UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Uma Análise de Large Language Models para a Rotulação de Polaridade de Revisões de Produtos Escritas em Português

Natália Sathler de Souza Cunha

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

Natália Sathler de Souza Cunha

Uma Análise de Large Language Models para a Rotulação de Polaridade de Revisões de Produtos Escritas em Português

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Dr. Rafael Geraldeli Rossi

Versão original

São Carlos

2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi, ICMC/USP, com os dados fornecidos pelo(a) autor(a)

S856m

Cunha, Natália Sathler de Souza
Uma Análise de Large Language Models para a Rotulação de Polaridade de Revisões de Produtos Escritas em Português / Natália Sathler de Souza Cunha ; orientador Rafael Geraldeli Rossi. – São Carlos, 2024.
45 p. : il. (algumas color.) ; 30 cm.

Monografia (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. LaTeX. 2. abnTeX. 3. Classe USPSC. 4. Editoração de texto. 5. Normalização da documentação. 6. Tese. 7. Dissertação. 8. Documentos (elaboração). 9. Documentos eletrônicos. I. Rossi, Rafael Geraldeli, orient. II. Título.

Natália Sathler de Souza Cunha

An Analysis of Large Language Models for Polarity Labeling of Product Reviews Written in Portuguese

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Dr. Rafael Geraldeli Rossi

Original version

São Carlos

2024

RESUMO

Cunha, N. S. S. **Uma Análise de Large Language Models para a Rotulação de Polaridade de Revisões de Produtos Escritas em Português**. 2024. 45p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

As compras online estão se tornando cada vez mais comum entre os brasileiros. Essa tendência é impulsionada pela popularização do varejo digital e a maior conectividade da sociedade. No entanto, desafios como entregas atrasadas, produtos danificados e a insegurança dos consumidores em não ver o produto pessoalmente ainda persistem. Diante desses problemas, as revisões de produtos (*reviews*) surgem como uma ferramenta essencial, oferecendo relatos de experiências dos consumidores. Os relatos não apenas ajudam os consumidores, mas também fornecem *insights* valiosos para fabricantes e vendedores, possibilitando melhorias e aumentando a relevância da marca. A mineração de opiniões ou análise de sentimentos busca estudar as emoções e opiniões em relação a produtos e serviços, e gerar *insights* relevantes para tomadores de decisão. A análise de sentimentos é mais acurada ao se utilizar técnicas de aprendizado de máquina supervisionadas, o que exige uma rotulação muitas vezes de uma grande quantidade de revisões, as quais geralmente sáo feitas de forma manual. Porém, a rotulação manual pode ser demorada e propensa a erros. O uso de Large Language Models (LLMs) tem se mostrado promissor nesse contexto, proporcionando uma abordagem mais eficiente na rotulação automática. Contudo, a eficácia desses modelos ainda é um tema de investigação. Dado isso, o objetivo deste trabalho é avaliar a performance dos LLMs na rotulação de *reviews*, buscando identificar se é possível alcançar resultados comparáveis à rotulação humana sem ajustes específicos nos modelos. Os resultados mostram que tanto o GPT-4o-mini quanto o Gemini-1.5-Flash apresentaram acurácias semelhantes, com 79% e 80%, respectivamente, indicando que os modelos podem efetivamente ajudar na análise de sentimentos em português. Tais resultados poderiam ser melhores evitando os erros de rotulação na base de referência, o quais fizeram com que a performance real desses modelos fossem menor.

Palavras-chave: Análise de Sentimentos. LLMs. GPT. Gemini.

ABSTRACT

Cunha, N. S. S. **An Analysis of Large Language Models for Polarity Labeling of Product Reviews Written in Portuguese**. 2024. 45p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

Online shopping is becoming increasingly common among Brazilians, driven by the popularization of digital retail and greater connectivity in society, especially with the entry of Generation Z into the job market. However, challenges such as delayed deliveries, damaged products, and consumer insecurity about not seeing the product in person still persist. In light of these issues, product reviews emerge as an essential tool, providing accounts of consumer experiences. These reports not only assist consumers but also provide valuable insights for manufacturers and sellers, enabling improvements and increasing brand relevance. Opinion mining or sentiment analysis seeks to study emotions and opinions related to products and services. This process is vital for data classification, but manual labeling can be time-consuming and prone to errors. The use of Large Language Models (LLMs) has proven promising in this context, offering a more efficient approach to automated labeling. However, the effectiveness of these models in Portuguese remains a topic of investigation. The goal of this work is to evaluate the performance of LLMs in classifying review texts, seeking to identify whether it is possible to achieve results comparable to human labeling without specific adjustments to the models. Preliminary results show that both GPT-4o-mini and Gemini-1.5-Flash achieved similar accuracies of 79% and 80%, respectively, indicating that the models can effectively assist in sentiment analysis in Portuguese. Future studies are recommended to explore data treatment and compare with other LLMs.

Keywords: Sentimnet ANalysis. LLMs. GPT. Gemini.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

LLMs	Large Language Models
GPT	Generative Pre-trained Transformer
SVM	Support Vector Machine
NLP	Natural Language Processing
BERT	Bidirectional Encoder Representations from Transformers
RNNs	Recurrent Neural Networks
LLaMA	Large Language Model Meta AI
NER	Named Entity Recognition
FewRel	Few-shot Relation Extraction

SUMÁRIO

1 INTRODUÇÃO

Esta cada vez mais comum a compra de produtos online por parte dos brasileiros. Segundo o portal de notícias g1, 61% dos consumidores brasileiros compram mais pela internet do que em lojas físicas, sendo que a maioria afirma que prefrem essa modalidade devido aos descontos e à praticidade de comprar sem sair de casa (G1, 2022). De um lado temos a facilidade, praticidade e conforto dos consumidores em realizar as compras com apenas alguns cliques. Do outro, temos uma maior visibilidade dos vendedores em anunciar e ofertar os produtos. De acordo com o site Exame, a popularização das vendas onlines e os recordes anuais do varejo digital se devem a expansão do acesso à internet, fazendo com que a sociedade esteja mais digital e concetada. Além disso a inserção da geração Z no mercado de trabalho e a cultura *data driven* contribuem para esse avanço (EXAME, 2023). Porém ainda é muito comum encontrar alguns problemas com as compras online, como entregas atrasadas, produtos errados ou danificados e anúncios que não condizem com a realidade do produto. Além dos problemas, muitos consumidores ainda tem o receio de realizar a compra por não conseguir ver o produto pessoalmente e entender se ele realmente atende as necessidades.

Diante dos problemas e receios, algo que vem se tornando comum na internet são os *reviews* dos produtos, que nada mais é do que o relato da experiência dos clientes que já adquiriam determinada mercadoria, no qual o produto é avaliado destacando-se as qualidades e os defeitos, se realmente condiz com o que foi anunciado, se a expriência de compra naquele site foi satisfatória, dentre outros. Segundo a pesquisa feita pelo Reclame AQUI, 74,1% dos consumidores brasileiros possuem o hábito de ler as avaliações antes de realizarem a compra (RECLAME AQUI, 2019).

