Is ChatGPT an effective solver of sentiment analysis tasks in Portuguese? A Preliminary Study

Gladson Araújo[1], **Tiago de Melo**[1], and **Carlos Maurício**[1]

[1]Universidade do Estado do Amazonas
{gsda.eng20, tmelo, cfigueiredo}@uea.edu.br

Abstract

This paper presents an in-depth investigation into the capabilities of GPT-3.5 version for zero-shot sentiment analysis in Brazilian Portuguese, focusing on: i) identifying opinionated sentences; ii) calculating polarity; and iii) identifying comparative sentences. Results show that ChatGPT stands out in determining polarity but has challenges with subjective and comparative sentences. Despite this, we discovered that ChatGPT can be a valuable tool for annotating dataset labels, offering a practical solution for training alternative models with minimal performance impact. Representing a pioneering effort in this area, our study highlights ChatGPT's promise in Portuguese sentiment analysis and paves the way for future endeavors aimed at optimizing model efficacy and assessing other Large Language Models (LLMs) in sentiment analysis contexts.

1 Introduction

Large language models (LLMs) have showcased their ability to tackle a variety of natural language processing (NLP) tasks without the need for specific training data, a phenomenon named as zero-shot learning. This is achieved by conditioning the model with suitable prompts (Brown et al., 2020). The ability to undertake new tasks via instruction marks a significant stride towards artificial general intelligence. While contemporary LLMs exhibit commendable performance in certain scenarios, they remain prone to errors in zero-shot learning (Chang et al., 2023). Moreover, various configurations, such as temperature settings, can profoundly influence the model's effectiveness. These constraints imply that current LLMs may not truly serve as all-encompassing language systems.

The recent release of ChatGPT by OpenAI has garnered significant attention from the NLP community. ChatGPT, popular in GPT-3.5 version, is a model based on Transformer Neural Networks (Vaswani et al., 2023) trained with reinforcement learning from human feedback (RLHF) (Christiano et al., 2023). RLHF training consists of three steps: first, training a language model with self-supervised learning; second, gathering comparison data based on human preferences and training a reward model; and third, optimizing the language model against the reward model through reinforcement learning. As a result of this training, ChatGPT has demonstrated impressive capabilities such as generating high-quality responses to human input, rejecting inappropriate questions, and correcting previous errors based on subsequent conversations.

Although ChatGPT has demonstrated impressive conversational capabilities, the NLP community is still uncertain about its ability to achieve superior zero-shot generalization compared to existing LLMs, especially in languages other than English (Chang et al., 2023). Specifically, its efficacy in Brazilian Portuguese has not been thoroughly explored. To address this research gap, we conducted a comprehensive investigation into ChatGPT's zero-shot learning capacity by assessing its performance on a broad range of NLP datasets in Brazilian Portuguese, including three relevant sentiment analysis tasks: i) identification of opinionated sentences; ii) polarity calculation; and iii) identification of comparative sentences. These three tasks are important tasks in NLP regarding to problems of detecting information from comments from people's reviews for any subject, from any textual media, and mainly from Internet. Thus, these contribution can be applied to several data mining problems. More specifically, our research questions are:

Research Question 1 (RQ1): How does ChatGPT perform as a resolver for the three sentiment analysis tasks mentioned above? To address this, we will empirically compare the performance of ChatGPT against methods that are considered state

of the art.

Research Question 2 (RQ2): How does the annotation generated by ChatGPT influence the training data for different classifiers addressing the three mentioned sentiment analysis tasks? To address this, we will empirically compare the annotation generated by ChatGPT for training data for different classifiers addressing the three mentioned sentiment analysis tasks.

To the best of our knowledge, this is the first work that investigates the problem of using an LLM to address relevant sentiment analysis tasks in Portuguese. Our main contributions can be summarized as follows:

- We conducted experiments to evaluate the impact of the temperature hyperparameter on the performance of ChatGPT in NLP tasks.
- In our experiments, we identify that ChatGPT exhibit exceptional performance in sentiment analysis tasks, specifically in the identification of subjectivity and polarity in sentences. In terms of comparative sentences identification, ChatGPT demonstrate a lower performance compared with baselines.
- We conduct comprehensive analysis of the feasibility of leveraging ChatGPT for data annotation for complex NLP task.

