Proposta recebida em Junho 2024 e aceite para publicação em Outubro 2024.

Explorando a Eficácia dos Modelos Generativos em Tarefas de Análise de Sentimentos no Português Brasileiro

Exploring the Effectiveness of Generative Models in Sentiment Analysis Tasks in Brazilian Portuguese

Gladson de Araújo
Universidade do Estado do Amazonas

Tiago de Melo
Universidade do Estado do Amazonas

Carlos Maurício S. Figueiredo
Universidade do Estado do Amazonas

Resumo

Modelos de linguagem de grande escala (LLMs) têm se destacado em diversas tarefas de processamento de linguagem natural (PLN). Este artigo investiga sua eficácia em tarefas de análise de sentimentos no contexto do português brasileiro, explorando a identificação de frases opinativas, polaridade e frases comparativas. O estudo avalia o desempenho de modelos como ChatGPT e Sabiá em diferentes tarefas e conjuntos de dados, comparando-os com métodos da literatura. Ainda, exploramos o uso de LLMs na anotação automática de dados. Os resultados demonstram o potencial dos LLMs na análise de sentimentos, especialmente na identificação de polaridade, e discutem suas limitações e aplicações em tarefas de anotação de dados.

Palavras chave

modelos de linguagem de grande escala, análise de sentimentos, anotação automática de dados

Abstract

Large language models (LLMs) have been successfully applied in various natural language processing (NLP) tasks. This paper investigates their effectiveness in sentiment analysis tasks in the context of Brazilian Portuguese, exploring the identification of opinionated sentences, polarity, and comparative sentences. The study evaluates the performance of models such as ChatGPT and Sabiá on different tasks and datasets, comparing them with methods from the literature. Furthermore, we explore the use of LLMs in automatic data annotation. The results demonstrate the potential of LLMs in sentiment analysis, especially in polarity identification, and discuss their limitations and applications in data annotation tasks.

Keywords

large language models, sentiment analysis, automatic data annotation

1. Introdução

Modelos de linguagem de grande escala (LLMs) demonstraram sua capacidade de lidar com uma variedade de tarefas de processamento de linguagem natural (PLN) sem a necessidade de dados de treinamento específicos, um fenômeno denominado aprendizado auto-supervisionado (*self-learning*). Isso é alcançado submetendo o modelo com prompts adequados (Brown et al., 2020). A capacidade de realizar novas tarefas por instrução representa um grande avanço em direção à inteligência artificial para múltiplos propósitos. Embora os LLMs contemporâneos exibam desempenho louvável em certos cenários, elas permanecem propensas a erros no aprendizado *zero-shot* (Chang et al., 2023) e *few-shot* (Li et al., 2023). Além disso, várias configurações, como as definições de temperatura, podem influenciar profundamente a eficácia do modelo. Essas limitações implicam que os LLMs atuais podem não servir verdadeiramente como sistemas de linguagem abrangentes.

O lançamento recente do ChatGPT pela OpenAI atraiu atenção significativa da comunidade de PLN. O ChatGPT[1] é um modelo baseado em Redes Neurais Transformer (Vaswani et al., 2023) treinado com aprendizado por reforço a partir de *feedback* humano (RLHF[2]) (Christiano et al., 2023). O treinamento RLHF consiste de três etapas: primeiro, treinar um modelo de linguagem com aprendizado auto-supervisionado; segundo, reunir dados de comparação baseados em preferências humanas e treinar um modelo de recompensa; e terceiro, otimizar o modelo de linguagem contra o modelo de recompensa através de aprendizado por reforço. Como resultado desse treinamento, o ChatGPT demonstrou

[1] https://platform.openai.com
[2] Reinforcement Learning from Human Feedback.

DOI: 10.21814/lm.16.2.433

capacidades impressionantes, como gerar respostas textuais de alta qualidade para entradas humanas, rejeitar perguntas inadequadas e corrigir erros anteriores com base em conversas subsequentes. Após o lançamento do ChatGPT, muitas empresas e pesquisadores iniciaram o desenvolvimento de modelos de linguagem generativos especializados em diferentes idiomas. Para português, por exemplo, o Sabiá é um modelo que foi treinado exclusivamente na língua portuguesa (Pires et al., 2023) e que está disponível através da MariTalk API[3] como um *chatbot*. Embora a maioria dos modelos de linguagem generativos tenha demonstrado capacidades conversacionais impressionantes, a comunidade de PLN ainda está incerta sobre sua capacidade de alcançar uma generalização *zero-shot* ou *few-shot* superior em comparação com os LLMs existentes, especialmente em idiomas diferentes do inglês (Chang et al., 2023). Especificamente, sua eficácia em português do Brasil ainda não foi explorada com maior profundidade.

Para abordar essa lacuna de pesquisa, este trabalho apresenta uma investigação abrangente sobre a capacidade de aprendizado de diferentes modelos LLMs, avaliando os seus desempenhos em uma ampla gama de conjuntos de dados de PLN em português do Brasil, incluindo três tarefas relevantes de análise de sentimentos: i) identificação de sentenças opinativas; ii) cálculo de polaridade; e iii) identificação de sentenças comparativas. Particularmente, trata-se de uma ampliação do trabalho publicado inicialmente no artigo de Araujo et al. (2024). Neste trabalho, ampliou-se a pesquisa com a comparação das técnicas de prompt *zero-shot* e *few-shot*, a introdução da avaliação de uma versão mais recente do ChatGPT (versão 4), a avaliação de um modelo treinado em português, o Sabiá, e a geração de novos resultados de modelos da literatura treinados com dados anotados pela melhor dos LLMs testados.

As três tarefas escolhidas para avaliação são importantes em PLN no que diz respeito a problemas de detecção de informações de comentários de avaliações de pessoas sobre qualquer assunto, de qualquer mídia textual e, principalmente, da Internet. Assim, essas contribuições podem ser aplicadas a vários problemas de mineração de dados. Mais especificamente, foram investidas as seguintes questões de pesquisa neste trabalho:

Questão de Pesquisa 1 (QP1): Como os modelos ChatGPT e Sabiá se comportam como solucionadores para as três tarefas de análise de sentimentos mencionadas acima? Para abordar isso, foi comparado empiricamente o desempenho do ChatGPT e do Sabiá com métodos considerados estado da arte.

Questão de Pesquisa 2 (QP2): Como a anotação gerada pelo melhor modelo de LLM avaliado neste trabalho influencia os dados de treinamento para diferentes classificadores abordando as três tarefas de análise de sentimentos mencionadas acima? Para isso, foi comparada empiricamente a anotação gerada pelo melhor modelo avaliado para dados de treinamento de diferentes classificadores que abordam as três tarefas de análise de sentimentos mencionadas.

Até onde se sabe, este é o primeiro trabalho que investiga o problema de usar um LLM para abordar tarefas relevantes de análise de sentimentos em português. As principais contribuições deste trabalho podem ser resumidas da seguinte forma:

- Realização de experimentos para avaliar o impacto do hiperparâmetro de temperatura no desempenho do ChatGPT em tarefas de PLN.
- Avaliação das versões do ChatGPT 3.5 e ChatGPT 4.0 como solucionadores em três diferentes tarefas de análise de sentimentos. Através da execução dos experimentos, foi possível identificar que o ChatGPT exibe desempenho excepcional em tarefas de análise de sentimentos, especificamente na identificação de subjetividade e polaridade em frases. Em termos de identificação de frases comparativas, o ChatGPT demonstra um desempenho inferior em comparação com os *baselines*.
- Realização de uma análise comparativa entre modelos treinados em múltiplos idiomas, como o ChatGPT, e um modelo treinado exclusivamente com textos em português, como é o caso do Sabiá.
- Análise abrangente da viabilidade de aproveitar o ChatGPT para anotação de dados para tarefas complexas de PLN.

O restante do artigo está organizado da seguinte forma. A Seção 2 apresenta conceitos preliminares importantes para o entendimento do trabalho. A Seção 3 apresenta uma revisão dos trabalhos relacionados sobre Modelos de Linguagem de Grande Escala (LLMs) e a Seção 4 apresenta uma visão geral da metodologia aplicada no presente estudo. A Seção 5 apresente e discute a avaliação experimental da abordagem proposta. Finalmente, a Seção 6 apresenta e discute nossas principais conclusões, limitações e direções futuras de pesquisa.

[3] https://github.com/maritaca-ai/maritalk-api

2. Conceitos Preliminares

2.1. Análise de Sentimentos

Análise de sentimentos é uma subárea do Processamento de Linguagem Natural (PLN) que busca identificar e extrair informações subjetivas de textos, tais como emoções, opiniões e atitudes expressas por usuários em diferentes contextos, incluindo redes sociais, comentários sobre produtos e *feedbacks* de clientes (Liu, 2020).

As tarefas centrais na análise de sentimentos incluem a identificação de características específicas ou atributos dos produtos e serviços mencionados nos textos, conhecida como extração de aspectos (Rana & Cheah, 2016). Estes aspectos extraídos podem ser utilizados em uma miríade de domínios, tais como análise de debate político (Seno et al. (2024)), sumarização de opiniões sobre produtos eletrônicos (de Melo et al., 2018; Hayatin et al., 2024) ou para a identificação de notícias falsas (Hou et al., 2024). Para isso, é necessário realizar a identificação de sentenças opinativas, que se diferenciam de sentenças factuais por expressarem uma opinião ou julgamento (Pandey & Deorankar, 2019; de Oliveira & de Melo, 2021). Essa tarefa de identificação de sentenças opinativas é frequentemente acompanhada pela análise de polaridade, onde a frase é classificada como positiva ou negativa com base na emoção expressa (Oliveira & de Melo, 2020). Por exemplo, em um comentário de um cliente sobre um restaurante, a extração de aspectos permite identificar termos como "comida" ou "atendimento" e associá-los às opiniões positivas ou negativas expressas. Além dessas tarefas, a identificação de sentenças comparativas também é relevante, pois envolve a detecção de frases que comparam dois ou mais elementos, identificando as preferências e percepções dos consumidores (Carvalho et al., 2017). Essas tarefas, em conjunto, fornecem uma visão mais detalhada e específica das opiniões dos usuários, permitindo uma análise mais profunda e direcionada dos sentimentos expressos nos textos.