Porém engana-se quem acredita que os *reviews* auxiliam apenas os consumidores na hora da compra. Os relatos de experiência são extremamente ricos para os fabricantes e vendedores, pois neles contém informações relevantes de como está sendo a aceitação do produto por parte dos conumidores, características que podem ser melhoradas em versões futuras, defeitos recorrentes, comparações com marcas concorrentes, entre outras vantagens. O site *E-commerce* Brasil, ainda cita que as avaliações aumentam a relevância da marca e a taxa de conversão (E-commerce Brasil, 2021). A empresa que consegue realizar a coleta desses dados e fazer uma análise a partir deles tem em mãos uma vantagem competitiva em relação às concorrentes que não levam em consideração esses relatos. Um dos motivos de certas empresas não realizarem análises com esse tipo de dados é que sem as técnicas corretas a análise manual se torna extremamente custosa e morosa. Para mitigar o problema da análise manual, técnicas de mineração de opiniões podem ser empregadas para a extração de informações e classificação das *reviews* (FELDMAN, 2023),

(PATWARDHAN; MARRONE; SANSONE, 2023).

A mineração de opiniões ou análise de sentimentos, tem como objetivo estudar as opiniões, sentimentos, avaliações e emoções das pessoas em relação às entidades e seus respectivos aspectos (atributos) (LIU, 2012) citado por (YUGOSHI, 2018). Segundo Samha, Li and Zhang (2014) o processo de resumir opiniões baseia-se principalmente na identificação e extração de informações opinativas vitais do texto, sendo que a eficiência do processo e qualidade do resumo resultante depende da extração de informações-chave e exclusão de informações supérfluas. No artigo de Cambria *et al.* (2013), os autores argumentam que a análise de sentimentos vai além da polaridade simples (positivo ou negativo), abrangendo uma ampla gama de emoções que impactam a tomada de decisão, tanto para consumidores quanto para empresas. Eles destacam os avanços no uso de aprendizado de máquina e inteligência artificial, que tornam a análise mais precisa e escalável, permitindo que as empresas obtenham *insights* em tempo real sobre percepções de marca e satisfação do cliente. O artigo também aborda desafios técnicos, como a interpretação de ironia, ambiguidades linguísticas e o contexto em que as opiniões são expressas, e discute o impacto dessas técnicas em áreas como marketing, monitoramento de reputação e até política.

A rotulação de uma grande quantidade de dados é uma etapa crucial para treinar modelos de análise de sentimentos. Essa tarefa, no entanto, pode ser bastante demorada e sujeita a erros, resultando em falhas na rotulação que comprometem a qualidade do modelo final. Estudos anteriores mostram que a rotulação manual é um processo que demanda tempo e recursos, podendo impactar a eficiência de projetos de análise de sentimentos (ALZUBAIDI; AL., 2023) e (FELDMAN, 2023).

Recentemente, o surgimento dos Large Language Models (LLMs) trouxe novas perspectivas para o campo do processamento de linguagem natural. Esses modelos têm sido amplamente utilizados para diversas tarefas, incluindo a rotulação de textos, oferecendo uma alternativa potencialmente mais eficiente em comparação com a rotulação manual (HAGOS; AL., 2024) e (GUDIVADA; RAGHAVAN, 2024). No entanto, apesar do avanço na utilização de LLMs, a avaliação de sua eficácia para rotulação de textos em português ainda é escassa, especialmente em contextos que envolvem análise de sentimentos.

Diante desse contexto, o objetivo deste trabalho é avaliar a performance da rotulação de LLMs na tarefa de classificar textos de revisões de produtos escritos em português, atribuindo polaridades negativa, neutra e positiva. Essas polaridades são fundamentais para a análise de sentimentos. Para isso, o trabalho busca responder às seguintes perguntas:

- É possível obter um desempenho de rotulação semelhante ao humano para a tarefa de análise de sentimentos de textos escritos em português utilizando LLMs de propósito geral, isto é, sem fine-tuning no domínio específico?

- Existem LLMs mais adequadas para esta tarefa?

Essas questões visam contribuir para o entendimento do potencial dos LLMs na rotulação automática de dados em português, oferecendo insights valiosos para futuros trabalhos na área. Devido ao grande volume de dados disponível, optou-se por utilizar versões otimizadas de modelos estado-da-arte, que são reconhecidas por serem mais rápidas e menos custosas em termos de processamento. A escolha desses modelos se deve à necessidade de equilibrar tempo e custo, fatores críticos para a rotulação em larga escala.

Os principais resultados mostraram que os dois modelos analisados apresentaram acurácias muito semelhantes, com o GPT-4o-mini alcançando 79% e o Gemini-1.5-Flash obtendo 80%. Em termos de médias de F1-Score, ambos os modelos apresentaram os mesmos valores. Ao examinar o desempenho em classes específicas, observou-se que os dois modelos tiveram resultados superiores na classe positiva, que era a mais prevalente no conjunto de dados. No entanto, o Gemini-1.5-Flash demonstrou uma leve vantagem em relação a essa classe.

Este estudo foi divido em 6 capítulos, sendo no Capítulo 1, já abordado, a introdução na qual tem-se a contextulaização do tema abordado, juntamente com as motivações e objteivos. No Capítulo 2 tem-se o referncial teórico, apresentando os conceitos utilizados nesse trabalho. No Capítulo 3, a revisão bibliográfica, traz os estudos mais recentes sobre o tema. No Capítulo 4, tem-se a metologia, abordando a base de dados trabalhada, modelos utilizados e as métricas de avaliação. No Capítulo 5, tem-se os resultados e por fim no Capítulp 6 as conclusões do estudo.

2 REFERENCIAL TEÓRICO

Nesta seção serão abordados os principais conceitos de análise de sentimentos e suas abordagens, como também uma breve discussão sobre os Large Language Models trazendo alguns modelos mais conhecidos, parâmetros utilizados e suas arquiteturas.

2.1 Análise de Sentimentos

A análise de sentimentos, também conhecida como mineração de opiniões, é o processo de identificar, extrair e classificar sentimentos expressos em um texto (LIU, 2012). Este campo da mineração de texto (ou processamento de linguagem natural) tem como objetivo avaliar se uma opinião em um texto é positiva, negativa ou neutra. Tal análise tem aplicações em diversas áreas, como a análise de avaliações de produtos, serviços e o monitoramento de redes sociais (FELDMAN, 2013). As principais abordagens para análise de sentimentos são as baseadas em regras, em aprendizado supervisionado e as que são baseadas em *deep learning* (GUPTA; RANJAN; SINGH, 2024).

As abordagens baseadas em regras utilizam léxicos de sentimentos e regras gramaticais ou sintáticas para atribuir polaridade ao texto. (LIU, 2012) define a abordagem de léxico como o uso de listas de palavras (léxicos) que já estão previamente classificadas como positivas, negativas ou neutras. Essas listas são aplicadas ao texto, associando as palavras encontradas à polaridade do sentimento. (LIU, 2012) também enfatiza o uso de regras gramaticais para considerar contextos que podem alterar o sentido, como negações (e.g., "não bom"transforma uma palavra positiva em negativa). De acordo com (PANG; LEE, 2008), as abordagens supervisionadas consistem em treinar modelos de aprendizado de máquina utilizando conjuntos de dados rotulados previamente com sentimentos. Entre os métodos mais comuns estão algoritmos como Naive Bayes, Support Vector Machines (SVM) e redes neurais. Essas abordagens requerem a extração de características do texto, como frequência de palavras ou presença de bigramas, para alimentar os algoritmos. Um aspecto importante destacado por (PANG; LEE, 2008) é o uso de técnicas de processamento de linguagem natural (NLP) para transformar o texto em uma representação numérica compreensível para os modelos de aprendizado de máquina.