The remainder of the paper is organized as follows. Section 2 provides a review of the related work on Large language models (LLMs) and Section 3 presents an overview of the methodology applied in our study. Section 4 includes experimental evaluation of the proposed approach. Finally, Section 5 discusses our main conclusions, limitations, and future research directions.

2 Related Work

The main goal of this study is to investigate the ability of ChatGPT for dealing with classic sentiment analysis tasks across a wide range of datasets in Brazilian Portuguese.

2.1 ChatGPT

ChatGPT[1] is a language model developed by OpenAI, based on the GPT-3.5 architecture, that can generate coherent and contextually relevant text given a prompt. It has 175 billion parameters, making it one of the largest language models today (Brown et al., 2020). According to OpenAI, ChatGPT can perform various tasks such as question answering, summarization, and translation without any additional training. The model was trained on a large corpus of text from various sources, including books, articles, and websites.

With the launch of the GPT-4 engine, the translation performance of ChatGPT is significantly boosted, becoming comparable to commercial translation products, even for distant languages(Jiao et al., 2023).

Several applications of intelligent chatbots has emerged in different areas showing, with some care, powerful results anda advantages (Bahrini et al., 2023). For instance, (Sallam et al., 2023) lists the following pros of chatGPT integration in the medical educational process: Improved personalized learning, improved clinical reasoning, and assistance to understand complex medical concepts.

2.2 Sentiment Analysis Tasks

Sentiment analysis is such a research area which identifies and extracts information about the opinions, attitudes, emotions, and sentiments expressed in text. A lot of research has been developed addressing opinions expressed in the English language. However, studies involving the Portuguese language still need to be advanced to make better use of the specificities of the language (Pereira, 2021). Our study aims to cover the state of the art research related some of the main tasks regarded to sentiment analysis in Portuguese: a) identifying opinionated from factual sentences (de Oliveira and de Melo, 2021); b) identifying the polarity of opinion sentences as positive or negative (Oliveira and de Melo, 2020); c) identifying comparative from regular sentences (Kansaon et al., 2020).

2.3 Annotators

In NLP applications, the utilization of labeled data is often necessary, which involves the manual process of data annotation. Traditionally, there have been two primary strategies employed for this purpose. Firstly, researchers can recruit and train coders, such as research assistants, to perform the annotation task. Secondly, they can rely on crowdworkers available on platforms like Amazon Mechanical Turk (MTurk) to annotate the data (Gilardi et al., 2023).

In a recent analysis conducted by Gilardi et al. (Gilardi et al., 2023), it was demonstrated that

[1]https://openai.com/blog/chatgpt

ChatGPT outperformed human workers for text-annotation in several tasks. Furthermore, other studies by Ding et al. (Ding et al., 2022) have shown that the performance of ChatGPT models is slightly lower when compared to human-labeled data. However, the utilization of ChatGPT models significantly reduces the cost and time required for the annotation process when compared to relying solely on human annotators.

Particularly, works such as those presented by Qin et al. (Qin et al., 2023) share similar objectives with our research; however, they are primarily focused on the English language. In contrast, our work provides an additional contribution by evaluating the performance of ChatGPT models on Portuguese texts.

These findings indicate that ChatGPT presents promising capabilities in accurately performing text data annotation task with many benefits, such as performance or costs, when compared to relying solely on human annotators. For these reasons, we have decided to investigate the use of ChatGPT in automatic training data generation (RQ2).

3 Methodology

The main goal of this study is to investigate the potential of ChatGPT's generalization across several sentiment analysis tasks, specifically in the context of Brazilian Portuguese. This research is centered around two principal research questions.

The research question (RQ1) seeks to empirically validate the performance of ChatGPT as a competent resolver for relevant sentiment analysis tasks. To validate this research question, we conducted evaluations on three crucial sentiment analysis tasks described as follows, where the Figure 1 shows the summary of our zero-shot prompt designs.

The first task (Task 1) is a sentence classification as either factual or opinionated, where the prompt design is showed in Figure 1 (a). For instance, the sentence "*o restaurante tem um ambiente agradável*" ("the restaurant has a pleasant atmosphere") would be classified as opinionated, whereas "*o restaurante abre às 14 horas*" ("the restaurant opens at 2 p.m.") would be classified as a factual sentence. This study adopted the methodology outlined in (de Oliveira and de Melo, 2021) as the baseline, and also utilized the datasets made available by the authors of this paper.