2.2. Modelos de Aprendizagem de Máquina

A aprendizagem de máquina é um subcampo da Inteligência Artificial que se concentra no desenvolvimento de algoritmos capazes de aprender padrões a partir de dados e realizar previsões ou tomar decisões sem serem explicitamente programados para isso. Esses modelos têm sido amplamente utilizados em uma variedade de aplicações, como reconhecimento de imagem, processamento de linguagem natural e análise preditiva. Entre os modelos mais comuns, o Naive Bayes (NB) se destaca por sua simplicidade e eficácia, especialmente em tarefas de classificação de textos (McCallum & Nigam, 1998). O NB é baseado na aplicação do teorema de Bayes com a suposição de independência entre as características, o que o torna eficiente mesmo em grandes conjuntos de dados. Apesar de sua simplicidade, o NB é robusto e frequentemente utilizado em situações onde a interpretação rápida e a escalabilidade são essenciais.

Outro modelo amplamente utilizado é o Gradient Boosting Trees (GBT), que é um método de aprendizado em conjunto que constrói um modelo preditivo forte a partir de uma combinação de modelos fracos, como árvores de decisão. O GBT funciona iterativamente, ajustando as previsões dos modelos anteriores para minimizar o erro residual, o que o torna particularmente eficaz em tarefas de regressão e classificação (Friedman, 2001). Devido à sua flexibilidade e alto desempenho, o GBT tem sido aplicado com sucesso em áreas como análise de sentimentos, previsão de risco de crédito, detecção de fraudes e outras tarefas que requerem alta precisão preditiva.

Com o aumento da complexidade dos modelos de aprendizado de máquina e a necessidade de otimização de hiperparâmetros, os modelos de *Automated Machine Learning* (AutoML) têm se tornado uma alternativa promissora. O AutoML automatiza o processo de desenvolvimento de modelos, desde o pré-processamento de dados até a seleção do modelo e a otimização de hiperparâmetros, permitindo que os pesquisadores criem modelos robustos e eficientes de maneira mais ágil (Elshawi et al., 2019). Ferramentas como AutoGluon exemplificam essa abordagem, utilizando técnicas avançadas como otimização bayesiana e aprendizado por reforço para explorar o espaço de hiperparâmetros e identificar configurações que maximizam o desempenho do modelo (Erickson et al., 2020). Além disso, AutoGluon integra múltiplos modelos e técnicas de pré-processamento para construir *pipelines* personalizados, otimizados para problemas específicos. Particularmente, para a tarefa de análise de sentimentos em textos, o AutoGluon possui classes como o *TextPredictor* e *TabularPredictor*, que empregam modelos pré-treinados baseados em Transformers, como o ELECTRA, RoBERTa, ou Multilingual BERT, para extração de características de textos, e treinam um modelo agregador para a tarefa de classificação em questão.

Diante do exposto, esses modelos citados têm sido empregados como estado-da-arte em tarefas de análise de sentimentos, e são usados neste artigo como *baseline*.

2.3. Modelos de Linguagem Generativos

Modelos de linguagem generativos são uma classe de modelos de aprendizagem de máquina projetados para gerar texto coerente e relevante em resposta a um determinado *prompt*. Baseados em arquiteturas como as redes neurais do tipo Transformer, esses modelos, como o *Generative Pre-trained Transformer* (GPT), são treinados em grandes quantidades de dados textuais e são capazes de prever a próxima palavra em uma sequência, permitindo-lhes criar sentenças, parágrafos ou até mesmo textos longos que imitam a linguagem humana. Esses modelos têm aplicações em diversas áreas, desde a criação de conteúdo automatizado até a tradução de idiomas, análise de sentimentos e a interação em chatbots.

Um elemento relevante dos modelos de linguagem generativos é o conceito de "temperatura." A temperatura é um parâmetro que controla a aleatoriedade das previsões do modelo. Valores mais baixos (próximos de zero) fazem com que o modelo produza saídas mais determinísticas e conservadoras, enquanto que valores mais altos tornam a saída mais diversificada e criativa. Por exemplo, em uma tarefa onde a precisão é importante, uma temperatura baixa pode ser preferível, enquanto uma tarefa que requer criatividade pode se beneficiar de uma temperatura mais alta.

Além disso, esses modelos podem ser usados em configurações de prompt *zero-shot* e *few-shot*. No modo *zero-shot*, o modelo realiza uma tarefa sem ter sido explicitamente treinado nela, baseando-se apenas no conhecimento adquirido em seu pré-treino e no contexto fornecido pelo prompt. Isso é particularmente útil quando não há dados específicos de treinamento disponíveis para uma tarefa. Já no modo *few-shot*, o modelo recebe alguns exemplos no prompt antes de realizar a tarefa, o que pode ajudar a melhorar a precisão da resposta, permitindo que o modelo ajuste sua compreensão da tarefa com base nesses exemplos. Esses conceitos são fundamentais para explorar o potencial dos modelos de linguagem generativos em uma ampla gama de aplicações.

3. Trabalhos Relacionados

Neste trabalho, investiga-se a capacidade das atuais linguagens generativas em lidar com tarefas clássicas de análise de sentimentos em uma extensa variedade de conjuntos de dados em português do Brasil. Além disso, investiga-se o uso dessas linguagens generativas na tarefa de anotação de dados em substituição ao trabalho humano.

3.1. Tarefas de Análise de Sentimentos

O tema de análise de sentimentos tem despertado o interesse de muitos pesquisadores ao longo dos últimos anos, especialmente para o idioma inglês. Nesse contexto, os trabalhos de Wankhade et al. (2022) e de Tan et al. (2023) fornecem boas referências e um panorama abrangente do estado da arte da área. Apesar de ser menos explorada que para a língua inglesa, a análise de sentimentos em português tem recebido atenção crescente nos últimos anos. Diversos trabalhos têm investigado as principais tarefas envolvidas, como a identificação de subjetividade (de Oliveira & de Melo, 2021), classificação de polaridade (Oliveira & de Melo, 2020) e identificação de sentenças comparativas (Kansaon et al., 2020). Para a tarefa de identificação de polaridade, técnicas que utilizam abordagens baseadas em léxicos (Lazarini et al., 2023; de Melo, 2022), aprendizado de máquina supervisionado (Aires et al., 2018) e, mais recentemente, modelos de linguagem pré-treinados, têm sido amplamente estudadas (Lopes et al., 2021).

Outro aspecto relevante é a tarefa de identificação de subjetividade, que foca em distinguir frases opinativas de factuais. Esta tarefa é essencial para compreender a natureza da opinião expressa em textos. Estudos como o de Oliveira & de Melo (2020) forneceram uma análise pormenorizada sobre a tarefa e conjuntos de dados específicos para o português, auxiliando no desenvolvimento de técnicas mais eficazes para essa tarefa. Adicionalmente, trabalhos sobre a identificação de sentenças comparativas, como o de Kansaon et al. (2020)), exploram a classificação de frases que contêm comparações explícitas ou implícitas entre dois ou mais itens. Esse tipo de análise é particularmente relevante em contextos de comércio eletrônico e avaliações de produtos.

Este trabalho propõe avançar no tema avaliando o impacto de modelos mais recentes, particularmente os modelos de linguagem de larga escala (LLMs), que podem trazer vantagens ao permitir a realização de tarefas de PLN com baixo custo de treino.

3.2. ChatGPT

ChatGPT[4] é um modelo de linguagem generativo desenvolvido pela OpenAI, baseado na arquitetura GPT-3.5, que pode gerar textos coerentes e contextualmente relevantes a partir de um *prompt* (Brown et al., 2020). A OpenAI lançou em setembro de 2022 o ChatGPT 3.5 Turbo e em março de 2023 a versão ChatGPT 4. De acordo com a literatura (Rudolph et al., 2023), o ChatGPT 3.5 Turbo possui 175 bilhões de parâmetros, enquanto a versão ChatGPT 4 possui estimativa de 1,7 trilhões de parâmetros (Schreiner, 2024). Segundo a OpenAI, o ChatGPT pode realizar várias tarefas, como gerar respostas a perguntas, e sumarizar e traduzir textos sem nenhum treinamento adicional. O modelo foi treinado em um grande *corpus* de texto de várias fontes, incluindo livros, artigos e sites.

Desde seu surgimento, vários trabalhos da literatura têm abordado o uso do ChatGPT em tarefas de PLN. De forma geral, explora-se a capacidade do modelo de linguagem de entender e gerar textos por meio de prompts, onde comandos são direcionados a, por exemplo, classificar textos. Esse tipo de pesquisa pode ser observado no trabalho de Zhao et al. (2023), onde tal modelo mostra-se competitivo aos métodos tradicionais para classificar textos relacionados à área de agricultura. De forma similar, Loukas et al. (2023) mostram desempenho similar a modelos não generativos, a custos muito menores, para classificação de textos na área de finanças.

No campo da análise de sentimentos, encontram-se trabalhos mostrando a capacidade desses modelos, como, por exemplo, o de Fatouros et al. (2023), onde a abordagem *zero-shot* do ChatGPT superou até mesmo um modelo especializado na área de finanças, o FinBERT. Já Belal et al. (2023) mostram que, considerando tweets em inglês, o ChatGPT supera técnicas baseadas em análise léxica na tarefa de anotação de dados de sentimentos.

