e implícitos como tarefas independentes, ou seja, usando *prompts* específicos para cada tipo de aspecto. Com o intuito de analisar a influência do *prompt* na identificação de aspectos, neste artigo analisamos a eficácia dos LLMs ao considerar a identificação de aspectos explícitos e implícitos em uma mesma tarefa e comparamos com os resultados reportados por (SILVA et al., 2024).

MATERIAIS E MÉTODOS

Para os experimentos realizados nesta pesquisa foram usados os cinco LLMs investigados por (SILVA et al., 2024), para que se pudesse comparar os resultados dos modelos ao considerar a detecção de aspectos explícitos e implícitos em uma mesma tarefa e como tarefas independentes. Dos cinco modelos, três são multilíngues – GPT-3.5 Turbo, o GPT-4o e o GPT-4o mini[1] – e dois monolíngues treinados para o português – Sabiá-2-medium e Sabiá-3[2]. Em experimentos reportados por (ALMEIDA et al., 2024), o Sabiá-2-medium, lançado em março de 2024, é comparado a vários outros LLMs, alcançando desempenho igual ou melhor que GPT-3.5 Turbo e Gemini 1.0 Pro em 96,9% dos exames analisados. O Sabiá-3, por sua vez, foi lançado no final de julho de 2024 e, até o momento da escrita deste artigo, não houve divulgação de resultados sobre o modelo. Segundo informações

[1]Fornecidos pela OpenAI por meio da API do ChatGPT (<https://openai.com/api>)
[2]Disponibilizados pela Maritaca AI por meio da MariTalk API (<https://www.maritaca.ai/>)

da OpenAI[3], o GPT-4o ('o' de omini), lançado em maio de 2024, é o seu modelo mais avançado e inteligente para tarefas mais complexas. O GPT-4o mini, lançado em junho de 2024, é o modelo mais avançado na categoria de modelos pequenos (é multimodal), que também inclui o GPT-3.5 Turbo, um dos primeiros modelos disponibilizados pela empresa em setembro de 2022.

Os LLMs investigados são modelos generativos baseados em *prompt*, que recebem como entrada um texto (*prompt*) com a descrição da tarefa a ser realizada e geram as saídas conforme solicitado. O maior desafio em lidar com esses modelos consiste em definir um *prompt* que gere as saídas exatamente como se espera para a tarefa, uma vez que esses modelos estão sujeitos a comportamentos diferentes para *prompts* distintos. Nos experimentos reportados por (SILVA et al., 2024) foram adotados *prompts* específicos para cada tipo de aspecto. Diferentemente daquele trabalho, neste artigo a detecção de aspectos explícitos e implícitos é abordada em um mesmo *prompt*, apresentado a seguir:

Prompt: *Dada a sentença EXEMPLO com os aspectos alvos de opinião, responda no formato [alvo] o(s) alvo(s) de opinião explícito(s) e implícito(s) na sentença, quando existir(em). EXEMPLO: "A comida estava deliciosa, mas achei cara." SAÍDA: [exp - comida] [imp - preço]*

A fim de se obter modelos mais determinísticos e reduzir a possibilidade de alucinações dos modelos, o parâmetro temperatura de todos os LLMs foi ajustado em zero, conforme sugerido por outros trabalhos da literatura (OLIVEIRA et al., 2023; SANTOS; PARABONI, 2023).

As avaliações foram feitas usando o mesmo *corpus*[4] de avaliação usado por (SILVA et al., 2024), composto de 374 sentenças opinativas sobre diferentes restaurantes brasileiros. Esse *corpus* contém no total 432 aspectos anotados manualmente, sendo 88,6% deles explícitos e 11,4% implícitos. A Tabela 1 apresenta exemplos de sentenças e os respectivos aspectos explícitos (em negrito) e implícitos do *corpus*.

Tabela 1: Exemplos de anotação de aspectos explícitos (em negrito) e implícitos.