The main goal of the second task (Task 2) is to classify each sentence as either positive or negative sentiment, where the prompt desing is showed in Figure 1 (b). The sentence "*a comida estava deliciosa*" ("the food was delicious") exhibits a positive sentiment, while "*o preço era muito salgado*" ("the price was very steep") conveys a negative sentiment about the restaurant's pricing. The methodologies elaborated in (Oliveira and de Melo, 2020) were employed as the baseline for this task, and the datasets published by the respective authors were also used.

The third task (Task 3) consists of classifying sentences as either comparative or direct, where the prompt design is showed in Figure 1 (c). For instance, the sentence "*o restaurante tem um ambiente agradável*" ("the restaurant has a pleasant atmosphere") is a direct sentence, while the sentence "*o sorvete da McDonald's é melhor*" ("McDonald's ice cream is better") is comparative. The methods outlined in (Kansaon et al., 2020) served as the baseline for this task, and the datasets published by the authors were also employed.

The second research question (RQ2) aims to validate the feasibility of using ChatGPT models for automating dataset labeling. To address RQ2, firstly, we utilized ChatGPT to label our data, as obtained from RQ1. We employed the labeled data from ChatGPT to train models using AutoGluon~(Erickson et al., 2020). Finally, we compared the results obtained from these models with baselines and with ChatGPT itself to assess their performance and effectiveness.

3.1 Exploration of ChatGPT Models

OpenAI offers a diverse range of models via their API, each tailored for distinct purposes and performance benchmarks. For our study, we focused on GPT 3.5-Turbo, the Large Language Model (LLM) encompassing 175B parameters, which also powers the online ChatGPT — hereafter referred to as ChatGPT. This model, within the GPT-3.5 series, stands out for its robustness and is optimized for chat functionalities, rendering it ideal for tasks centered around dialogue interaction. Moreover, ChatGPT delivers performance on par with other models from OpenAI but at roughly one-tenth of the computational expense, making it a cost-effective alternative for researchers and developers[2]. Our experiments were consistently conducted using OpenAI's official API, with the same parameters and

[2]https://platform.openai.com/docs/models/gpt-3-5

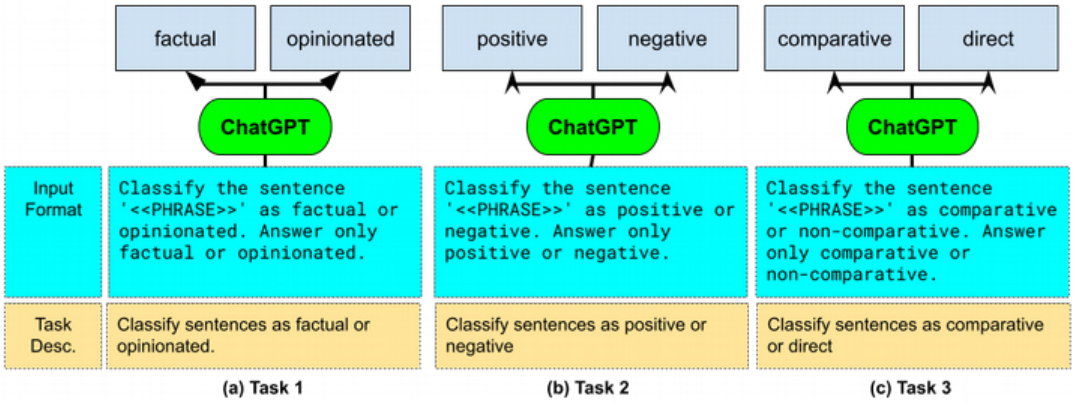

Figure 1: Zero-shot prompt designs.

model version, unless otherwise specified.

In order to evaluate the impact of ChatGPT's temperature parameter, which controls the degree of randomness of the model's output, we performed the tasks with the value of 0, which implies more deterministic, as well as with a value of 1.0, which implies higher randomness. As noted by Gilardi (Gilardi et al., 2023), employing lower temperatures values yields superior outcomes in sentiment analysis task when leveraging ChatGPT.

3.2 Prompts

According to Liu et al. (Liu et al., 2023), a prompt serves as a set of instructions given to an LLM, effectively programming the LLM by customizing, enhancing, or refining its capabilities. Selecting an appropriate prompt is essential for ChatGPT to provide the desired answer accurately. Initially, we made some attempts with prompts that had more detailed instructions, but we observed that prompts with direct instructions yield better results. Below, the selected prompt for each task is presented.