Na tarefa de tradução automática, Com o lançamento do GPT-4, o desempenho do ChatGPT melhorou significativamente, tornando-se comparável a produtos comerciais de tradução, mesmo para idiomas distantes (Jiao et al., 2023).

Várias aplicações de chatbots inteligentes surgiram em diferentes áreas, mostrando, com alguns cuidados, resultados e vantagens significativas (Bahrini et al., 2023). Por exemplo, Sallam et al. (2023) lista as seguintes vantagens da integração do ChatGPT no processo educacional médico: melhoria na aprendizagem personalizada, melhoria no raciocínio clínico e assistência para entender conceitos médicos complexos.

[4] https://openai.com/blog/ChatGPT

3.3. Modelos Generativos em Português

Recentemente, a pesquisa sobre Modelos de Linguagem de Grande Escala (LLMs) tem avançado significativamente. Enquanto muitos dos modelos de linguagem foram treinados em múltiplos idiomas, visando a abrangência e a flexibilidade, estudos recentes têm demonstrado que o pré-treinamento monolíngue pode trazer benefícios substanciais em termos de desempenho para tarefas específicas de um determinado idioma (Pires et al., 2023). Essa abordagem permite que o modelo capture melhor as nuances linguísticas e culturais inerentes à língua alvo. Em particular, para o português do Brasil, a utilização de LLMs otimizados para a língua, como o Sabiá-2, tem mostrado resultados promissores, superando modelos multilíngues em diversas avaliações (Almeida et al., 2024). Particularmente, a arquitetura do Sabiá consiste de modelos Transformers similares ao ChatGPT, no entanto, treinados em conjuntos de dados em português. Este trabalho investiga essa tendência, explorando o potencial e as vantagens de empregar LLMs monolíngues em tarefas de análise de sentimentos no contexto brasileiro, comparando seu desempenho com modelos multilíngues como GPT-3.5 e GPT-4.

Para o objetivo descrito acima, foi escolhido o Sabiá-2 devido às suas capacidades linguísticas e de custo-benefício. O modelo foi desenvolvido especificamente para português e treinado em uma vasta gama de textos no idioma, resultando em uma compreensão mais profunda das nuances linguísticas. Isso é fundamental para tarefas de análise de sentimentos, onde a interpretação correta de emoções e sentimentos é crucial. Os resultados do Sabiá-2, conforme descrito por Almeida et al. (2024), mostraram que ele supera ou iguala modelos avançados como GPT-3.5 e GPT-4 em diversos exames, destacando sua eficácia. O modelo Sabiá-2 está disponível, seguindo padrão disponibilizado pela OpenAI, através de uma API nomeada de MariTalk.

A escolha do Sabiá-2 também se justifica pela sua vantagem econômica. O Sabiá-2 Medium, por exemplo, oferece um desempenho comparável ao GPT-4 a um custo significativamente menor por *token*. Esta característica torna o Sabiá-2 uma opção viável para aplicações de análise de sentimentos em larga escala, onde a relação custo-eficácia é um fator crucial. Além disso, o foco do Sabiá-2 em um corpus monolíngue e es-

pecializado no português permite uma integração mais eficiente de conhecimentos específicos do domínio, o que é essencial para capturar corretamente as nuances culturais e contextuais presentes nas expressões sentimentais do português brasileiro. Em suma, o Sabiá-2 não só pode proporcionar uma análise de sentimentos mais precisa, mas também o faz de maneira mais econômica, justificando assim sua escolha para este estudo.

3.4. Anotadores

Em aplicações de PLN, o uso de dados rotulados é frequentemente necessário, o que envolve o processo manual de anotação de dados. Tradicionalmente, duas estratégias principais têm sido empregadas para esse fim. Primeiramente, pesquisadores podem recrutar e treinar codificadores, como assistentes de pesquisa, para realizar a tarefa de anotação. Uma outra forma de anotação consiste em confiar nas pessoas que se disponibilizam a realizar essa tarefa em plataformas como Amazon Mechanical Turk (MTurk) para anotar os dados (Gilardi et al., 2023).

Em uma análise recente (Gilardi et al., 2023), foi demonstrado que o ChatGPT superou trabalhadores humanos na anotação de texto em várias tarefas. Além disso, outros estudos (Ding et al., 2022) mostraram que o desempenho dos modelos ChatGPT é ligeiramente inferior quando comparado a dados rotulados por humanos. No entanto, a utilização de modelos ChatGPT reduz significativamente o custo e o tempo necessários para o processo de anotação em comparação com a dependência exclusiva de anotadores humanos.

Particularmente, trabalhos como os apresentados por Qin et al. (2023) têm objetivos semelhantes aos objetivos desta pesquisa; no entanto, eles estão principalmente focados na língua inglesa. Em contraste, este trabalho fornece uma contribuição adicional ao avaliar o desempenho de modelos ChatGPT em textos em português.

Essas descobertas indicam que o ChatGPT apresenta capacidades promissoras em realizar a tarefa de anotação de dados de texto com muitas vantagens, como desempenho ou custos, quando comparado à dependência exclusiva de anotadores humanos. Por essas razões, decidiu-se investigar o uso do ChatGPT na geração automática de dados de treinamento (QP2).

4. Metodologia

O objetivo principal deste estudo é investigar o potencial de generalização dos modelos de linguagem de grande escala (LLMs) em diversas tarefas de análise de sentimentos, especificamente no contexto do português brasileiro. Esta pesquisa se concentra em duas questões principais.

A primeira questão de pesquisa (QP1) busca validar empiricamente o desempenho dos modelos de linguagem escolhidos como solucionadores competentes para tarefas relevantes de análise de sentimentos. Para validar esta questão de pesquisa, foram conduzidas avaliações em três tarefas cruciais de análise de sentimentos descritas a seguir. Além de adotar a abordagem *zero-shot* de usar os modelos de linguagem para classificação dos textos, foi avaliado o impacto da abordagem *few-shot* de engenharia de prompt.

A primeira tarefa (Tarefa 1) consiste em classificar sentenças como factuais ou opinativas. O design do prompt para essa tarefa é mostrado na Figura 1 (a). Por exemplo, a frase "o restaurante tem um ambiente agradável" seria classificada como opinativa, enquanto a sentença "o restaurante abre às 14 horas" seria classificada como factual. Este estudo adotou como base a metodologia descrita por de Oliveira & de Melo (2021) e também utilizou os conjuntos de dados disponibilizados pelos autores desse artigo. No trabalho de de Oliveira & de Melo (2021), os autores analisaram diversos modelos clássicos de aprendizado de máquina, e o modelo GBT alcançou o melhor resultado ao utilizar a classificação de tags de partes do discurso (POS tags) das palavras, sendo o número de adjetivos a principal característica.

A segunda tarefa (Tarefa 2) tem como objetivo principal classificar cada sentença como possuindo sentimento positivo ou negativo. O *design* do prompt para esta tarefa é mostrado na Figura 1 (b). A frase "a comida estava deliciosa" expressa um sentimento positivo, enquanto "o preço era muito salgado" transmite um sentimento negativo sobre o preço do restaurante. As metodologias elaboradas por Oliveira & de Melo (2020) foram empregadas como base para esta tarefa, e os conjuntos de dados publicados pelos respectivos autores também foram utilizados. No trabalho de Oliveira & de Melo (2020), os autores compararam diversos métodos que representavam o estado da arte no tratamento de cálculo de polaridade de texto em português com um classificador GBT, utilizando as seguintes características: número de palavras, adjetivos, substantivos, advérbios, superlativos, comparativos e termos modificadores de substantivos. O classificador GBT apresentou resultado superior e será adotado como modelo de referência (*baseline*) nesta tarefa. Neste trabalho, assim como no trabalho de Oliveira & de Melo (2020)) foram consi-

deradas apenas os sentimentos positivos e negativos. Apesar de alguns trabalhos na literatura considerarem a classificação de sentenças neutras, estas costumam ter uma menor relevância (Wankhade et al., 2022), pois o maior interesse é na descoberta de quais sentenças transmitem um sentimento positivo ou negativo.

A terceira tarefa (Tarefa 3) consiste em classificar sentenças como comparativas ou diretas. O *design* do prompt para esta tarefa é mostrado na Figura 1 (c). Por exemplo, a frase "o restaurante tem um ambiente agradável" é uma frase direta, enquanto a frase "o sorvete da McDonald's é melhor" é comparativa. Os métodos descritos por Kansaon et al. (2020) serviram como base para esta tarefa, e os conjuntos de dados publicados pelos autores também foram utilizados. Neste trabalho, os autores analisaram diversos modelos clássicos de aprendizado de máquina aplicados em combinação com três diferentes características textuais: Frequência do Termo e da Frequência Inversa do Documento (TF-IDF[5]) de palavras, TF-IDF de bigramas de palavras e TF-IDF de trigramas de palavras, sendo que o modelo Naive Bayes (NB) alcançou o melhor resultado.

A segunda questão de pesquisa (QP2) tem como objetivo validar a viabilidade do uso de modelos ChatGPT para automatizar a rotulagem de conjuntos de dados. Inicialmente, foi empregado o ChatGPT para rotular os dados obtidos na QP1. Em seguida, os dados rotulados pelo ChatGPT foram utilizados para treinar modelos usando o AutoGluon (Erickson et al., 2020). Por fim, os resultados obtidos desses modelos foram comparados, tanto com *baselines* da literatura quanto com o próprio ChatGPT, para avaliar seus desempenhos e eficácia.