Recentemente, modelos de redes neurais profundas têm sido utilizados para melhorar a análise de sentimentos, em particular com o uso de arquiteturas como LSTM (Long Short-Term Memory) e transformers. (DEVLIN *et al.*, 2019) apresentam o modelo BERT (Bidirectional Encoder Representations from Transformers), que introduziu uma nova forma de capturar o contexto bidirecional de palavras em uma sentença. Em vez de processar o texto de forma sequencial, o BERT analisa todo o contexto de uma palavra, tanto antes quanto depois, proporcionando uma compreensão mais rica e precisa dos sentimentos

expressos. (DEVLIN *et al.*, 2019) sugerem que o uso de modelos pré-treinados, como o BERT, melhora significativamente o desempenho em tarefas de NLP, como análise de sentimentos, devido à sua capacidade de transferir aprendizado de uma tarefa para outra. Em (CHUMAKOV; KOVANTSEV; SURIKOV, 2023), além de destacarem a eficiência dos LLMs na análise de sentimentos baseada em aspectos, os autores ressaltam que esses modelos são capazes de processar dados de forma eficiente em cenários de grande volume de informação. Eles também mencionam que os LLMs, como o GPT, conseguem generalizar bem mesmo em contextos desconhecidos, o que os torna vantajosos em aplicações onde não há dados rotulados suficientes ou previamente conhecidos para treinamento.

Além dessas abordagens, existem métodos não supervisionados, como *clustering* e aprendizado de tópicos, que são aplicados quando não há rótulos disponíveis nos dados. Esses métodos tentam agrupar dados similares ou identificar tópicos sem a necessidade de um conjunto de treinamento rotulado, o que é útil em cenários onde os dados rotulados são escassos.

2.2 Large Language Models

Os Large Language Models (LLMs) são modelos de linguagem natural treinados em grandes quantidades de dados textuais para realizar diversas tarefas de processamento de linguagem natural (NLP), como tradução, resumo, geração de texto e análise de sentimentos (MIN *et al.*, 2023). Esses modelos utilizam arquiteturas profundas de redes neurais, como a arquitetura *transformer*, para processar e gerar linguagem natural de forma contextual. A arquitetura *transformer*, introduzida por (VASWANI *et al.*, 2017), é o alicerce dos LLMs modernos. Essa arquitetura se distingue pelo uso do mecanismo de atenção, que permite que o modelo "preste atenção"em partes importantes da entrada ao processar texto, capturando relacionamentos de longo alcance entre palavras em uma sequência. (VASWANI *et al.*, 2017) demonstraram que, ao contrário das redes neurais recorrentes, que processam dados sequencialmente, o transformer pode processar todos os *tokens* de entrada em paralelo, resultando em uma grande eficiência computacional. De acordo com (ZHANG *et al.*, 2023) a abordagem dos *trasnformers* permite que o modelo capture relações complexas e dependências de longo alcance nos dados, o que representa um avanço significativo em comparação com arquiteturas anteriores, como as redes neurais recorrentes (RNNs). Essa capacidade de atenção e análise contextual é fundamental para melhorar o desempenho em diversas tarefas, incluindo a análise de sentimentos. (DEVLIN *et al.*, 2019) destacam que os LLMs revolucionaram o campo de NLP ao permitirem o uso de representações contextuais bidirecionais das palavras, o que aumentou significativamente o desempenho em tarefas complexas. Alguns exemplos desses modelos são: BERT (DEVLIN *et al.*, 2019), GPT (Generative Pre-trained Transformer) (RADFORD *et al.*, 2018), Gemini (DEEPMIND, 2023), LLaMA (Large Language Model Meta AI) (TOUVRON *et al.*, 2023)e

o Claude (ANTHROPIC, 2023).

De acordo com (DEVLIN *et al.*, 2019), o BERT utiliza uma arquitetura transformer bidirecional para capturar o contexto de uma palavra, tanto à esquerda quanto à direita, durante o treinamento. Isso permite que o modelo tenha uma compreensão mais profunda do significado das palavras no contexto em que são usadas. (DEVLIN *et al.*, 2019) mostram que o BERT-base tem cerca de 110 milhões de parâmetros, enquanto sua versão maior, o BERT-large, possui 340 milhões.

Desenvolvido pela OpenAI, o GPT é um modelo de linguagem unidirecional focado na geração de texto ao prever a próxima palavra em uma sequência. Conforme explicado por (RADFORD *et al.*, 2018), o GPT se destaca pela sua capacidade de gerar textos coerentes e contínuos. A versão GPT-3, por exemplo, contém 175 bilhões de parâmetros (BROWN *et al.*, 2020). Já o modelo GPT-4o-mini possui aproximadamente 1.5 bilhões de parâmetros. Este modelo é uma versão compacta do GPT-4o, projetada para ser mais leve e eficiente, mantendo a capacidade de realizar tarefas de processamento de linguagem natural de forma rápida e eficaz. Além disso, ele é otimizado para funcionar em ambientes com recursos computacionais limitados, tornando-o uma opção viável para uma ampla gama de aplicações (OPENAI, 2024).

O Gemini é um modelo recente da Google (DEEPMIND, 2023). Também é construído sobre a arquitetura *transformer*, e é amplamente utilizado em tarefas de análise de sentimentos e outras aplicações de NLP (PATWARDHAN; MARRONE; SANSONE, 2023), com uma arquitetura otimizada para eficiência e alta performance em grandes volumes de dados. O modelo Gemini-1.5- Flash possui 8 bilhões de parâmetros.

De acordo com (TOUVRON *et al.*, 2023), o LLaMA é projetado para fornecer modelos de linguagem eficientes, oferecendo alto desempenho mesmo com um número relativamente menor de parâmetros em comparação com gigantes como GPT-3. Seu design visa ser mais acessível, utilizando menos recursos computacionais. Desenvolvido pela Anthropic, o Claude é um LLM com foco em segurança e alinhamento (ANTHROPIC, 2023). Ele é projetado para gerar respostas controladas, buscando evitar outputs que possam ser prejudiciais ou inadequados.

Os LLMs podem ser utilizados de duas formas principais, rodando localmente ou através de serviços de API. Na primeira opção, é possível baixar modelos e executá-los localmente, desde que haja capacidade computacional suficiente para lidar com o tamanho do modelo e a quantidade de dados. Isso requer GPUs ou TPUs potentes, já que modelos como falado anteriormente, possuem muitos parâmetros tornando-os, extremamente grandes e demandam uma grande quantidade de memória. A segunda opção, é por meio de APIs baseadas em nuvem, como o OpenAI API para o GPT ou a API do Gemini da Google Cloud. (BROWN *et al.*, 2020) explicam que as APIs permitem que os desenvolvedores acessem os modelos preditivos sem a necessidade de grandes recursos

computacionais, enviando requisições HTTP e recebendo respostas geradas pelo modelo, tornando essa abordagem mais acessível e prática.