Sentença	**Implícito**
O **baião de dois** é caprichado, a **carne de sol** é saborosa, o **dadinho de tapioca** não tem como ser ruim, o **caldo de mocofava** é rico, equilibrado, o **atolado de cupim** é gostoso, há ampla **oferta de cachaças**, o **café** é correto.	–
O **bacalhau grelhado**, o prato mais pedido da casa, é muito delicado.	–
Não é demais lembrar: a casa só aceita dinheiro ou cheque – costume fora de moda, também trazido de outros tempos.	forma de pagamento
Bonus track: das caixas de som da casa sai uma agradável sequência de baião, xote, forró.	música

A seção a seguir apresenta os resultados alcançados por cada modelo.

RESULTADOS E DISCUSSÃO

A Tabela 2 apresenta os resultados obtidos em termos de Precisão (P), Cobertura (C) e F-*score* (F) por cada LLM na detecção de aspectos explícitos e implícitos. Os resultados são apresentados em termos de Precisão (P), Cobertura (C) e F-*score* (F). Seja c uma classe qualquer, a Precisão

[3]https://platform.openai.com/docs/models

[4]*Corpus* (corpora no plural) é um conjunto de textos em uma dada língua que serve de base de análises.

considera a quantidade de sentenças da classe *c* corretamente classificadas em relação ao total de sentenças classificadas como sendo da classe *c*. A Cobertura representa a quantidade de sentenças da classe *c* corretamente classificadas sobre o total de sentenças classificadas como classe *c* na anotação de referência. F-*score* representa a média harmônica entre a Precisão e a Cobertura. Macro-F, por sua vez, representa a média aritmética entre os valores de F-*score* das duas classes. Para fins de comparação, os resultados reportados por (SILVA et al., 2024) são reproduzidos na Tabela 3.

Tabela 2: Resultados obtidos na detecção de aspectos explícitos e implícitos usando um único *prompt*.

	Explícitos			**Implícitos**			
LLM	**Precisão**	**Cobertura**	**F-*score***	**Precisão**	**Cobertura**	**F-*score***	**Macro-F**
Sabiá-2-medium	**54,33%**	65,71%	**59,48%**	3,61%	19,23%	6,08%¨	**32,78%**
GPT-4o	42,13%	**69,37%**	52,42%	**6,16%**	**32,69%**	**10,37**%	31,40%
Sabiá-3	47,20%	61,78%	53,51%	1,10%	7,69%	1.92%	27,71%
GPT-4o mini	**52,70**%	**53,66**%	**53,18**%	0,78%	5,77%	1,37%	27,27%
GPT-3.5 turbo	35,38%	48,17%	40,80%	3,30%	5,77%	4,20%	22,50%

Tabela 3: Resultados para a detecção de aspectos explícitos e implícitos usando *prompts* independentes (SILVA et al., 2024).

	Explícitos			**Implícitos**		
LLM	**Precisão**	**Cobertura**	**F-*score***	**Precisão**	**Cobertura**	**F-*score***
Sabiá-2-medium	35,11%	77,75%	48,37%	1,60%	26,00%	3,01%
GPT-4o	21,51%	76,18%	33,54%	**3,90%**	**34,00%**	**7,00**%
Sabiá-3	33,21%	67,80%	44,58%	2,21%	32,00%	4,13%
GPT-4o mini	**40,90%**	**15,82%**	**22,81%**	2,23%	20,00%	4,01%
GPT-3.5 turbo	31,53%	76,18%	44,60%	1,95%	32,00%	3,68%

Ao comparar os resultados das Tabela 2 e 3, nota-se que todos os LLMs apresentaram um ganho de desempenho em termos de precisão e de F-*score*, na identificação de aspectos explícitos, ao considerar os dois tipos de aspectos em um único *prompt*. Contudo, com exceção do GPT-4o mini, todos os modelos tiveram uma diminuição nos valores de cobertura. O GPT-4o mini obteve uma melhora de cerca de 12 pontos percentuais na precisão (isto é, saltou de 40,90% para 52,70%) e um ganho de aproximadamente 38 pontos percentuais na cobertura (saltando de 15,82% foi para 53,66%). Por consequência, o modelo obteve um aumento de cerca de 30 pontos percentuais no F-*score*, saltando de 22,81% para 53,18%.