4.1. Exploração de Modelos de Linguagem

A OpenAI disponibiliza uma ampla gama de modelos através de sua API, cada um voltado para propósitos e *benchmarks* de performance distintos. Neste estudo, foram avaliados o GPT 3.5-Turbo, o Modelo de Linguagem Grande (LLM) com 175 bilhões de parâmetros, que também é a base do ChatGPT *online* - referenciado como ChatGPT 3.5 e em sua versão mais recente e comercial, o ChatGPT 4.0. Estes modelos se destacam pelo pioreirismo e revolução causada na área de Processamento de Linguagem Natural, e são otimizados para funcionalidades de chat, tornando-os ideal para tarefas centradas na interação por diálogo. Os experimentos foram realizados através da API oficial da OpenAI, com os mesmos parâmetros e versão do modelo, a menos que especificados de outra forma.

Para avaliar o impacto do parâmetro de temperatura dos modelos, que controla o grau de aleatoriedade da saída do modelo, as tarefas foram executadas com o valor de 0, que implica maior determinismo, e também com o valor de 1.0, que implica maior aleatoriedade. Conforme observado por Gilardi et al. (2023), o uso de valores de temperatura mais baixos resulta em melhores resultados do ChatGPT na tarefa de análise de sentimentos.

4.2. Prompts

De acordo com Liu et al. (2023), um prompt funciona como um conjunto de instruções fornecidas a um modelo de linguagem, efetivamente programando-o por meio da customização, aprimoramento ou refinamento de suas capacidades. Selecionar um prompt adequado é essencial para que o modelo forneça a resposta desejada com precisão. Inicialmente, foram experimentados prompts com instruções mais detalhadas, mas se observou que prompts com instruções diretas apresentaram melhores resultados. Abaixo, apresenta-se o prompt selecionado para cada tarefa seguindo a abordagem *zero-shot* descrita na Figura 1.

Para a Tarefa 1, foi escolhido o seguinte prompt: *Classifique a sentença* "`FRASE`" *em factual ou opinativa. Responda somente factual ou opinativa*, onde a frase que se quer avaliar fica entre aspas simples. Com este prompt, espera-se que o modelo responda apenas com as palavras "*factual*" ou "*opinativa*."

Para a Tarefa 2, foi escolhido o seguinte prompt: *Classifique a sentença* "`FRASE`" *em positiva ou negativa. Responda somente positiva ou negativa*, onde a frase que se quer avaliar fica entre aspas simples. Com este prompt, espera-se que o modelo responda apenas com a palavra "*positiva*" ou "*negativa*."

Por fim, para a Tarefa 3, foi escolhido o seguinte prompt: *Classifique a sentença* "`FRASE`" *em comparativa ou não comparativa. Responda somente comparativa ou não comparativa*, onde a frase que se quer avaliar fica entre aspas simples. Com este prompt, espera-se que o modelo responda apenas com a palavra "*comparativa*" ou "*não comparativa*."

[5]Em inglês, TF-IDF significa *Term Frequency-Inverse Document Frequency*.

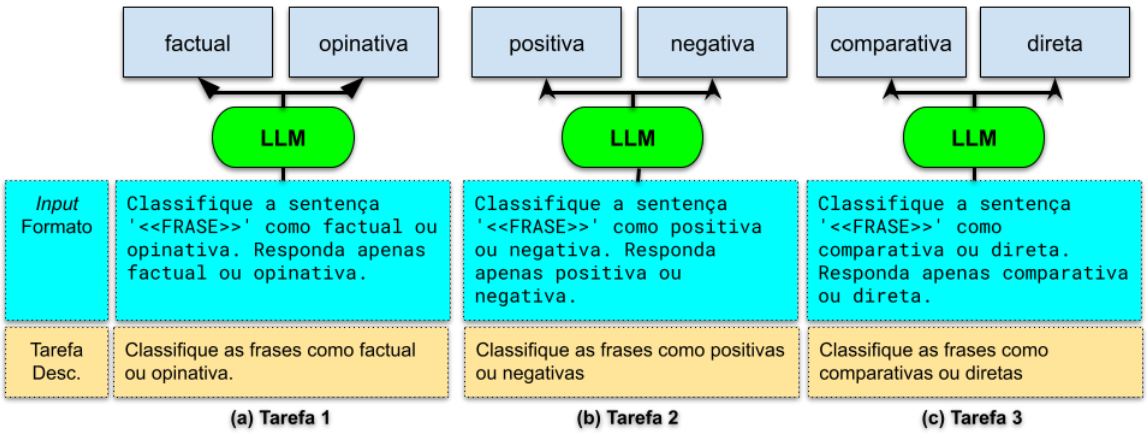

Figura 1: Design de prompts *zero-shot* (ZS).

Para a avaliação dos modelos seguindo a abordagem *few-shot*, adotou-se o *design* de prompt da Figura 2. Nela, as chaves "`classe_1`" ou "`classe_2`" são substituídas conforme as tarefas descritas anteriormente. A chave "`sentença`" é a frase sendo avaliada. E as chaves "`exemplo_classe_1`" e "`exemplo_classe_2`" são frases escolhidas aleatoriamente dentre os datasets avaliados, sendo uma de cada classe para a respectiva tarefa. As frases de exemplo são apresentados nas tabelas 1, 2 e 3, conforme *datasets* usados e melhor descritos na próxima seção.

5. Experimentos

Nesta seção, detalha-se o ambiente de avaliação, o que inclui a descrição dos dados utilizados (*dataset*) e as métricas de avaliação adotadas. Em seguida, os resultados experimentais são apresentados e discutidos.

5.1. Datasets

Diversos conjuntos de dados foram empregados para testar a robustez das linguagens generativas em face de diversos desafios linguísticos e contextuais inerentes ao português brasileiro, garantindo uma validação abrangente para tarefas variadas de análise de sentimentos e alinhamento com *benchmarks*.

Para a Tarefa 1, foram utilizados três conjuntos de dados distintos compostos por sentenças factuais e subjetivas. Os detalhes de cada conjunto de dados empregados na Tarefa 1 são apresentados na Tabela 4. ReLi[6] consiste em uma coleção de resenhas de livros em português, recuperadas da internet e anotadas manualmente (Freitas et al., 2012). Neste trabalho, foi utilizado uma quantidade menor de instâncias do ReLi, pois representa o conjunto de dados empregados nos experimentos do artigo adotado como referência (Oliveira & de Melo, 2020). TA-Restaurantes[7] contém sentenças em português relacionadas a comentários de restaurantes coletados do TripAdvisor[8] (Oliveira & de Melo, 2020). Computer-BR[9] é um conjunto de tuítes em português e abrange uma ampla gama de tópicos relacionados a computadores (Moraes et al., 2016).

Para a Tarefa 2, foram utilizados os mesmos conjuntos de dados da Tarefa 1. No entanto, anotações adicionais para polaridade de sentimento (positiva ou negativa) foram adicionadas. Além disso, foi considerando também o *corpus* do Google Play[10] (Stiilpen Junior & Merschmann, 2016)). Este *corpus* consiste em 1.630 frases selecionadas aleatoriamente de um conjunto original de 10.000 avaliações de aplicativos móveis na Google Play Store. As frases do *corpus* do Google Play são divididas igualmente entre sentimentos positivos e negativos. Os detalhes de cada conjunto de dados empregados na Tarefa 2 são apresentados na Tabela 5.

Para a Tarefa 3, foram utilizados dois conjunto de dados sumarizados na Tabela 6. O Twitter é um *corpus* de frases comparativas relacionadas a produtos eletrônicos (Kansaon et al., 2020) e o Buscapé consiste em avaliações de produtos coletadas do site Buscapé[11] (Kansaon et al., 2020). Os conjuntos de dados são anotados como sentenças comparativas ou diretas.[12]

[6] https://www.linguateca.pt/Repositorio/ReLi

[7] https://data.mendeley.com/datasets/hsn6g3dbsk

[8] https://www.tripadvisor.com.br

[9] http://tiagodemelo.info/datasets.html

[10] http://tiagodemelo.info/datasets.html

[11] https://www.buscape.com.br

[12] https://zenodo.org/records/4124410

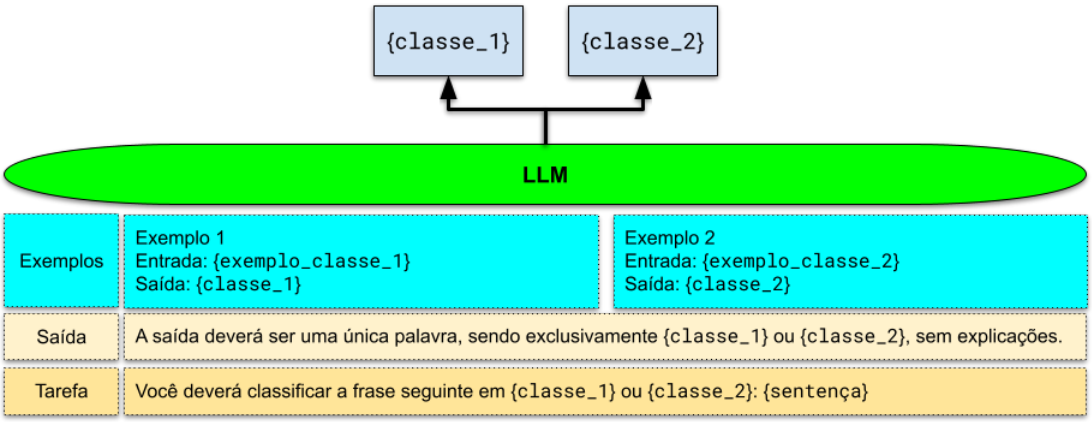

Figura 2: Design de prompts *few-shot* (FS).

Dataset	Exemplos
ReLi	factual: "Indicado para crianças e pré-adolescentes!" opinativa: "Não tenho palavras para explicar o quanto eu amei a experiência de ler finale e todo hush hush. Vou guardar essa história comigo para sempre."
TA-Restaurants	factual: "Vários tipos de molho regional, com tucupi e outros." opinativa: "Os pratos, petiscos e bebidas são nota 10!"
Computer-Br	factual: "Notebook Dell Vostro V14T-5470-B30 Intel Core i5 4GB (2GB de Memória Dedicada) 500GB LED 14"Touchscreen Windows 8... http://fb.me/J7liX8dJ" opinativa: "Amooo meu notebook #Dell #corei7 ????"