3 REVISÃO BIBLIOGRÁFICA

O trabalho de (DING *et al.*, 2023) tem como objetivo avaliar o desempenho do GPT-3 como anotador de dados, comparando-o com métodos tradicionais de anotação de dados em tarefas de processamento de linguagem natural (NLP). O estudo busca entender se o GPT-3 pode ser utilizado de forma eficaz para a anotação de dados, visando reduzir custos e oferecer uma alternativa acessível para organizações com recursos limitados. O modelo de linguagem utilizado no estudo foi o GPT-3, desenvolvido pela OpenAI. O artigo explora 3 abordagens baseadas no GPT-3 para realizar a anotação de dados, investigando a sua viabilidade e desempenho em comparação com abordagens tradicionais. O estudo conclui que o GPT-3 pode ser usado de forma eficiente para anotação de dados, principalmente em tarefas com pequeno espaço de rótulos, como análise de sentimentos com uma acurácia de 91,3% enquanto a anotação humana é considerada o padrão de ouro, com desempenho geralmente em torno de 93%. No Reconhecimento de Entidades Nomeadas (NER) desempenho do GPT-3 na anotação de dados de NER mostrou F1-score de 87.1%, em comparação com 88-90% de anotações humanas. Embora esteja próximo, a anotação humana ainda supera ligeiramente o modelo em precisão ao lidar com nuances mais sutis no reconhecimento de entidades. Na Extração de Relações (FewRel), o modelo GPT-3 atingiu F1-score de 84.5%, comparado aos 86-88% dos dados anotados por humanos. A capacidade do GPT-3 de generalizar bem com base em prompts direcionados foi considerada um fator decisivo para o seu desempenho robusto. As abordagens baseadas na geração de dados demonstraram ser mais econômicas do que a anotação direta. No entanto, os autores observam que, apesar de suas vantagens em termos de custo e tempo, a qualidade dos dados anotados pelo GPT-3 ainda precisa melhorar para se equiparar à anotação manual.

O estudo conduzido por (GILARDI; ALIZADEH; KUBLI, 2023) tem como objetivo explorar a eficácia do ChatGPT em comparação com anotadores humanos (trabalhadores em plataformas como MTurk) em tarefas de anotação de texto. O estudo se concentra em avaliar se o ChatGPT pode superar anotadores humanos em tarefas como detecção de relevância, postura, tópicos e quadros, buscando também entender se ele pode reduzir significativamente os custos de anotação. O modelo utilizado foi o ChatGPT (versão GPT-3.5-turbo). O estudo utilizou ChatGPT em configurações de zero-shot (sem treinamento adicional) para realizar as tarefas de anotação, comparando seu desempenho com trabalhadores do MTurk e anotadores humanos treinados. Os resultados mostram que o ChatGPT supera significativamente os anotadores humanos de MTurk em termos de precisão, com uma média de 25 pontos percentuais a mais de acurácia em diversas tarefas. Além disso, o acordo entre codificadores (intercoder agreement) do ChatGPT foi maior do que o dos trabalhadores do MTurk e dos anotadores humanos treinados em todas as

tarefas. O custo por anotação com o ChatGPT foi extremamente baixo, cerca de $0.003 por anotação, tornando-o cerca de 30 vezes mais barato que o MTurk. O estudo conclui que o ChatGPT tem o potencial de transformar como a anotação de dados é conduzida, especialmente em termos de eficiência e economia.

O objetivo do trabalho de (QIN *et al.*, 2023) é avaliar se o ChatGPT, particularmente o modelo GPT-3.5-turbo, pode ser considerado um solucionador generalista de tarefas de processamento de linguagem natural (NLP) sem a necessidade de adaptação para tarefas específicas (zero-shot learning). O estudo busca analisar o desempenho do ChatGPT em uma ampla gama de tarefas de NLP, como análise de sentimentos, raciocínio lógico, inferência textual, e sumarização, entre outras, utilizando dados de 20 conjuntos de dados representativos. O modelo utilizado foi o ChatGPT (GPT-3.5-turbo), comparado com outros modelos, como GPT-3.5, FLAN, PaLM, e T5, em diferentes tarefas de NLP. A avaliação foi feita principalmente em cenários de zero-shot learning, em que o modelo não recebe treinamento adicional para as tarefas. O estudo demonstrou que o ChatGPT apresentou um bom desempenho em várias tarefas que requerem raciocínio, como raciocínio aritmético e inferência textual. No entanto, o ChatGPT ainda enfrenta desafios em tarefas mais específicas, como etiquetagem de sequência (ex.: reconhecimento de entidades nomeadas) e sumarização. Em tarefas de análise de sentimentos, o ChatGPT superou o GPT-3.5, especialmente na classificação de textos negativos. Embora o modelo tenha mostrado potencial como um solucionador de tarefas generalistas, ele frequentemente foi superado por modelos que foram especificamente ajustados para essas tarefas.

O artigo realizado por (BAWA; KUMAR; SINGH, 2022) teve como objetivo comparar o desempenho de modelos de linguagem baseados em *transformers*, como BERT e GPT-3, com o de humanos na análise de sentimentos. Os autores buscam entender se esses modelos podem substituir os anotadores humanos na tarefa de classificar sentimentos em textos, investigando a precisão e a eficiência de ambos os métodos. Os modelos utilizados no estudo incluem BERT, GPT-3, e abordagens baseadas em *transformers*. (BAWA; KUMAR; SINGH, 2022) avaliam como esses modelos realizam a análise de sentimentos, comparando seu desempenho com o de anotadores humanos em várias tarefas de classificação de sentimentos. Os resultados mostraram que, embora os modelos de linguagem baseados em *transformers*, como o BERT, possam atingir um desempenho competitivo em tarefas de análise de sentimentos, os humanos ainda superam os modelos em casos que requerem uma interpretação mais sutil e complexa das emoções expressas nos textos. O estudo constatou que, em algumas tarefas específicas, os modelos não conseguiram capturar nuances que os humanos perceberam. Em termos de eficiência, os modelos demonstraram ser mais rápidos na análise de grandes volumes de dados, mas a qualidade das anotações humanas ainda se destacou em contextos mais complexos. Os autores concluem que, embora os transforms tenham potencial para auxiliar na análise de sentimentos, eles não podem completamente substituir humanos devido às suas limitações em captar nuances emocionais.

O trabalho de (KRUGMANN; HARTMANN, 2024) teve como objetivo avaliar a performance de LLMs (Large Language Models) de última geração, como GPT-3.5, GPT-4, e Llama 2, em tarefas de análise de sentimentos, comparando-os com modelos de *transfer learning* (aprendizado por transferência). Os autores realizaram três experimentos para comparar esses LLMs em tarefas de classificação binária e de três classes, sem treinamento prévio (zero-shot), utilizando um total de 20 conjuntos de dados de avaliações de produtos online. Os resultados indicaram que o GPT-4 foi o modelo com o melhor desempenho geral, alcançando uma acurácia média de 93% em classificação binária e 83% em classificação de três classes, superando outros modelos, exceto o SiEBERT no experimento binário. O Llama 2 também apresentou resultados sólidos, superando o GPT-3.5 em tarefas binárias, com 91% de acurácia, apesar de seu tamanho de parâmetro menor (70B comparado a 175B do GPT-3.5). Os resultados mostraram que os LLMs são promissores para análise de sentimentos sem necessidade de treinamento prévio, com o GPT-4 sendo o mais robusto, seguido pelo Llama 2 e GPT-3.5.