Com relação aos aspectos implícitos, apenas o Sabiá-2-medium e o GPT-4o obtiveram uma melhora no desempenho em termos de F-*score*. De modo similar ao observado por (SILVA et al., 2024), os resultados mostram uma grande dificuldade dos LLMs em identificar esse tipo de aspecto, sendo que o GPT-4o, o mais avançado para tarefas complexas, segundo a OpenAI, manteve-se como o modelo que apresentou o melhor desempenho. Segundo (SILVA et al., 2024), a dificuldade em identificar aspectos implícitos também foi relatada pelos humanos na anotação do *corpus* de avaliação de restaurantes.

Na comparação entre os resultados obtidos por cada LLM, apresentados na Tabela 2, o Sabiá-2-medium obteve o melhor desempenho geral em termos de Macro-F (i.e. 32,78%). O segundo melhor desempenho foi alcançado pelo GPT-4o, ou seja, 31,40% de Macro-F. Testes estatísticos devem ser realizados no futuro para comprovar se essa diferença é estatísticamente significativa.

No que diz respeito a identificação de aspectos explícitos, o melhor desempenho em termos de F-*score* (59,48%) e a melhor precisão (54,33%) foram obtidos pelo modelo monolíngue Sabiá-2-medium. Contudo, a maior cobertura (i.e. 69,37%) foi alcançada pelo modelo multilíngue GPT-4o.

CONCLUSÕES

Este artigo investigou a eficácia de cinco LLMs ao identificar aspectos explícitos e implícitos em uma mesma tarefa (usando o mesmo *prompt*) em avaliações de restaurantes em português. Nos experimentos, o LLM monolíngue Sabiá-2-medium, treinado para o português e disponibilizado pelo Maritaca AI, mostrou um potencial maior na detecção de aspectos explícitos, do que todos os modelos investigados, incluindo três modelos multilíngues da OpenAI como o GPT-3.5 turbo, GPT-4o e GPT-4o mini. Vale destacar que os modelos da família Maritaca AI, além de apresentarem resultados superiores ou similares em alguns casos aos obtidos pelos modelos da OpenAI, eles têm um custo bem menor do que o custo de uso do ChatGPT API[5].

Os experimentos também mostraram uma dificuldade maior de todos LLMs na identificação dos aspectos implícitos, isto é, que não são mencionados de forma explícita no texto, mas que podem ser inferidos pelo contexto. Essa dificuldade também foi apontada pelos humanos durante a anotação do *corpus* usado nos experimentos (SILVA et al., 2024).

Os resultados obtidos neste trabalho foram também comparados com os resultados de experimentos anteriores que avaliaram o uso dos mesmos LLMs investigados nesta pesquisa ao lidar com a detecção de aspectos explícitos e implícitos de forma independente, ou seja, tratando como duas tarefas distintas e executadas com *prompts* específicos (SILVA et al., 2024). Os resultados indicaram um aumento em termos de F-*score* de aproximadamente 30 pontos percentuais no melhor caso (alcançado pelo modelo GPT-4o mini) na detecção de aspectos explícitos ao abordar a identificação dos dois tipos de aspectos em uma mesma tarefa. Esses resultados comprovam que os LLMs estão sujeitos a apresentar comportamentos diferentes para diferentes prompts de entrada e sugerem a necessidade de uma cuidadosa engenharia de prompt para se alcançar um melhor resultado em determinada tarefa.

Como trabalhos futuros pretende-se investigar o uso de modelos pré-treinados e ajustados para a tarefa de análise de sentimentos em português, como o modelo de (HAMMES; FREITAS, 2021) e comparar os resultados obtidos com o desempenho dos LLMs.

CONTRIBUIÇÕES DOS AUTORES

O primeiro e o segundo autor realizaram os experimentos com os LLMs e com o auxílio da professora orientadora compilaram os resultados e fizeram as análises. O primeiro autor escreveu este texto também auxiliado pela orientadora.

AGRADECIMENTOS

Agradecemos ao CNPq pelo apoio financeiro concedido por meio de bolsa na modalidade PIBIC e ao Instituto Federal de São Paulo pelo apoio financeiro concedido por meio de bolsa na modalidade PIBIFSP. Agradecemos também ao Guilherme A. Faller, graduando do curso de letras da UFRGS, e a profa. Dra. Rozane R. Rebechi da UFRGS pela valiosa colaboração na anotação do corpus de estudo.