Tabela 1: Exemplos usados em prompts *few-shot* para a Tarefa 1.

Dataset	Exemplos
ReLi	positiva: "Eu adorei esse livro quando li, foi um conselho de um garoto que é escritor eu conheci em a internet, foi ótimo." negativa: "Acho que este livro não merece nem mesmo uma estrela."
TA-Restaurants	positiva: "Ambiente amplo, confortável e muito agradável." negativa: "atendimento deixa a desejar."
Computer-Br	positiva:" Selok meo not é mol rapido kk DELL é zikaa" negativa: "Reclame Aqui - Dell Computadores do Brasil - Dell erra na entrega e não cumpre prazo http://fb.me/6oW8KuPSt"
Google Play	positiva: "Execelente Muito bom esse jogo!!! Adoreei...recomendo." negativa: "Max O grafico é ruim"

Tabela 2: Exemplos usados em prompts *few-shot* para a Tarefa 2.

Dataset	Exemplos
Buscapé	não-comparativa: "completa tem suas vantagens mas fui lezado comprei esta tv no dia 26"' comparativa: "tenho uma brastemp 11 kg bwa11 que é muito inferior a essa consul."'
Twitter	não-comparativa: "já tentei assistir black mirror três vezes mas paro no 2 ep pq fico sem entender nada!"' comparativa: "a glock ou vc escolho a glock"'

Tabela 3: Exemplos usados em prompts *few-shot* para a Tarefa 3.

	Factual	Subjectiva	Total
ReLi	175	175	350
TA-Restaurantes	591	458	1.049
Computer-BR	604	1.677	2.281

Tabela 4: *Dataset* para a Tarefa 1.

	Positivo	Negativo	Total
ReLi	85	85	170
TA-Restaurantes	505	56	561
Computer-BR	198	400	598
Google Play	815	815	1.630

Tabela 5: *Dataset* para a Tarefa 2.

	Direta	Comparativa	Total
Buscapé	1.282	1.472	2.754
Twitter	918	1.135	2.053

Tabela 6: *Dataset* para a Tarefa 3.

5.2. Métricas de Avaliação

Para avaliar os modelos nas tarefas investigadas neste artigo, foram utilizadas as métricas de precisão (P), revocação (R) e F-measure (F1) (Baeza-Yates & Ribeiro-Neto, 1999). Seja A o conjunto de respostas corretas, de acordo com um conjunto de referência, e seja B o conjunto de respostas produzidas pelo método que está sendo avaliado. Definiu-se precisão (P), revocação (R) e F-score (F_1) como:

$$P = \frac{|A \cap B|}{|B|} \quad R = \frac{|A \cap B|}{|A|} \quad F_1 = \frac{2 \times (P \times R)}{P + R}$$

5.3. Resultados

Nesta seção, são apresentados os resultados das questões de pesquisa levantadas QP1 e QP2 para os diferentes *datasets* e modelos das três tarefas.

5.3.1. Questão de Pesquisa 1 (QP1)

Inicialmente, foi avaliada a influência do hiperparâmetro temperatura no desempenho em todas as tarefas. Para isso, foi escolhido o ChatGPT 3.5 como referência, por ser o modelo de acesso livre e amplamente utilizado em pesquisas com LLMs. Foram consideradas a temperatura de 0, onde o modelo é totalmente determinístico, e a temperatura de 1, onde o modelo gera respostas mais criativas. A Figura 3 exibe os valores do score F1 para as diferentes tarefas (em cores diferentes) e para cada conjunto de dados de uma determinada tarefa. É importante notar que o modelo com temperatura 0 produziu resultados melhores ou, no mínimo, iguais ao modelo com temperatura 1. A justificativa para isso é que o objetivo da classificação de texto é produzir uma saída única para uma dada entrada. Portanto, a liberdade de escolher respostas mais variadas e criativas tende a gerar resultados piores em tarefas de classificação de texto.

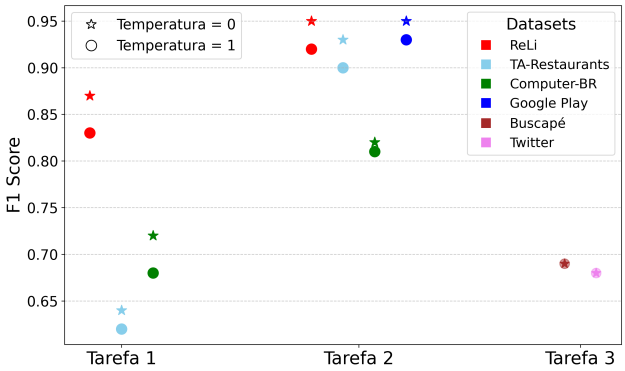

Figura 3: Desempenho do ChatGPT 3.5 conforme hiperparâmetro de temperatura.

Os resultados para todas as tarefas, descritos a seguir, têm os modelos com temperatura igual a zero, e são comparados com métodos da literatura em cada uma das tarefas avaliadas. Os modelos configurados na abordagem *zero-shot* estão representados por ZS e os modelos configurados na abordagem *few-shot* estão representados por FS.

Os resultados para a Tarefa 1 (identificação de subjetividade) são apresentados na Tabela 7. A análise mostra que, na configuracão de prompt *zero-shot* (indicado por ZS nos modelos) e no *dataset* ReLi, os modelos de linguagem ChatGPT 4, ChatGPT 3.5 e MariTalk alcançaram, nesta ordem, resultados muito próximos ao GBT, que é o estado da arte para esta tarefa. Todos os modelos de linguagem apresentaram desempenho inferior no *dataset* TA-Restaurantes, com destaque negativo para o MariTalk. Já no *dataset* Computer-BR, a situação se inverteu, onde todos os modelos de linguagem superaram o GBT, com destaque positivo para o MariTalk. Considerando a metodologia de prompt *few-shot* (indicado por FS nos modelos), pode-se observar que não houve ganhos nos modelos ChatGPT 3.5 e ChatGPT 4 em relação às suas versões *zero-shot*, mas o desempenho do MariTalk subiu consideravelmente, superando o GBT nos *datasets* ReLi e Computer-BR.

	ReLi			**TA-Restaurantes**			**Computer-BR**		
	P	*R*	F_1	*P*	*R*	F_1	*P*	*R*	F_1
GBT	0,76	0,68	0,71	**0,71**	**0,91**	**0,80**	0,39	0,34	0,36
ChatGPT 3.5$_{ZS}$	0,68	0,68	0,68	0,63	0,63	0,63	0,54	0,54	0,54
ChatGPT 4$_{ZS}$	0,69	0,69	0,69	0,64	0,64	0,64	0,58	0,58	0,58
MariTalk$_{ZS}$	**0,82**	0,57	0,67	0,47	0,39	0,43	0,78	0,73	0,75
ChatGPT 3,5$_{FS}$	0,58	0,58	0,58	0,59	0,59	0,59	0,32	0,32	0,32
ChatGPT 4$_{FS}$	0,63	0,63	0,63	0,62	0,62	0,62	0,54	0,54	0,54
MariTalk$_{FS}$	**0,82**	**0,79**	**0,81**	0,47	0,46	0,46	**0,79**	**0,76**	**0,78**

Tabela 7: Identificação de subjetividade (Tarefa 1).

Em relação à abordagem de prompt *few-shot* (FS), percebe-se que houve uma piora nos modelos ChatGPT 3.5 e 4 em relação às versões *zero-shot* (ZS). Isso pode ser explicado pelo aumento do tamanho dos prompts, causando mais dificuldades de interpretação dos mesmos sem agregar novos conhecimentos. No entanto, o modelo MariTalk se beneficiou pelo prompt com exemplos da tarefa. De fato, observou-se que o MariTalk na versão ZS produziu muitos erros de formatação em sua saída (palavras ou formatações diferentes das anotações de saída solicitadas, causando erros de classificação). Isso se observa pela diferença entre as métricas de precisão e revocação em todos os *datasets*, indicando a ocorrência de muitos falsos negativos. Neste modelo, a abordagem *few-shot* ajudou a produzir saídas melhor formatadas, melhorando o resultado geral e aproximando as métricas de precisão e revocação do modelo.

Em relação às diferenças de desempenho nos diferentes *datasets*, nota-se que tanto o ReLi quanto o TA-Restaurantes são compostos por textos mais descritivos e formais se comparados ao Computer-BR, que é composto por tuítes. Esses tuítes são, geralmente, escritos de forma abreviada, utilizando jargões ou linguagens coloquiais. Assim, pode-se ver que o modelo da literatura, treinado especificamente para o *dataset*, teve um desempenho muito melhor nos dois primeiros casos. No entanto, os modelos de linguagem mostraram-se mais resilientes aos dados ruidosos do último conjunto. Ao se analisar os resultados dos experimentos, percebe-se que as versões do ChatGPT podem não entender bem a subjetividade de uma frase na maioria dos casos, mas é muito mais capaz de lidar com diferentes tipos de texto devido à grande quantidade de dados e a diversidade de textos usados em seu treinamento. Já o MariTalk, parece ter se beneficiado de sua base de treinamento de textos em português, mostrando superioridade evidente no ReLi e no Computer-BR.