Como pode-se observar, nos últimos anos diversos trabalhos da literatura vem fazendo uso de LLMs para rotular dados ou realizar a análise de sentimentos. Entretanto, há uma carência de trabalhos utilizando versões mais recentes, ou o uso em textos escritos na língua portuguesa. Dados isso, esse trabalho de conclusão de curso visa atacar essas lacunas encontradas na literatura. Mais detalhes sobre o método utilizado neste trabalho serão apresentados no próximo capítulo.

4 METODOLOGIA

Neste capítulo são relatados os passos realizados para atingir os objetivos deste trabalho de conclusão de curso. Foram realizados 3 passos: i) coleta da base de dados; ii) aplicação das LLMs; e iii) avaliação dos resultados. Os detalhes de cada um desses passos são apresentados nas próximas seções.

4.1 Base de Dados

A base de dados utilizada no estudo pertence a um conjunto de dados públicos de comércio eletrônico brasileiro disponibilizado pela Olist, podendo ser acessada no Kaggle[1]. Tal conjunto de dados consta com aproximadamente 100 mil reviews de produtos feitos no período de 2016 a 2018 em diversos *marketplaces* no Brasil. Desse conjunto de dados, foi selecionado para as análises a base com 84.991 análises de produtos em português coletadas no site Buscapé[2] em 2013.

A base contém colunas com a identificação de cada *review*, o texto do *review* e a nota de 1 a 5 dada pela pessoa que escreveu a análise do produto no site. Para realizar a análise dos modelos propostos, foi criada uma quarta coluna denominada *target*, contendo as transformações das notas em um sentimento, sendo as notas 1 e 2 classificadas como negativo, 3 classificada como neutro e 4 e 5 classificadas como positivo. A partir daí, foi feita uma análise da distribuição das classes. O gráfico de distribuição é apresentado na Figura 1. Pode-se observar que para essa base, a classe positiva corresponde a 79% do total da base, seguido de 13% da classe neutra e 8% da classe negativa.

[1] Disponível em: <www.kaggle.com/datasets/fredericods/ptbr-sentiment-analysis-datasets/data>. Acesso em: [21 de setembro de 2024].

[2] Disponível em: <https://www.buscape.com.br>.

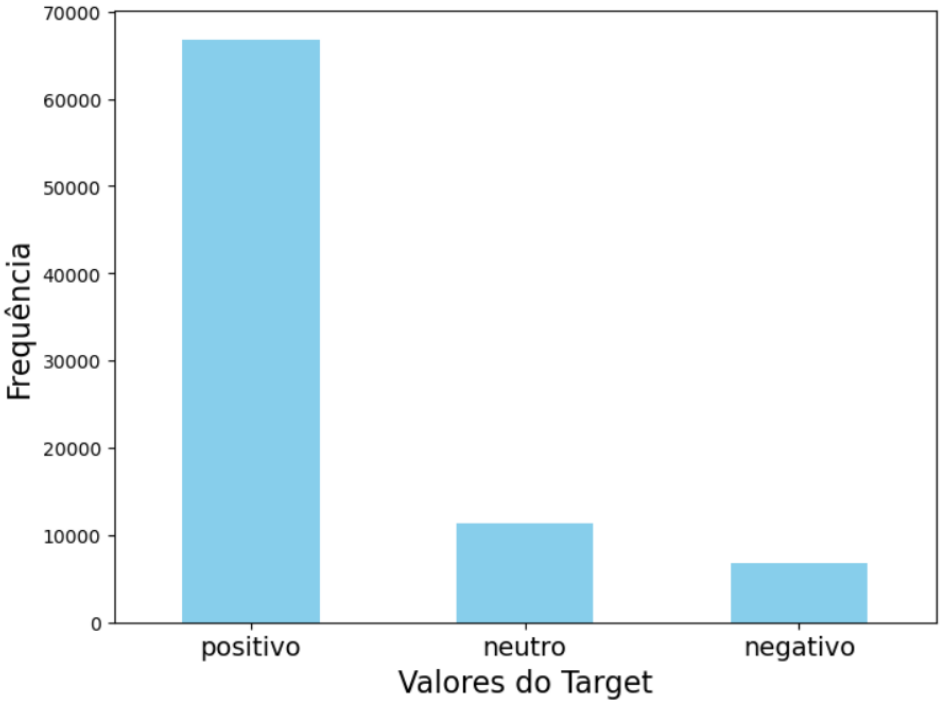

Figura 1 – Distribuição das classes de sentimentos

Em relação aos *reviews*, observou-se uma média de 46 palavras por texto, sendo que a média de palavras únicas era de 35. Na Figura 2. é apresentada uma nuvem de palavras, após a retirada das *stopwords*, sendo as palavras "gostei", "produto", "bom"e "qualidade", que mais se destacam.

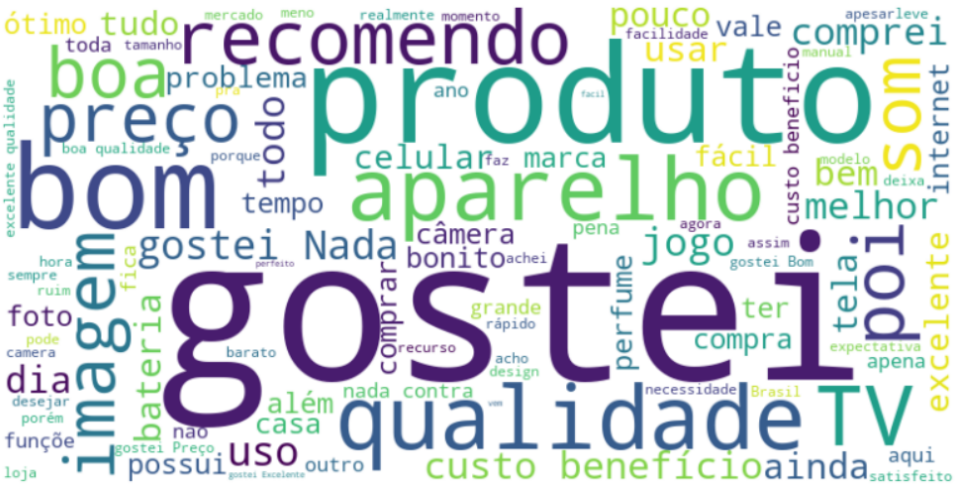

Figura 2 – Nuvem de palavras

Dado o custo financeiro ao se utilizar as LLMs e ao tempo limitado para a execução dos experimentos, foi realizada uma amostragem estratificada, garantindo que a base trabalhada contenha a mesma distribuição das classes da base principal. A amostra representa 10% do total da base contendo 8.500 avaliações.

4.2 Aplicação das LLMs

Os dois modelos utilizados no presente estudo foram GPT-4o-mini[3] e o Gemini-1.5-Flash[4]. O primeiro é uma variante mais compacta e otimizada da família GPT-4, desenvolvida pela OpenAI. O objetivo da escolha da versão "mini"foi devido a ele oferecer um modelo mais leve, rápido e com menor consumo de recursos. O segundo, é um modelo avançado de linguagem natural desenvolvido pelo Google, que se destaca pela sua capacidade de análise em tempo real, com foco em eficiência e precisão. O modelo é conhecido por sua capacidade de processamento rápido e por equilibrar a qualidade das respostas com a velocidade. Em ambos os modelos, cada parte da base foi processada em lotes de 100 *reviews* e foram utilizados os seguintes parâmetros:

- temperatura: determina a aleatoriedade da geração do texto e varia de zero a um. O objetivo de utilizar a temperatura igual a zero foi de sempre utilizar a resposta mais provável provida pela LLM e para evitar variabilidade nos resultados para execuções diferentes.