For Task 1, we choose the following prompt: *Classifique a sentença* "FRASE" *em factual ou opinativa. Responda somente factual ou opinativa* (Classify the sentence "SENTENCE" as factual or opinionated. Respond only with factual or opinionated), where the sentence that we want to evaluate is positioned between apostrophes. For this prompt, it is expected that ChatGPT responds only with *"factual"* (factual) or *"opinativa"* (opinionated).

For Task 2, we choose the following prompt: *Classifique a sentença* "FRASE" *em positiva ou negativa. Responda somente positiva ou negativa* (Classify the sentence "SENTENCE" as positive or negative. Respond only with positive or negative), where the sentence that we want to evaluate is positioned between apostrophes. For this prompt, it is expected that ChatGPT responds only with *"positiva"* (positive) or *"negativa"* (negative).

Finally, for Task 3, we choose the following prompt: *Classifique a sentença* "FRASE" *em comparativa ou não comparativa. Responda somente comparativa ou não comparativa* (Classify the sentence "SENTENCE" as comparative or direct. Respond only with comparative or direct), where the sentence that we want to evaluate is positioned between apostrophes. For this prompt, it is expected that ChatGPT responds only with *"comparativa"* (comparative) or *"não comparativa"* (direct).

4 Experiments

In this section, we detail the experimental setup, encompassing the description of the datasets used and the evaluation metrics adopted. Subsequently, we present and discuss the experimental results.

4.1 Datasets

For the Task 1, we utilized three distinct datasets comprising both factual and subjective sentences. Different datasets were employed to test ChatGPT's robustness across diverse linguistic and contextual challenges inherent in Brazilian Portuguese, ensuring comprehensive validation for varied sentiment analysis tasks and alignment with standard benchmarks.

The details of each dataset are presented in the Table 1. ReLi consists of a collection of book reviews in Portuguese, retrieved from the internet and manually annotated (Freitas et al., 2012). TA-Restaurants contains sentences in Portuguese related to restaurant reviews collected from TripAdvisor[3] (Oliveira and de Melo, 2020). Computer-BR

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

is a set of tweets in Portuguese and covers a wide range of topics related to computers (Moraes et al., 2016).

	Factual	Subjective	Total
ReLi	175	175	350
TA-Restaurants	591	458	1,049
Computer-BR	604	1,677	2,281

Table 1: Dataset for Task 1.

For Task 2, we used the same datasets as in Task 1, but with added annotations for sentiment polarity (either positive or negative). Furthermore, we incorporated the Google Play corpus annotated by Junior and Merschmann (Stiilpen Junior and Merschmann, 2016). This corpus consists of 1,630 sentences, randomly selected from an original set of 10,000 mobile application reviews on the Google Play Store. The sentences in the Google Play corpus are evenly split between positive and negative sentiments.

	Positive	Negative	Total
ReLi	85	85	170
TA-Restaurants	505	56	561
Computer-BR	198	400	598
Google Play	815	815	1,630

Table 2: Dataset for Task 2.

Lastly, the Table 3 presents two additional datasets for the Task 3. Twitter is a corpus of comparative sentences mined from related to electronic products (Kansaon et al., 2020) and Buscapé consists of product evaluations collected from the Buscapé[4] website (Kansaon et al., 2020). The datasets are annotated as comparative or direct sentences.

	Direct	Comparative	Total
Buscapé	1,282	1,472	2,754
Twitter	918	1,135	2,053

Table 3: Dataset for Task 3.

4.2 Evaluation Metrics

We use the metrics of precision (P), recall (R) and F-measure (F_1) to evaluate the models in the tasks investigated in this paper (Baeza-Yates et al., 1999). Let A be the set of correct answers, according to a reference set, and let B be the set of responses produced by the method that is being evaluated. We define precision (P), recall (R) and F-score (F_1) as:

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

4.3 Results

In this section, we show the results of both stated research questions for the different datasets and models of Tasks 1 to 3.