Os resultados para a Tarefa 2 (identificação de polaridade) são apresentados na Tabela 8. Os dados tabulados mostram que os modelos de linguagem alcançaram resultados significativamente melhores que o modelo GBT, que representa o estado da arte, nos conjuntos de dados ReLi, Computer-BR e Google Play, enquanto apresentaram um F1-score similar no TA-Restaurants. O ChatGPT 4 foi melhor que os demais, obtendo os maiores valores de F1-score em sua versão *zero-shot*. Já o MariTalk perdeu para as versões do ChatGPT na maioria dos *datasets*, ficando com desempenho bem próximo no TA-Restaurants. A abordagem de prompts *few-shot* apresentou pouca vantagem na maioria dos casos, tendo ajudado levemente o desempenho do ChatGPT 3.5 nos *datasets* Computer-BR e Google Play, mas ao contrário da Tarefa 1, prejudicando mais o MariTalk, como observado no F1 desses mesmos dados.

Os resultados sugerem que os modelos de linguagem e, em especial, o ChatGPT, são altamente capazes de determinar a polaridade das sentenças. Mesmo não sendo ajustados para esses conjuntos de dados específicos, é plausível que a análise de sentimentos e polaridade seja comum nos diversos textos utilizados para o treinamento desses modelos. Por exemplo, é esperado que textos de conversas e literatura abordem a positividade ou não das ideias muito mais do que a subjetividade. Além disso, o treinamento dos ChatGPTs incorporou avaliações de usuários relacionadas a produtos e serviços de diversas plataformas. Esse *feedback* geralmente inclui um sistema de classificação por estrelas: comentários com 1 ou 2 estrelas são interpretados como negativos, enquanto comnentários com 4 ou 5 estrelas são positivos. Isso permite aos modelos discernir efetivamente a polaridade de termos e frases dentro dessas avaliações. Essas observações podem ajudar a esclarecer a dificuldade maior dos modelos de linguagem na Tarefa 1. Por último, prompts que buscam o sen-

	Reli			**TA-Restaurants**			**Computer-BR**			**Google Play**		
	P	*R*	F_1	*P*	*R*	F_1	*P*	*R*	F_1	*P*	*R*	F_1
GBT	0,47	0,64	0,59	0,90	**0,99**	**0,95**	0,44	0,44	0,44	0,69	0,68	0,69
ChatGPT 3.5$_{ZS}$	**0,96**	**0,96**	**0,96**	0,93	0,93	0,93	0,82	0,82	0,82	0,95	0,95	0,95
ChatGPT 4$_{ZS}$	**0,96**	**0,96**	**0,96**	**0,94**	0,94	0,94	**0,92**	**0,92**	**0,92**	**0,98**	**0,98**	**0,98**
MariTalk$_{ZS}$	0,85	0,85	0,85	**0,94**	0,94	0,94	0,89	0,89	0,89	0,94	0,94	0,94
ChatGPT 3.5$_{FS}$	0,95	0,95	0,95	0,93	0,93	0,93	0,89	0,89	0,89	0,97	0,97	0,97
ChatGPT 4$_{FS}$	**0,96**	**0,96**	**0,96**	**0,94**	0,94	0,94	0,91	0,91	0,91	0,97	0,97	0,97
MariTalk$_{FS}$	0,82	0,82	0,82	0,93	0,93	0,93	0,69	0,69	0,69	0,82	0,82	0,82

Tabela 8: Identificação de polaridade (Tarefa 2).

timento do texto tendem a ser mais diretos do que aqueles que investigam a subjetividade (factual ou opinativo). Essa clareza intrínseca nos prompts de sentimento pode reduzir as chances de interpretações equivocadas pelo modelo.

Os resultados para a Tarefa 3 (identificação de frases comparativas) são apresentados na Tabela 9. Todos os modelos de linguagem apresentaram desempenho inferior ao método NB, que é o estado-da-arte. Mas o ChatGPT 4 apresentou um desempenho superior em relação aos demais modelos de linguagem, mostrando uma evolução de entendimento em relação à sua versão anterior. Já o MariTalk, mesmo treinado em português, teve desempenho similar ao ChatGPT 3.5. Novamente, a abordagem *few-shot* não se mostrou significativa, tendo ajudado levemente o ChatGPT 4 no *dataset* Twitter, mas deteriorando o desempenho dos modelos nos demais casos.

	Buscape			**Twitter**		
	P	*R*	F_1	*P*	*R*	F_1
NB	**0,87**	**0,88**	**0,87**	**0,86**	**0,86**	**0,86**
ChatGPT 3.5$_{ZS}$	0,67	0,67	0,67	0,61	0,61	0,61
ChatGPT 4$_{ZS}$	0,75	0,75	0,75	0,70	0,70	0,70
MariTalk$_{ZS}$	0,67	0,67	0,67	0,68	0,68	0,68
ChatGPT 3.5$_{FS}$	0,51	0,51	0,51	0,58	0,58	0,58
ChatGPT 4$_{FS}$	0,75	0,75	0,75	0,73	0,73	0,73
MariTalk$_{FS}$	0,47	0,46	0,46	0,48	0,47	0,48

Tabela 9: Identificação de sentenças comparativas (Tarefa 3).

Diferentemente das Tarefas 1 e 2, onde modelos de linguagem generativos tiveram desempenhos competitivos ou superiores, foram observadas maiores diculdades desses no reconhecimento de frases comparativas. Essa limitação possivelmente se deve ao fato desses modelos não terem sido treinado nesses conjuntos de dados específicos. Além disso, é possível que os textos comuns utilizados durante seu treinamento não apresentem frequentemente julgamentos comparativos explícitos, ponto já discutido no contexto da Tarefa 1, e que contrasta com as expectativas para a Tarefa 2. Por exemplo, frases como "acho um ótimo smartphone em relação ao seu preço com muitas funções" e "preço poderia ser mais acessível já que a Caloi é no brasil" são identificadas como frases comparativas pelo ChatGPT 3.5, apesar de não haver comparação explícita entre dois produtos.

Como pode-se observar, os modelos de linguagem demonstraram elevado desempenho na análise de sentimentos, particularmente na identificação de subjetividade e polaridade dentro das frases. Na tarefa de identificação de polaridade, o desempenho do ChatGPT 4 se destaca como o melhor no geral, sugerindo que ele pode lidar com essas tarefas de forma confiável. Para a identificação de frases comparativas, embora os modelos de linguagem não tenham obtido os melhores resultados, é possível observar uma evolução do ChatGPT em sua versão 4. Ainda, a seleção de um prompt mais adequado poderia melhorar os resultados. A adição de mais *tokens* poderia refinar ainda mais as respostas, mas isso traria a contra-partida de aumentar o custo por requisição. O MariTalk mostrou-se competitivo com o ChatGPT 3.5, ultrapassando seu desempenho em vários casos, o que evidencia o potencial de um modelo treinado exclusivamente em português para tarefas nessa língua. Os resultados experimentais indicam que o ChatGPT 4 poderia ser utilizado como um método adequado para abordar as tarefas analisadas.

Na Tabela 10, pode-se observar ainda, alguns exemplos de frases que os modelos não conseguiram classificar corretamente, algumas frases apresentam ambiguidade, o que dificultaria o desempenho do modelo. Outras frases como, por exemplo, "meu pai é fofin dms, diferente da minha mãe" demonstram que apesar de serem bons classificadores, os LLMs ainda não são capazes de entender perfeitamente as nuances do idioma português.

Datasets	Exemplos
ReLi	"Em o final, me envolvi, me vi vendo tudo que os personagens não viam" Resposta (ChatGPT 3.5): negativa Resposta esperada: positiva
TA-Restaurantes	"Sucos com frutas locais além dos tradicionais." Resposta (ChatGPT 3.5): positiva Resposta esperada: negativa
Computer-BR	"Entro no site da dell com a intenção de comprar o notebook mais barato, aí eles fazem a comparação com os mais caros..." Resposta (ChatGPT 3.5): positiva Resposta esperada: negativa
Google Play	"Bacana Bom na hora do almoço :(" Resposta (ChatGPT 4): negativa Resposta esperada: positiva
Buscapé	"outros smartphones android de parecido não possuem" Resposta (ChatGPT 3.5): comparativa Resposta esperada: não-comparativa
Twitter	"meu pai é fofin dms, diferente da minha mãe" Resposta (ChatGPT 3.5): não-comparativa Resposta esperada: comparativa

Tabela 10: Exemplos de anotações incorretas dos modelos.

Foi ainda realizada uma análise comparativa do número de *tokens* e do custo da realização dos experimentos com as linguagens generativas. A Tabela 11 apresenta a quantidade total de tokens usados nas tarefas, considerando as abordagens *zero-shot* e *few-shot*. Os valores apresentados dentro dos parênteses representam a média do número de *tokens* por sentença em cada domínio.

Tarefas	Datasets	zero-shot	few-shot
Tarefa 1	ReLi	20.574 (59)	59.043 (169)
	TA-Restaurantes	47.740 (46)	143.084 (136)
	Computer-BR	145.312 (64)	436.985 (192)
Tarefa 2	ReLi	10.205 (60)	28.552 (168)
	TA-Restaurantes	26.765 (48)	72.716 (130)
	Computer-BR	35.427 (59)	104.716 (175)
	Google Play	79.286 (49)	219.067 (134)
Tarefa 3	Buscapé	139.182 (51)	421.895 (153)
	Twitter	102.807 (50)	301.346 (147)

Tabela 11: Número de *tokens* empregados nos experimentos.

É possível observar que a abordagem *few-shot* usou cerca de três vezes mais *tokens* quando comparada com a abordagem *zero-shot*. Essa quantidade a mais resulta em um maior tempo de processamento e também em um custo mais elevado. A Tabela 12 apresenta uma estimativa de custos[13] no uso das linguagens generativas tomando-se como referências os custos apresentados nos sites do ChatGPT e MariTalk. Observa-se que apesar do modelo ChatGPT ter obtido resultados superiores na maioria das tarefas, o seu custo é mais que o triplo da sua versão 3.5 e o dobro do MariTalk. Com base nesses valores, é possível concluir que a depeneder da tarefa e do volume de dados que precisem ser processados, usar modelos anteriores ou um modelo nacional poderiam trazer resultados próximos ao estado da arte, mas com um custo bastante reduzido.

	zero-shot	few-shot
ChatGPT 3.5	1,2 dólares	3,4 dólares
ChatGPT 4	3,6 dólares	10,2 dólares
MariTalk	1,8 dólares	5,1 dólares

Tabela 12: Estimativa dos custos no uso dos LLMs.