- max-tokens: determina o tamanho da resposta que o modelo irá gerar, no caso deste trabalho foi escolhido a valor 100.

- n: determina quantas respostas diferentes o modelo soltará para cada requisição. A ideia era ter somente uma resposta para cada avaliação, logo o n foi igual a um.

Além dos parâmetros, foi utilizada a técnica de *zero-shot learning*, uma técnica de aprendizagem de máquina, a qual permite que o modelo classifique classes ou categorias sem a utilização de exemplos. A ideia principal é que o modelo seja capaz de generalizar para novos conceitos sem precisar de dados de treinamento específicos para essas classes ou sem o uso de um especialista de domínio. Para descrever a tarefa a ser realizada foi utilizado o mesmo prompt para ambos os modelos:

"Olá. Você é um rotulador de dados de análise de sentimentos de reviews de vários tipos de produtos vendidos na internet. Dada a review, você tem a tarefa de extrair o sentimento do review. O sentimento é algo para falar se o review é positivo, negativo ou neutro. Gere apenas um sentimento para cada texto da revisão. Dado um identificador e uma revisão associada ao identificador, gere a resposta em formato JSON no seguinte formato:

```
{
```

[3] Disponível em: <https://www.google.com/url?q=https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/&sa=D&source=docs&ust=1727965890171368&usg=AOvVaw3p4CLb-Z44TsYJa4YQ6FIy>.

[4] Disponível em: <https://deepmind.google/technologies/gemini/flash/?hl=pt-br>.

```
   "respostas" : [
        {
            "id" : id do documento,
            "sentimento" : sentimento da review
        }
   ]
}

[texto contendo os ids e revisões]:
"
```

4.3 Avaliação

Para a avaliação do desempenho dos modelos foram utilizadas as métricas de acurácia e F1-score, juntamente com as macro averaging e weighted averaging. As técnicas de precisão e recall foram descritas nesta seção para um melhor entendimento da métrica F1-score. Considere

- TP = Verdadeiros Positivos
- TN = Verdadeiros Negativos
- FP = Falsos Positivos
- FN = Falsos Negativos

para os cálculos das métricas.

A acurácia mede a proporção de previsões corretas em relação ao total de previsões feitas, abrangendo tanto os verdadeiros positivos quanto os verdadeiros negativos. O cálculo dela é dado por

$$Acurácia = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

Porém, quando a base é desbalanceada, não é interessante avaliar a acurácia isolada.

A precisão é utilizada para avaliar a qualidade das previsões feitas, indicando a proporção de verdadeiros positivos em relação ao total de exemplos classificados como positivos. Ou seja, ela mede a exatidão das previsões positivas do modelo, mostrando a capacidade do modelo de evitar falsos positivos. O cálculo é dado por

$$Precisão = \frac{TP}{TP + FP} \tag{4.2}$$

O recall, também conhecido como sensibilidade ou taxa de verdadeiro positivo, mede a capacidade do modelo de identificar todos os casos positivos relevantes em um

conjunto de dados. O recall é a proporção de verdadeiros positivos em relação ao total de casos que realmente são positivos. Essa métrica é calculada por

$$Recall = \frac{TP}{TP + FN} \tag{4.3}$$

O F1-score é uma métrica que combina tanto a precisão quanto o recall em uma única medida por meio de uma média harmônica, fornecendo uma visão mais equilibrada do desempenho do modelo. Ele é especialmente útil em cenários onde há uma classe positiva desbalanceada, já que leva em conta tanto os falsos positivos quanto os falsos negativos. O valor dessa métrica varia de 0 a 1, onde 1 indica um modelo perfeito. O cálculo é dado por

$$F1 = 2 \times \frac{Precisão \times Recall}{Precisão + Recall} \tag{4.4}$$

A *macro average* é calculada ao tomar a média aritmética das métricas (precision, recall, f1-score) para cada classe, sem considerar o suporte, que é o número de ocorrências de cada classe. O cálculo é dado por

$$MacroAverage = \frac{1}{n} \sum_{i=1}^{n} M_i \tag{4.5}$$

A *weighted average* leva em consideração o suporte de cada classe ao calcular a média. Cada classe contribui para a média de acordo com o número de instâncias que ela representa. O cálculo é dado por

$$WeightedAverage = \frac{\sum_{i=1}^{n} (M_i \times S_i)}{\sum_{i=1}^{n} S_i} \tag{4.6}$$

sendo

- n é o número total de classes,
- M_i é a métrica (precisão, recall ou F1-score) para cada classe
- S_i é o número de instâncias (suporte) para cada classe

5 RESULTADOS

Apesar do prompt indicar para os modelos classificar os sentimentos em apenas 3 tipos de classes (positivo, neutro ou negativo) o modelo GPT-4o-mini apresentou para uma *review* a classificação 'misturado'. Sendo assim essa linha foi excluída para o cálculo das métricas. Na Tabela 1 são apresentados os resultados das métricas Precison, Recall e F1-Score para cada uma das classes e na Tablea 2 são apresentados os resultados sumarizando todas as classes.

Modelo	Métrica	Negativo	Neutro	Positivo
GPT-4o-mini	Precision	0.46	**0.33**	**0.91**
	Recall	0.77	**0.23**	0.89
	F1-Score	0.58	**0.27**	0.90
	Support	681	1136	6682
Gemini-1.5-Flash	Precision	**0.48**	0.32	0.90
	Recall	0.77	0.19	**0.91**
	F1-Score	**0.59**	0.24	**0.91**
	Support	681	1137	6682

Tabela 1 – Comparação das métricas entre as classes para os dois modelos

	Acurácia	Macro Avg F1	Weighted Avg F1
GPT-4o-mini	0.79	0.58	0.79
Gemini-1.5-Flash	**0.80**	0.58	0.79

Tabela 2 – Tabela de métricas de desempenho geral dos modelos

Ambos os modelos obtiveram uma boa acurácia, sendo o Gemini-1.5-Flash (80%) um pouco maior que o GPT-4o-mini (79%). Porém, analisando as outras métricas em cada classe, o modelo não tem um desempenho equilibrado nas classes "negativo"e "neutro". A alta acurácia se deve, em grande parte, ao bom desempenho na classe "positivo", que tem o maior suporte (6682 exemplos), logo o modelo tende a acertar mais nesta classe porque ela domina o conjunto de dados. Em relação a precisão:

- Negativo: o segundo modelo tem uma precisão um pouco maior (0.48) em relação ao primeiro (0.46), sugerindo que o segundo modelo faz previsões ligeiramente mais corretas para a classe "negativa".
- Neutro: o primeiro modelo tem uma leve melhora na precisão para a classe "neutro"(0.33 vs. 0.32), mas ambas as precisões são baixas.

- Positivo: ambos os modelos têm precisões muito semelhantes para a classe "positivo"(0.91 e 0.90), indicando que ambos são muito bons em identificar corretamente os exemplos dessa classe.