4.3.1 Research Question 1

Initially, we assessed the influence of the temperature hyperparameter on ChatGPT's performance across all the tasks. We considered a temperature of 0, where the model is entirely deterministic, and a temperature of 1, where the model generates more creative responses. Figure 2 displays the F1 score values for the different tasks (in different colors), and for each dataset of a given task. It is noteworthy that the model with a temperature of 0 produced results that were better or, at the very least, equal to the model with a temperature of 1. The rationale behind this is that the objective of text classification is to produce a singular output for a given input. Therefore, the freedom to choose more varied and creative answers tends to yield poorer results in text classification tasks.

The results for all the tasks are better described as follows by considering ChatGPT with temperature of 0, and comparing it with with the respective state-of-the-art methods for each task.

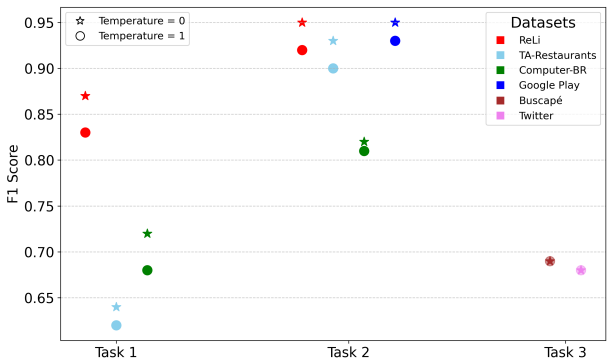

Figure 2: ChatGPT performance for different temperatures.

The results for the Task 1 (subjectivity identification) are presented in Table 4. The analysis shows that the ChatGPT model achieved results very close to GBT on the ReLi dataset, which is the

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

state of the art for this task. While ChatGPT underperformed on the TA-Restaurants dataset, it surpassed performance on the Computer-BR dataset. It is noteworthy to record that ChatGPT is doing the classification of the datasets without any training, in a zero-shot manner.

Both the ReLi and TA-Restaurants are datasets with more descriptive and formal texts when compared to Computer-BR, which is composed of tweets. These tweets are often written in abbreviated forms, using jargon or colloquial language. Thus, we can see that the dataset-specific trained model from literature performed much better on the first two cases, but ChatGPT showed to be more resilient to the noisy data from the last dataset. Based on our observations, it appears that ChatGPT may not understand well the subjectivity of a sentence in most cases, but it is much more capable of dealing with different types of texts due to the huge and diverse data used during its training.

The results for Task 2 (polarity identification) are presented in Table 5. It shows that the ChatGPT model achieved much more superior results on the ReLi, Computer-BR, and Google Play datasets than GBT, while it presented similar F1-score on TA-Restaurants.

The results suggest that ChatGPT is very capable of determining the polarity of sentences. Despite not being fine-tuned on those specific datasets, it is plausible that sentiment and polarity analysis are common in the diverse texts used for ChatGPT's training. For instance, it is expected that texts from conversations and literature talk about the positivity or not of ideas much more than subjectivity. Furthermore, ChatGPT's training incorporated user reviews related to products and services from various platforms. Such feedback typically includes a star rating system: comments with 1 or 2 stars are interpreted as negative, while those with 4 or 5 stars are positive. This allows ChatGPT to effectively discern the polarity of terms and phrases within these reviews. These observations might shed light on ChatGPT's comparatively lower performance on Task 1. Lastly, prompts seeking text sentiment tend to be more straightforward compared to those probing subjectivity (factual or opinionated). This intrinsic clarity in sentiment prompts may reduce the chances of misinterpretation.

The results for Task 3 (identification of comparative sentences) are presented in Table 6. The ChatGPT model exhibits inferior performance compared to the state-of-the-art method NB. Such as in Task 1, ChatGPT notably struggles in recognizing comparative sentences. This limitation is potentially attributed to the fact that ChatGPT was not trained on these specific datasets. Furthermore, common texts used during its training might not frequently feature explicit comparative judgments, a point previously discussed in the context of Task 1 and contrasting the expectations for Task 2. For instance, sentences such as *"acho um ótimo smartphone em relação aos eu preço com muitas funções"* (I think it's a great smartphone for its price with many features) and *"preço poderia ser mais acessível já a Caloi é no brasil"* (the price could be more affordable since Caloi is in Brazil) are identified as comparative sentences by ChatGPT, despite there is no explicit comparison between two products.