REFERÊNCIAS

ALMEIDA, T. S. et al. Sabiá-2: A new generation of portuguese large language models. *ArXiv*, abs/2403.09887, 2024. Disponível em: <https://api.semanticscholar.org/CorpusID:268509975>.

[5]Os valores de uso das APIs podem ser consultados em <https://openai.com/api/pricing/> e <https://www.maritaca.ai/>

ASSI, F. M. et al. Ufscar's team at ABSAPT 2022: using syntax, semantics and context for solving the tasks. In: MONTES-Y-GÓMEZ, M.; AL. et (Ed.). *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022)*. CEUR-WS.org, 2022. (CEUR Workshop Proceedings, v. 3202). Disponível em: <https://ceur-ws.org/Vol-3202/absapt-paper5.pdf>.

BALAGE, P. P. *Aspect extraction in sentiment analysis for portuguese language*. Tese (Doutorado), São Carlos - SP, 2017.

CATHARIN, L.; FELTRIM, V. D. Finding opinion targets in news comments and book reviews: 13th international conference, propor 2018, canela, brazil, september 24–26, 2018, proceedings. In: _____. [S.l.: s.n.], 2018. p. 375–384. ISBN 978-3-319-99721-6.

COSTA, R.; PARDO, T. Métodos baseados em léxico para extração de aspectos de opiniões em português. In: *Anais do IX Brazilian Workshop on Social Network Analysis and Mining*. Porto Alegre, RS, Brasil: SBC, 2020. p. 61–72. ISSN 2595-6094. Disponível em: <https://sol.sbc.org.br/index.php/brasnam/article/view/11163>.

HAMMES, L.; FREITAS, L. Utilizando BERTimbau para a classificação de emoções em português. In: *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. Porto Alegre, RS, Brasil: SBC, 2021. p. 56–63. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/stil/article/view/17784>.

LOPES, E.; CORREA, U.; FREITAS, L. Exploring BERT for aspect extraction in portuguese language. *The International FLAIRS Conference Proceedings*, v. 34, Apr. 2021. Disponível em: <https://journals.flvc.org/FLAIRS/article/view/128357>.

MACHADO, M. et al. Learning rules for automatic identification of implicit aspects in portuguese. In: *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. Porto Alegre, RS, Brasil: SBC, 2021. p. 82–91. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/stil/article/view/17787>.

MACHADO, M. T.; PARDO, T. A. S. NILC at ABSAPT 2022: Aspect extraction for portuguese. In: MONTES-Y-GÓMEZ, M.; AL. et (Ed.). *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022)*. CEUR-WS.org, 2022. (CEUR Workshop Proceedings, v. 3202). Disponível em: <https://ceur-ws.org/Vol-3202/absapt-paper5.pdf>.

OLIVEIRA, A. et al. How good is ChatGPT for detecting hate speech in portuguese? In: *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. Porto Alegre, RS, Brasil: SBC, 2023. p. 94–103. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/stil/article/view/25441>.

SAIAS, J.; MOURãO, M.; OLIVEIRA, E. Detailing sentiment analysis to consider entity aspects: An approach for portuguese short texts. *Transactions on Machine Learning and Artificial Intelligence*, Society for Science and Education, United Kingdom, v. 6, n. 2, p. 26–35, 2018.

SANTOS, W.; PARABONI, I. Predição de transtorno depressivo em redes sociais: Bert supervisionado ou ChatGPT zero-shot? In: *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. Porto Alegre, RS, Brasil: SBC, 2023. p. 11–21. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/stil/article/view/25433>.

SILVA, L. H. N. et al. Identificação de aspectos explícitos e implícitos em críticas gastronômicas em português: avaliando o potencial dos LLMs. In: *Proceedings of the 15th Brazilian Symposium in Information and Human Language Technology*. [S.l.: s.n.], 2024.

VARGAS, F.; PARDO, T. Aspect clustering methods for sentiment analysis: 13th international conference, propor 2018, canela, brazil, september 24–26, 2018, proceedings. In: _____. [S.l.: s.n.], 2018. p. 365–374. ISBN 978-3-319-99721-6.