Em relação ao recall:

- Negativo: O recall para a classe "negativo"é idêntico nos dois modelos (0.77), o que significa que ambos conseguem identificar a mesma quantidade de exemplos negativos.
- Neutro: O primeiro modelo (0.23) tem um desempenho superior em relação ao segundo (0.19), indicando que ele consegue encontrar mais exemplos da classe "neutro", ainda que o desempenho seja baixo em ambos os casos.
- Positivo: O segundo modelo (0.91) tem um recall ligeiramente maior do que o primeiro (0.89), o que indica que ele está capturando mais exemplos da classe "positivo".

E por fim, em relação ao F1-Score:

- Negativo: O F1-score, que equilibra precisão e recall, é praticamente o mesmo em ambos os modelos (0.58 no primeiro e 0.59 no segundo).
- Neutro: O primeiro modelo (0.27) melhora levemente em relação ao segundo (0.24) na classe "neutro", mas ambos ainda apresentam baixo desempenho.
- Positivo: Ambos os modelos têm um F1-score quase idêntico para a classe "positivo"(0.90 e 0.91), refletindo o equilíbrio entre precisão e recall elevado para essa classe.

No geral, ambos os modelos têm desempenho semelhante, com uma ligeira vantagem para o segundo modelo, especialmente na classe "positivo"(com recall e F1-score ligeiramente maiores).

Foram analisadas algumas avaliações em que os modelos erraram para entender o motivo da divergência da resposta do modelo com o *target*. Na Tabela 3 o target era positivo e o modelo classificou como neutro, como o texto tem partes positivas e negativas a respeito do produto é compreensível que o modelo tenha classificado como neutro. Na Tbela 4, apesar da nota do cliente ser neutra, o texto dele contém mais aspectos negativos, o que é compreensível entender o que levou o modelo a classificar como negativo. Na Tabela 5, a avaliação do cliente é extremamente positiva apesar da nota ter sido neutra, logo a classificação do modelo foi positiva. Por fim, na Tabela 6 a avaliação do cliente foi neutro, porém o texto em si demonstra que ele não está totalmente satisfeito com o

id	review	rating	target	sentimento
2961	Bom custo benefício. Executa chamadas e envio de mensagens sem problemas. Acesso à internet muito limitado. O que gostei: Bom preço, diversidade de funções e ferramentas. O que não gostei: Pouca memória interna e sempre que acesso à internet ocorrem algumas incompatibilidades.	4	positivo	neutro

Tabela 3 – Exemplo 1 do modelo GPT-4o-mini

id	review	rating	target	sentimento
8428	Deveria aceitar cartão de memória de 4GB. O que gostei: Toca músicas em muitas configurações nas quais outros celulares não possuem. O que não gostei: A bateria não dura muito mesmo configurado para economia de energia. Não aceita cartão de memória além de 1GB (pelo menos o meu não aceita).	3	neutro	negativo

Tabela 4 – Exemplo 2 do modelo GPT-4o-mini

id	review	rating	target	sentimento
4658	Estou plenamente satisfeito com o produto. Se por acaso você ficar na dúvida entre comprar essa ou outra, compre esta. O que gostei: Perfeita. De extrema qualidade. Rica em recursos. Algumas pessoas criticam a ausência do wireless USB, levando outras a pensarem que ter o wireless USB é uma condição para ter a internet. Não é. É possível conectar a internet com um cabo LAN. E alguém comentou no Buscapé que a TV não vem com Skype. Vem. Eu acho que a pessoa não olhou todas as opções de menu. Alguém também disse que ela é frágil. Definitivamente, não é. A TV É PERFEITA. O que não gostei: A TV não faz pipoca.	5	neutro	positivo

Tabela 5 – Exemplo 1 do modelo Gemini-1.5-Flash

produto e aparentemente o objeto que ele não consegue comprar é importante para que ele consiga usufruir do produto.

id	review	rating	target	sentimento
1414	Gostaria de saber onde comprar as tiras medidoras. O que gostei: O preço do aparelho é atraente, mas não consigo comprar as tiras medidoras. O que não gostei: Adquiri o kit completo e não consigo mais comprar as tiras medidoras.	3	neutro	negativo

Tabela 6 – Exemplo 2 do modelo Gemini-1.5-Flash

De acordo com os resultados obtidos por ambos os modelos, e pelas análises realizadas de alguns *reviews* que foram classificados errados pelos modelos, observou-se que os modelos foram capazes sim de se aproximar da rotulação humana.

6 CONCLUSÕES

A análise de *reviews* online é uma prática comum entre as pessoas que procuram saber os benefícios dos produtos, preços, questões de entregas, qualidade do atendimento e prestação de serviços dos e-commerces. Como o cliente não consegue ver o produto e inspecionar da forma que gostaria para definir se a compra será benéfica ou não, as avaliações de outros consumidores se tornam uma boa fonte para ajudar na decisão. Da mesma forma que os consumidores tiram proveito desses *reviews*, as empresas também conseguem se beneficiar desse artifício, conseguindo entender melhor as necessidades dos clientes, pontos positivos dos produtos e dos serviços prestados e onde precisam realizar melhorias. Porém a coleta dessas informações e tratamento desses dados podem levar tempo e consumir recursos significativos para a empresa. O processo de classificação dos reviews para que os modelos sejam treinados e obter informações de negócios, normalmente são realizados por pessoas, podendo acontecer divergências nas classificações, pois está sujeito a diversas interpretações, além do tempo gasto para tal tarefa.

Nos últimos anos, os Large Language Models (LLMs) têm sido cada vez mais utilizados em processos de automação de processos linguísticos e para melhorar a interação humano-máquina, sendo que alguns trabalhos da literatura já utilizados LLMs para rotulação de dados. Porém, não foram encontrados trabalhos que analisam a capacidade de rotulação de dados desses modelos para a análise de sentimentos de avalições escritas em português. Dado isso, o objetivo deste trabalho foi realizar tais avaliações utilizando duas das LLMs mais utilizadas em suas versões mais rápidas e menos custosas.

Como principais resultados ambos os modelos estudados obtiveram acurácia bem próximas, sendo o modelo GPT-4o-mini com 79% e o Gemini-1.5-Flash 80%. Em relação as médias do F1-Score, ambos obtiveram as mesmas porcentagens. Analisando as classes individualmente, ambos os modelos obtiveram um melhor desempenho na classe positiva, a qual era predominante na base, porém o Gemini-1.5-Flash obteve uma melhor vantagem nessa classe. Outro ponto improtante a ser ressaltado é que os modelos obtiveram esses resultados sem nenhum tratamento prévio na base e sem neceesidade de treinamemto. Sendo assim, dada a pergunta principal deste trabalho, que era a de analisar a capacidade desses modelos de rotularem dados para a análise de sentimentos de avaliações escritas em portguês, observou-se que os dois modelos foram capazes de realizar essa classificação com desempenhos bem semelhantes.