5.3.2. Questão de Pesquisa 2

O objetivo da QP2 é verificar experimentalmente se a classificação de sentenças por modelos de linguagem pode ser usada para treinar um modelo AutoML. Os resultados apresentam a comparação entre os modelos da literatura, com o ChatGPT 4_{ZS}, na configuração de prompts *zero-shot*, que obteve os melhores resultados na avaliação anterior, e o AutoGluon. Para isso, utilizou-se o ChatGPT 4_{ZS} para anotar todos

[13]Estes custos foram calculados a partir de valores publicados no dia 06 de junho de 2024.

	Reli			**TA-Restaurantes**			**Computer-BR**		
	P	R	F_1	P	R	F_1	P	R	F_1
GBT	0.76	0.68	0.71	**0.71**	**0.91**	**0.80**	0.39	0.34	0.36
ChatGPT 4_{ZS}	0.69	0.69	0.69	0.64	0.64	0.64	0.58	0.58	0.58
AutoGluon	**0.88**	**0.86**	**0.86**	0.68	0.61	0.60	**0.73**	**0.79**	**0.72**

Tabela 13: Identificação de subjetividade (Tarefa 1) - usando ChatGPT 4_{ZS} como anotador.

	ReLi			**TA-Restaurantes**			**Computer-BR**			**Google Play**		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
GBT	0.57	0.64	0.59	0.90	**0.99**	**0.95**	0.44	0.44	0.44	0.69	0.68	0.69
ChatGPT 4_{ZS}	**0.96**	**0.96**	**0.96**	**0.94**	0.94	0.94	**0.92**	**0,92**	**0.92**	**0.98**	**0.98**	**0.98**
AutoGluon	0.90	0.86	0.87	0.76	0.64	0.68	0.91	0.81	0.84	0.93	0.92	0.92

Tabela 14: Identificação de polaridade (Tarefa 2) - usando ChatGPT$_{ZS}$ como anotador.

os datasets das três tarefas consideradas neste trabalho (seção 5.1). O AutoGluon foi treinado com esses rótulos e avaliou-se as suas métricas de treino de forma comparativa aos resultados dos demais modelos perante as anotações originais.

A Tabela 13 mostra os resultados comparativos para a identificação de subjetividade (Tarefa 1). Observa-se que o desempenho do AutoGluon nos *datasets* ReLi e Computer-BR superou o modelo GBT, considerado o estado da arte, e também foi superior ao próprio ChatGPT 4_{ZS}. No entanto, no *dataset* TA-Restaurantes, o AutoGluon apresentou desempenho inferior aos demais. Esses resultados indicam que as anotações feitas pelo ChatGPT podem ser usadas para treinar outros modelos e alcançar uma performance próxima à dele próprio ou mesmo superior. Mesmo com a possibilidade do ChatGPT 4 ter introduzido mais erros de anotações no *dataset*, ao treinar um modelo mais simples no AutoGluon, o modelo resultou em melhor generalização nos datasets Reli e Computer-BR, superando o próprio anotador.

Na Tabela 14, apresentam-se os resultados comparativos para a tarefa de identificação de polaridade (Tarefa 2). Pode-se notar que o desempenho do AutoGluon é inferior ao do ChatGPT 4_{ZS} em todos os conjuntos de dados examinados. Embora o desempenho do ChatGPT$_{ZS}$ tenha superado significativamente os padrões estabelecidos pelo modelo da literatura, a abordagem de utilizar o ChatGPT$_{ZS}$ como anotador automático para treinar o AutoGluon não funcionou tão bem quanto no caso anterior. É importante mencionar que o ChatGPT é baseado em um modelo muito grande e poderoso, treinado em uma vasta quantidade de dados textuais, então como seu desempenho foi muito superior nesta tarefa, um modelo mais simples treinado pelo AutoGluon não conseguiu superar seu desempenho. Apesar desse fato, os resultados do AutoGluon são melhores que o GBT em todos os casos, exceto para o TA-Restaurantes. Assim, pode-se concluir que o ChatGPT é uma potencial ferramenta de anotação útil em tarefas nas quais ele já apresenta um bom desempenho.

Na Tabela 15, apresentam-se os resultados comparativos para a Tarefa 3 (identificação de frases comparativas). O AutoGluon, treinado com anotações do ChatGPT 4_{ZS}, obteve um desempenho próximo a ele. Este resultado sugere que o AutoGluon conseguiu aprender efetivamente a partir das anotações fornecidas pelo modelo de linguagem. No entanto, seu desempenho foi inferior no conjunto de dados do Twitter, particularmente na pontuação F1, o que pode indicar que esse modelo teve dificuldades para generalizar em fontes de dados com conteúdo mais diversificado.

	Buscapé			**Twitter**		
	P	R	F_1	P	R	F_1
NB	**0.87**	**0.87**	**0.87**	**0.86**	**0.86**	**0.86**
ChatGPT 4	0.75	0.75	0.75	0.70	0.70	0.70
AutoGluon	0.75	0.71	0.70	0.72	0.57	0.51

Tabela 15: Identificação de sentenças comparativas (Tarefa 3) - usando ChatGPT$_{ZS}$ como anotador.

Uma possível explicação para o desempenho inferior do AutoGluon em relação ao ChatGPT pode estar relacionada às complexidades inerentes às arquiteturas dos modelos. Enquanto o ChatGPT foi extensivamente treinado em padrões linguísticos diversos e pode se adaptar a várias nuances dos dados, o AutoGluon pode

não extrapolar tão efetivamente a partir dos dados anotados apenas. Além disso, dados do Twitter, por serem mais informais e diversos, podem introduzir desafios adicionais que podem influenciar a capacidade de generalização de um modelo mais simples.

A partir dos resultados apresentados, pode-se deduzir que, mesmo com uma ligeira queda de desempenho, utilizar dados rotulados por um modelo de linguagem para treinar outros modelos de aprendizado de máquina continua sendo uma opção viável. Essa vantagem se torna particularmente evidente quando o modelo demonstra alto desempenho, como observado na análise de sentimentos de frases (Tarefa 2). Devido ao tamanho do ChatGPT 4, com estimados 1,7 trilhões de parâmetros, aproveitar suas capacidades para treinar modelos mais compactos, como o AutoGluon, pode fornecer uma vantagem significativa na implementação de soluções eficientes de aprendizado de máquina profundo.

6. Conclusões

Este artigo apresenta um estudo abrangente que investiga a eficácia de modelos de linguagem generativos em lidar com três tarefas relevantes de análise de sentimentos em português, utilizando diversos conjuntos de dados. Os resultados alcançados demonstram que os modelos ChatGPT, especialmente o GPT 4, podem ser utilizados com sucesso como modelos para identificação de polaridade em sentenças. Além disso, descobriu-se que o conjunto de dados anotado pelo ChatGPT pode ser usado para treinar modelos alternativos com impacto mínimo no desempenho, obtendo resultados comparáveis aos alcançados pelo próprio ChatGPT. Portanto, ele pode ser uma ferramenta útil quando tempo e custo são aspectos importantes na construção de modelos de aprendizado de máquina.

No entanto, para algumas outras tarefas, como identificação de subjetividade e sentenças comparativas, o ChatGPT não teve um bom desempenho como solução *zero-shot*. Sugere-se que isso ocorra devido a dois fatores: a dificuldade em construir prompts diretos e a menor ocorrência natural do assunto nos dados de treinamento do ChatGPT. Por exemplo, a identificação de sentimento em frases possui prompts mais precisos e é uma estrutura linguística muito comum em qualquer assunto textual, o que pode explicar o desempenho superior do ChatGPT nessa tarefa. Esse argumento se mostra plausível observando um melhor desempenho do MariTalk, treinado em português, especialmente na tarefa de identificação de subjetividade. De forma geral, observa-se que há um grande potencial de avanço em modelos de linguagem personalizados ao português, onde esperam-se melhores resultados conforme esses modelos ganharem capacidade e forem treinados em *datasets* maiores.

Em pesquisas futuras, há diversas vertentes a serem exploradas para aprimorar ainda mais os resultados. Uma área de foco será o aprimoramento das técnicas de engenharia de prompts para extrair resultados ainda melhores dos modelos testados. Além disso, planeja-se investigar o desempenho de outros modelos de linguagens generativas disponíveis na comunidade *Open Source*, expandindo a avaliação para abranger uma gama mais ampla de modelos e comparando sua eficácia em tarefas de análise de sentimento.

Agradecimentos

Este trabalho foi apoiado pelo Samsung Ocean Center, um programa de pesquisa e desenvolvimento da Universidade do Estado do Amazonas. Os autores também agradecem ao apoio financeiro da Fundação de Amparo à Pesquisa do Estado do Amazonas - FAPEAM através do Projeto NeuralBond (UNIVERSAL 2023 Proc. 01.02.016301.04300/2023-04).

Referências

Aires, João Paulo, Carlos Padilha, Christian Quevedo & Felipe Meneguzzi. 2018. A deep learning approach to classify aspect-level sentiment using small datasets. Em *International Joint Conference on Neural Networks (IJCNN)*, 1–8. 10.1109/IJCNN.2018.8489760

Almeida, Thales Sales, Hugo Abonizio, Rodrigo Nogueira & Ramon Pires. 2024. Sabiá-2: A new generation of Portuguese large language models. ArXiv:2403.09887 [cs.CL]. 10.48550/arXiv.2403.09887

de Araujo, Gladson, Tiago de Melo & Carlos Maurício S Figueiredo. 2024. Is ChatGPT an effective solver of sentiment analysis tasks in Portuguese? a preliminary study. Em *16th International Conference on Computational Processing of Portuguese*, 13–21.