ChatGPT demonstrates exceptional performance in sentiment analysis, particularly in identifying both subjectivity and polarity within sentences. In the task of polarity identification, ChatGPT's performance stands out as the best overall, suggesting it can reliably handle such tasks with minimal issues. For the identification of comparative sentences, although ChatGPT did not achieve the best results, the selection of a more appropriate prompt might improve outcomes. Adding more tokens could further refine the responses, but this might also increase the cost per request. The experimental results indicate that ChatGPT could be used as a suitable method to address the tasks analyzed.

4.3.2 Research Question 2

The goal of RQ2 is to experimentally verify if the classification of sentences by ChatGPT in zero-shot could be used to train an AutoML model. The results present the comparison of the state of the art models, ChatGPT and AutoGluon, in which only the last was trained with datasets automatically annotated by ChatGPT for all the three considered tasks evaluated before.

Table 7 shows the comparative results of identification of subjectivity (Task 1). We can observe that the performance of AutoGluon on the ReLi and Computer-BR datasets surpassed the state-of-the-art GBT, and in the first case, it also was superior to ChatGPT. However, in the other two datasets, AutoGluon's results underperformed compared to ChatGPT. This results indicate that ChatGPT annotations can be used to train other models to achieve a close performance than itself. And note that in the case of Computer-BR dataset, the trained model

	ReLi			TA-Restaurants			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	0.58	0.68	0.68	0.63	0.63	0.63	0.54	0.54	0.54

Table 4: Task 1 - Identification of subjectivity.

	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	0.96	0.96	0.96	0.93	0.93	0.93	0.82	0.82	0.82	0.95	0.95	0.95

Table 5: Task 2 - Identification of polarity.

	Buscape			Twitter		
	P	R	F_1	P	R	F_1
NB	0.87	0.88	0.87	0.86	0.86	0.86
ChatGPT	0.67	0.67	0.67	0.61	0.61	0.61

Table 6: Task 3 - Identification of comparative sentences.

also surpassed GBT by far.

In Table 8, we present the comparative results for the task of polarity identification (Task 3). It is evident that AutoGluon's performance is inferior than ChatGPT across all examined datasets. While ChatGPT's performance significantly exceeded the benchmarks set by the state-of-the-art model, the approach of utilizing ChatGPT as an automated annotator for training AutoGluon did not perform so well. Results indicate that either ChatGPT training present some loss when labeling data for training, or the AutoGluon trained model is not so good than ChatGPT to generalize data. It is important to mention that ChatGPT is based on a very large and powerful model trained over extensive textual data. Nevertheless, results from AutoGluon are better than GBT in all cases but TA-Restaurants. Thus, we can conclude that ChatGPT may be a useful annotation tool in tasks that it already presents a good performance.

In Table 9, we present the comparative results for Task 3 (identification of comparative sentences). AutoGluon, which was trained using ChatGPT annotations, showed a very close performance to ChatGPT. This result suggests that AutoGluon managed to learn effectively from the annotations provided by ChatGPT. However, its slight lower performance for the Twitter dataset, particularly in the F1-score, might indicate that the model had challenges generalizing across diverse data sources when relying on ChatGPT's annotations. One potential explanation for AutoGluon's inferior performance relative to ChatGPT could be caused by the inherent complexities of model architectures. While ChatGPT has been extensively trained on diverse linguistic patterns and can adapt to various data nuances, AutoGluon may not extrapolate as effectively from the annotated data alone. Furthermore, Twitter data, being more informal and diverse, might introduce additional challenges that could influence the model's ability to generalize.

From the presented results, we can deduce that, even with a slight decrease in performance, utilizing labeled data from ChatGPT to train other machine learning models remains a viable option. This advantage becomes particularly evident when ChatGPT demonstrates strong performance, as showed in the sentence sentiment analysis (Task 2). Given the sheer size of ChatGPT, boasting 175 billion parameters, leveraging its capabilities to train more compact models, such as AutoGluon, could provide a significant edge in deploying efficient deep learning solutions.

5 Conclusions

This paper presented a comprehensive study investigating the effectiveness of ChatGPT model in addressing three relevant sentiment analysis tasks in Portuguese using various datasets. Our findings demonstrate that ChatGPT models, particularly GPT 3.5-Turbo, can be successfully utilized as sentiment analysis solvers. Furthermore, we found out that the dataset annotated by ChatGPT can be used to train alternative models with minimal impact on performance, while still producing comparable results to those achieved by ChatGPT. Thus, it can be an useful tool when time and cost are important aspects on building machine learning

	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	0.68	0.68	0.68	0.64	0.64	0.64	0.72	0.72	0.72
AutoGluon	0.80	0.79	0.79	0.64	0.54	0.48	0.67	0.72	0.68

Table 7: Identification os subjectivity (Task 1) - using ChatGPT as annotator.