Algumas dificuldades encontradas durante as análises impediram que fossem realizados mais testes em modelos diferentes e em uma base maior de dados, para verificar se haveria resultados melhores que os que foram obtidos. Tais impedimentos foram os recursos para a utilização dos modelos, pois em cada um havia limites de tokens por requi-

sição e cada requisição era necessário um investimento financeiro para obter as respostas, investimentos esses que no momento eram limitados. Sendo assim, para futuros trabalhos seria interessante um tratamento mais aprofundado na base para melhorar a qualidade dos textos, testar técnicas de balanceamento para identificar se há melhoras nas classificações de sentimentos com menores ocorrências e testar outros modelos para uma comparação de desempenhos entre os modelos já analisados, como por exemplo o Large Language Model Meta AI (LLAMA) desenvolvido pelo Meta e o Claude desenvolvido pelo Anthropic. Outro ponto foi a limitação em analisar o sentimento inteiro da *reviews*, que pode ser falho em casos em que há uma análise de mais de um aspecto de um produto, como foi o caso dos exemplos analisados em que as classificações dos modelos divergiam do *target*. Nas *reviews* era comum o cliente falar tanto de pontos positivos quanto negativos dos produtos, o que acabou interferindo na classificação final do modelo. Portanto, uma extensão desse trabalho seria a avaliação de modelos de LMM para a análise de sentimentos baseada em aspectos.

REFERÊNCIAS

ALZUBAIDI, O.; AL. et. Challenges in sentiment analysis: The importance of data annotation. **Journal of Data Science and Artificial Intelligence**, v. 45, n. 2, p. 120–135, 2023.

ANTHROPIC. Claude: An ai language model by anthropic. **Anthropic Blog**, 2023. Available at: <https://www.anthropic.com/index/claude>.

BAWA, P.; KUMAR, A.; SINGH, R. Human versus transformer models in sentiment analysis: Can transformers replace humans? *In*: **Proceedings of the 2022 International Conference on Computational Linguistics (COLING)**. [*S.l.: s.n.*], 2022.

BROWN, T. B. *et al.* Language models are few-shot learners. **arXiv preprint arXiv:2005.14165**, 2020.

CAMBRIA, E. *et al.* Mineração de opiniões e análise de sentimentos: A revolução das emoções computacionais. **Revista Brasileira de Computação Aplicada**, v. 5, p. 77–89, 2013.

CHUMAKOV, S.; KOVANTSEV, A.; SURIKOV, A. Generative approach to aspect based sentiment analysis with gpt language models. **Procedia Computer Science**, Elsevier, v. 229, p. 284–293, 2023.

DEEPMIND, G. Gemini: A new paradigm for ai and machine learning. **Google DeepMind Blog**, 2023. Available at: <https://blog.google/products/ai/gemini-new-paradigm-ai-and-machine-learning/>.

DEVLIN, J. *et al.* Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2019.

DING, B. *et al.* Is gpt-3 a good data annotator? **arXiv preprint arXiv:2212.10450v2**, 2023.

E-commerce Brasil. **A importância de ter reviews no seu e-commerce**. 2021. Available at: <https://www.ecommercebrasil.com.br/noticias/importancia-reviews-e-commerce/>.

EXAME. **5 tendências para e-commerce em 2023: veja como se preparar para bons resultados**. 2023. Available at: <https://exame.com/negocios/5-tendencias-para-e-commerce-em-2023-veja-como-se-preparar-para-bons-resultados/>.

FELDMAN, R. Techniques and applications for sentiment analysis. **Communications of the ACM**, ACM New York, NY, USA, v. 56, n. 4, p. 82–89, 2013.

FELDMAN, R. Techniques and applications for sentiment analysis and opinion mining. **Communications of the ACM**, ACM, v. 66, n. 4, p. 55–64, 2023.

G1. **61físicas, aponta estudos**. 2022. Available at: <https://g1.globo.com/economia/noticia/2022/12/14/61percent-dos-brasileiros-compram-mais-pela-internet-do-que-em-lojas-fisicas-aponta-estudo.ghtml>.

GILARDI, F.; ALIZADEH, M.; KUBLI, M. Chatgpt outperforms crowd-workers for text-annotation tasks. **arXiv preprint arXiv:2303.15056**, 2023.

GUDIVADA, V.; RAGHAVAN, V. Large language models (llms): Survey, technical frameworks, and future challenges. **Artificial Intelligence Review**, v. 57, p. 356–370, 2024.

GUPTA, S.; RANJAN, R.; SINGH, S. N. Comprehensive study on sentiment analysis: From rule-based to modern llm based system. **arXiv preprint arXiv:2409.09989**, 2024.

HAGOS, B.; AL. et. Recent advances in generative ai and large language models: Current status, challenges, and perspectives. **Artificial Intelligence Review**, v. 57, n. 3, p. 340–357, 2024.

KRUGMANN, J. O.; HARTMANN, J. Sentiment analysis in the age of generative ai. **Customer Needs and Solutions**, Springer, v. 11, n. 3, 2024.

LIU, B. Sentiment analysis and opinion mining. **Synthesis lectures on human language technologies**, Morgan & Claypool Publishers, v. 5, n. 1, p. 1–167, 2012.

MIN, B. *et al.* Recent advances in natural language processing via large pre-trained language models: A survey. **ACM Computing Surveys**, ACM New York, NY, v. 56, n. 2, p. 1–40, 2023.

OPENAI. Gpt-4o mini unveiled: A cost-effective, high-performance alternative to claude haiku, gemini flash and gpt 3.5 turbo. 2024. Available at: <https://www.unite.ai/gpt-4o-mini-unveiled-a-cost-effective-high-performance-alternative-to-claude-haiku-gemini-flash-and-gpt-3->.

PANG, B.; LEE, L. Opinion mining and sentiment analysis. *In*: NOW PUBLISHERS INC. **Foundations and Trends in Information Retrieval**. [*S.l.: s.n.*], 2008. v. 2, n. 1-2, p. 1–135.

PATWARDHAN, N.; MARRONE, S.; SANSONE, C. Transformers in the real world: A survey on nlp applications. **Information**, MDPI, v. 14, n. 4, p. 242, 2023.

QIN, C. *et al.* Is chatgpt a general-purpose natural language processing task solver? **arXiv preprint arXiv:2302.06476**, 2023.

RADFORD, A. *et al.* Improving language understanding by generative pre-training. *In*: . [*S.l.: s.n.*], 2018.

RECLAME AQUI. **60% dos brasileiros não sabem o que são reviews, aponta pesquisa do Reclame AQUI**. 2019. Available at: <https://noticias.reclameaqui.com.br/noticias/60-dos-brasileiros-nao-sabem-o-que-sao-reviews-aponta-pesqui_3625/>.

SAMHA, A. K.; LI, Y.; ZHANG, J. Aspect-based opinion extraction from customer reviews. **arXiv preprint arXiv:1404.1982**, 2014.

TOUVRON, H. *et al.* **LLaMA: Open and Efficient Foundation Language Models**. [*S.l.*], 2023. Available at: <https://arxiv.org/abs/2302.13971>.

VASWANI, A. *et al.* Attention is all you need. *In*: **Advances in neural information processing systems**. [*S.l.: s.n.*], 2017. p. 5998–6008.

YUGOSHI, I. P. M. **Mineração de opiniões baseada em aspectos para revisões de produtos e serviços**. 2018. Tese (Doutorado) — Universidade de São Paulo, 2018.

ZHANG, E. Y. *et al.* From turing to transformers: A comprehensive review and tutorial on the evolution and applications of generative transformer models. **Procedia Computer Science**, Elsevier, v. 229, p. 284–293, 2023.