Baeza-Yates, Ricardo & Berthier Ribeiro-Neto. 1999. *Modern information retrieval*. ACM Press, Addison-Wesley

Bahrini, Aram, Mohammadsadra Khamoshifar, Hossein Abbasimehr, Robert J. Riggs,

Maryam Esmaeili, Rastin Mastali Majdabadkohne & Morteza Pasehvar. 2023. ChatGPT: Applications, opportunities, and threats. Em *Systems and Information Engineering Design Symposium (SIEDS)*, 274–279. 10.1109/SIEDS58326.2023.10137850

Belal, Mohammad, James She & Simon Wong. 2023. Leveraging ChatGPT as text annotation tool for sentiment analysis. ArXiv [cs.CL]. 10.48550/arXiv.2306.17177

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford & Ilya Sutskever. 2020. Language models are few-shot learners. Em *Neural Information Processing Systems*, 1877–1901.

Carvalho, Caio Magno Aguiar, Hitoshi Nagano & Allan Kardec Barros. 2017. A comparative study for sentiment analysis on election Brazilian news. Em *11th Brazilian Symposium in Information and Human Language Technology*, 103–111.

Chang, Yupeng, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang & Xing Xie. 2023. A survey on evaluation of large language models. ArXiv [cs.CL]. 10.48550/arXiv.2307.03109

Christiano, Paul, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg & Dario Amodei. 2023. Deep reinforcement learning from human preferences. ArXiv [stat.ML]. 10.48550/arXiv.1706.03741

Ding, Bosheng, Chengwei Qin, Linlin Liu, Lidong Bing, Shafiq Joty & Boyang Li. 2022. Is GPT-3 a good data annotator? ArXiv [cs.CL]. 10.48550/arXiv.2212.10450

Elshawi, Radwa, Mohamed Maher & Sherif Sakr. 2019. Automated machine learning: State-of-the-art and open challenges. ArXiv [stat.ML]. 10.48550/arXiv.1906.02287

Erickson, Nick, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li & Alexander Smola. 2020. Autogluon-tabular: Robust and accurate automl for structured data. ArXiv [stat.ML]. 10.48550/arXiv.2003.06505

Fatouros, Georgios, John Soldatos, Kalliopi Kouroumali, Georgios Makridis & Dimosthenis Kyriazis. 2023. Transforming sentiment analysis in the financial domain with ChatGPT. *Machine Learning with Applications* 14. 100508. 10.1016/j.mlwa.2023.100508

Freitas, Cláudia, Eduardo Motta, R Milidiú & Juliana César. 2012. Vampiro que brilha... rá! desafios na anotaçao de opiniao em um corpus de resenhas de livros. *Encontro de Linguística de Corpus* 11. 22.

Friedman, Jerome H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* 1189–1232. 10.1214/aos/1013203451

Gilardi, Fabrizio, Meysam Alizadeh & Maël Kubli. 2023. ChatGPT outperforms crowd-workers for text-annotation tasks. Em *Proceedings of the National Academy of Sciences*, e2305016120. 10.1073/pnas.2305016120

Hayatin, Nur, Suraya Alias & Lai Po Hung. 2024. Trends and challenges in sentiment summarization: a systematic review of aspect extraction techniques. *Knowledge and Information Systems* 1–47. 10.1007/s10115-024-02075-w

Hou, Ziwei, Bahadorreza Ofoghi, Nayyar Zaidi & John Yearwood. 2024. Aspect-based fake news detection. Em *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 95–107. 10.1007/978-981-97-2266-2_8

Jiao, Wenxiang, Wenxuan Wang, Jen tse Huang, Xing Wang & Zhaopeng Tu. 2023. Is ChatGPT a good translator? Yes with GPT-4 as the engine. ArXiv [cs.CL]. 10.48550/arXiv.2301.08745

Kansaon, Daniel, Michele A Brandão, Julio CS Reis, Matheus Barbosa, Breno Matos & Fabrício Benevenuto. 2020. Mining Portuguese comparative sentences in online reviews. Em *Brazilian Symposium on Multimedia and the Web*, 333–340. 10.1145/3428658.3431081

Lazarini, Lucas, Fábio S Igarashi Anno, Eloize R Marques Seno & Helena M Caseli. 2023. Abordagens baseadas em léxicos para a classificação de sentimentos orientada aos alvos de opinião em comentários do domínio político. Em *XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, 375–380. 10.5753/stil.2023.234206

Li, Rui, Guoyin Wang & Jiwei Li. 2023. Are human-generated demonstrations necessary for in-context learning? ArXiv [cs.LG]. 10.48550/arXiv.2309.14681

Liu, Bing. 2020. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press

Liu, Pengfei, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi & Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys* 55(9). 1–35. 10.48550/arXiv.2107.13586

Lopes, Émerson, Ulisses Correa & Larissa Freitas. 2021. Exploring BERT for aspect extraction in Portuguese language. Em *The International FLAIRS Conference*, vol. 34, 10.32473/flairs.v34i1.128357

Loukas, Lefteris, Ilias Stogiannidis, Prodromos Malakasiotis & Stavros Vassos. 2023. Breaking the bank with ChatGPT: Few-shot text classification for finance. ArXiv [cs.CL]. 10.48550/arXiv.2308.14634

McCallum, Andrew & Kamal Nigam. 1998. A comparison of event models for Naive Bayes text classification. Em *Workshop on Learning for Text Categorization*, vol. 752 1, 41–48

de Melo, Tiago. 2022. SentiLexBR: An automatic methodology of building sentiment lexicons for the Portuguese language. *Journal of Information and Data Management* 13(3). 10.5753/jidm.2022.2504

de Melo, Tiago, Altigran da Silva & Edleno S de Moura. 2018. An aspect-driven method for enriching product catalogs with user opinions. *Journal of the Brazilian Computer Society* 24. 1–19. 10.1186/s13173-018-0080-4

Moraes, Silvia, André LL Santos, Matheus Redecker, Rackel M Machado & Felipe R Meneguzzi. 2016. Comparing approaches to subjectivity classification: A study on Portuguese tweets. Em *Conference on Computational Processing of the Portuguese Language*, 86–94. 10.1007/978-3-319-41552-9_8

de Oliveira, Miguel & Tiago de Melo. 2021. An empirical study of text features for identifying subjective sentences in Portuguese. Em *10th Brazilian Conference on Intelligent Systems*, 374–388. 10.1007/978-3-030-91699-2_26

Oliveira, Miguel V & Tiago de Melo. 2020. Investigating sets of linguistic features for two sentiment analysis tasks in Brazilian Portuguese web reviews. Em *XXVI Simpósio Brasileiro de Sistemas Multimídia e Web*, 45–48. 10.5753/webmedia_estendido.2020.13060

Pandey, Shubham V & AV Deorankar. 2019. A study of sentiment analysis task and it's challenges. Em *International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, 1–5. 10.1109/ICECCT.2019.8869160

Pires, Ramon, Hugo Abonizio, Thales Sales Almeida & Rodrigo Nogueira. 2023. Sabiá: Portuguese large language models. Em *Brazilian Conference on Intelligent Systems*, 226–240. 10.48550/arXiv.2304.07880

Qin, Chengwei, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga & Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? ArXiv [cs.CL]. 10.48550/arXiv.2302.06476

Rana, Toqir A & Yu-N Cheah. 2016. Aspect extraction in sentiment analysis: comparative analysis and survey. *Artificial Intelligence Review* 46(4). 459–483. 10.1007/s10462-016-9472-z

Rudolph, Jürgen, Samson Tan & Shannon Tan. 2023. ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of applied learning and teaching* 6(1). 342–363. 10.37074/jalt.2023.6.1.9

Sallam, Malik, Nesreen Salim, Muna Barakat & Alaa Al-Tammemi. 2023. ChatGPT applications in medical, dental, pharmacy, and public health education: A descriptive study highlighting the advantages and limitations. *Narra J* 3(1). e103–e103. 10.52225/narra.v3i1.103

Schreiner, Maximilian. 2024. GPT-4 architecture, datasets, costs and more leaked. Acesso: 20-072024.

Seno, Eloize, Lucas Silva, Fábio Anno, Fabiano Rocha & Helena Caseli. 2024. Aspect-based sentiment analysis in comments on political debates in Portuguese: evaluating the potential of ChatGPT. Em *16th International Conference on Computational Processing of Portuguese*, 312–320

Stiilpen Junior, Milton & Luiz Henrique C Merschmann. 2016. A methodology to handle social media posts in Brazilian Portuguese for text mining applications. Em *22nd Brazilian Symposium on Multimedia and the Web*, 239–246. 10.1145/2976796.2976845

Tan, Kian Long, Chin Poo Lee & Kian Ming Lim. 2023. A survey of sentiment analysis: Approaches, datasets, and future research. *Applied Sciences* 13(7). 4550. 10.3390/app13074550

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser & Illia Polosukhin. 2023. Attention is all you need. ArXiv [cs.CL]. 10.48550/arXiv.1706.03762

Wankhade, Mayur, Annavarapu Chandra Sekhara Rao & Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review* 55(7). 5731–5780. 10.1007/s10462-022-10144-1

Zhao, Biao, Weiqiang Jin, Javier Del Ser & Guang Yang. 2023. ChatAgri: Exploring potentials of ChatGPT on cross-linguistic agricultural text classification. *Neurocomputing* 557. 126708. 10.1016/j.neucom.2023.126708