	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	0.96	0.96	0.96	0.93	0.93	0.93	0.82	0,82	0.82	0.95	0.95	0.95
AutoGluon	0.71	0.78	0.70	0.71	0.60	0.63	0.77	0.79	0.77	0.94	0.94	0.94

Table 8: Identification of polarity (Task 2) - using ChatGPT as annotator.

	Buscape			Twitter		
	P	R	F_1	P	R	F_1
NB	0.87	0.87	0.87	0.86	0.86	0.86
ChatGPT	0.67	0.67	0.67	0.61	0.61	0.61
AutoGluon	0.66	0.66	0.66	0.66	0.60	0.58

Table 9: Identification of comparative sentences (Task 3) - using ChatGPT as annotator.

models.

However, for some other tasks, as subjectivity and comparative identification of sentences, ChatGPT did not performed well in a zero-shot solution. We suggest that this occurs due to both the facility to build direct prompts and to natural occurrence of the subject in ChatGPT training data. For instance, sentiment identification of sentences has a more precise prompt and is a language structure very common to occur in any textual subject, which may explain the superior performance of ChatGPT.

In future research, there are several avenues to explore for further improvement. One area of focus will be enhancing prompt engineering techniques to extract even better results from the GPT 3.5-Turbo model. Additionally, we plan to investigate the performance of other LLM models available in the Open Source community, expanding our evaluation to encompass a wider range of models and comparing their effectiveness in sentiment analysis tasks.

Acknowledgements

This work was supported by Samsung Ocean Center, a research program in the State University of Amazonas. The authors also would like to acknowledge the financial support provided by Fundação de Amparo à Pesquisa do Estado do Amazonas - FAPEAM (FAPEAM UNIVERSAL N. 001/2023, Protocolo N. 66074.UNI961.4630.16032023).

References

Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. 1999. *Modern information retrieval*, volume 463. ACM press New York.

Aram Bahrini, Mohammadsadra Khamoshifar, Hossein Abbasimehr, Robert J. Riggs, Maryam Esmaeili, Rastin Mastali Majdabadkohne, and Morteza Pasehvar. 2023. Chatgpt: Applications, opportunities, and threats. In *2023 Systems and Information Engineering Design Symposium (SIEDS)*, pages 274–279.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences.

Miguel de Oliveira and Tiago de Melo. 2021. An empirical study of text features for identifying subjective sentences in portuguese. In *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II 10*, pages 374–388. Springer.

Bosheng Ding, Chengwei Qin, Linlin Liu, Lidong Bing, Shafiq Joty, and Boyang Li. 2022. Is gpt-3 a good data annotator? *arXiv preprint arXiv:2212.10450*.

Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. 2020. Autogluon-tabular: Robust and accurate automl for structured data.

Cláudia Freitas, Eduardo Motta, R Milidiú, and 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.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine.

Daniel Kansaon, Michele A Brandão, Julio CS Reis, Matheus Barbosa, Breno Matos, and Fabrício Benevenuto. 2020. Mining portuguese comparative sentences in online reviews. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*, pages 333–340.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and 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.

Silvia MW Moraes, André LL Santos, Matheus Redecker, Rackel M Machado, and Felipe R Meneguzzi. 2016. Comparing approaches to subjectivity classification: A study on portuguese tweets. In *International Conference on Computational Processing of the Portuguese Language*, pages 86–94. Springer.

Miguel V Oliveira and Tiago de Melo. 2020. Investigating sets of linguistic features for two sentiment analysis tasks in brazilian portuguese web reviews. In *Anais Estendidos do XXVI Simpósio Brasileiro de Sistemas Multimídia e Web*, pages 45–48. SBC.

Denilson Alves Pereira. 2021. A survey of sentiment analysis in the portuguese language. *Artificial Intelligence Review*, 54(2):1087–1115.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Malik Sallam, Nesreen Salim, Muna Barakat, and 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.

Milton Stiilpen Junior and Luiz Henrique C Merschmann. 2016. A methodology to handle social media posts in brazilian portuguese for text mining applications. In *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web*, pages 239–246.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.