TeamUFPR at ABSAPT 2024: Improving Sentiment Polarity Classification with Data Augmentation using ChatGPT

Bruno Dal Pontte[1], Anderson Aparecido do Carmo Frasão[1], Marcus Vinícius Reisdoefer Pereira[1], Tiago Heinrich[1] and Vinicius Fulber-Garcia[1]

[1]*Department of Informatics - Federal University of Paraná – Curitiba, Parana, Brazil, 81531-990*

Abstract

This paper describes the participation of TeamUFPR in the *Aspect-Based Sentiment Analysis in Portuguese* (ABSAPT 2024). The challenge proposed two tasks aimed at evaluating aspects and sentiments in Portuguese. Based on solutions from previous years, we explored the BERT model for sentiment classification of texts. However, we also focused on developing a solution that would be helpful in data generation. We chose this strategy because we were limited to 4.8k training samples, making it important to explore new ways to generate high-quality data. Thus, we explored ChatGPT to better understand how more samples could be generated and the potential behind using LLMs in the creation of training data for the proposed tasks.

Keywords

Sentimental Analysis, Natural Language Processing, Aspect Extraction

1. Introduction

Over the past decade, Natural Language Processing (NLP) techniques have gained significant prominence. These techniques aim to represent and extract characteristics from texts and sentences, mirroring the way humans understand text. NLP offers significant benefits, particularly in processing the vast amount of information available on the Internet, which would be impractical for humans to analyze quickly [1].

The study of Aspect-Based Sentiment Analysis (ABSA) involves various methods aimed at identifying and extracting specific aspects of sentences. These methods allow for detailed information retrieval, providing a deeper understanding of the sentiments expressed within the text. ABSA is particularly valuable in categorizing and characterizing content during analysis, emphasizing its pivotal role in data analysis across sectors such as market research and Data Science.

ABSAPT 2024 is the second edition of the IberLEF task dedicated to aspect-based sentiment analysis in the Portuguese language [2, 3]. The competition uses a dataset obtained from

bdp20@inf.ufpr.br (B. D. Pontte); aacf20@inf.ufpr.br (A. A. d. C. Frasão); mvrp21@inf.ufpr.br (M. V. R. Pereira); theinrich@inf.ufpr.br (T. Heinrich); vinicius@inf.ufpr.br (V. Fulber-Garcia)
0000-0002-8017-1293 (T. Heinrich); 0000-0003-1544-6315 (V. Fulber-Garcia)

TripAdvisor, consisting of user comments. ABSAPT is one of the tasks featured at IberLEF 2024, within the sentiment, posture, and opinions section.

Thus, in the context of ABSAPT 2024, our proposal primarily consists of implementing a strategy focused on classification and sentiment extraction. This strategy takes into account the results obtained in previous work [4].

Our team has experience addressing machine learning challenges focused on detecting and mitigating computer security breaches, as well as applying artificial intelligence in gaming environments. While we are not directly involved in Natural Language Processing (NLP) tasks, we are interested in exploring learning strategies tailored to this context and investigating the feasibility of incorporating prior knowledge into our solutions.

The rest of this article is structured as follows: Section 2 describes the tasks of ABSAPT 2024. Section 3 presents related work. Section 4 details the proposed solution and discusses our findings. Finally, Section 5 concludes the article.

2. Task Description

In 2024, two tasks were proposed for competitors to explore. Both tasks focus on the use of a new dataset, which aims to address the challenge of developing specific methods for Aspect-Based Sentiment Analysis in Portuguese [2].

The data used by the competitors were sourced from TripAdvisor reviews written in Portuguese. Table 1 shows the distribution of the data used to train the models, along with the two test sets employed to evaluate the effectiveness of the models proposed by each team (one test set for each task).

Data	Number of Samples
Train	4828
Test task 1	284
Test task 2	1176

Table 1
ABSAPT 2024 Dataset.

The objective of *Task 1* is to extract aspect terms. Each team must propose and test a solution for identifying one or more aspects within a sentence using the TripAdvisor dataset. Each evaluation may yield a list of aspects that will be considered during the model evaluation.

Task 2 focuses on extracting sentiment orientation. Each team is required to develop a solution for determining the polarity of TripAdvisor reviews, which can be categorized as positive, negative, or neutral. Teams must identify the polarity of each review using predefined templates.

3. Related Works

Previous experiences of our group in the competition explored different strategies [5, 6]. In 2021, we focused on exploring machine learning classification strategies, testing nine different types of models with various feature selections, achieving an F1-score of 78% for one of the datasets. In 2022, our group pursued two different strategies: one based on Conditional Random Fields (CRF) for aspect extraction, and another utilizing Bidirectional Encoder Representations from Transformers (BERT) for sentiment extraction.

The Aspect-Based Sentiment Analysis in Portuguese (ABSAPT 2022) task summarizes the overall results of the 2022 competition [4]. Competitors mainly focused on exploring BERT or CRF, with the top results achieved using BERT. The highest F1-score achieved was 81%.

4. Proposed Strategies

In this section, we describe the generation of new data and the strategies employed for each task. Given the nature of the proposed tasks, we considered and applied two strategies. These strategies used the entire dataset presented in Section 2, along with additional data generated using ChatGPT, to train the algorithms.

4.1. Generating Extra Data With ChatGPT

In an attempt to increase the size of the training dataset, and considering the pivotal role of data abundance in refining a model's accuracy, our team opted to use ChatGPT to generate synthetic reviews, thereby obtaining the necessary data to expand the dataset.

However, generating the data posed several challenges. While creating the reviews themselves was straightforward, extracting their (sometimes multiple) aspects and polarities proved difficult for the chatbot. Initially, our aim was to avoid any manual intervention and generate approximately five thousand additional entries for the training dataset, effectively doubling its size. However, compromises had to be considered: either (i) introduce more manual intervention than anticipated, or (ii) accept that some of the generated data might need to be discarded because it wouldn't meet the challenge's objectives.

Each batch of data generated exhibited some variation, aimed at introducing diversity in writing styles to make the new dataset resemble the original one. The general guideline was the following: instructing the bot to create synthetic reviews from TripAdvisor, then prompting it to extract the aspects and polarity and finally formatting everything into a CSV file, then continue the generation of this file.

generate a couple of synthetic reviews from tripadvisor in brazilian portuguese

Table 2
Input prompt 1: generating reviews

now extract the aspects reviewed from each text, make sure they're words that are in the review body

Table 3
Input prompt 2: extracting aspects

make sure they’re single words

Table 4
Input prompt 3: correcting the output of the previous instruction

alongside the aspects put a number to represent the polarity of the aspect: 1 for positive, 0 for neutral and -1 for negative

Table 5
Input prompt 4: extracting the polarities

The first four prompts (Tables 2, 3, 4, and 5) consistently focused on generating the data: the reviews, aspects, and polarities. Occasionally, it was necessary to reiterate a rule for aspects, as they were required to be single words present within the review body itself. The last two prompts (Tables 6 and 7) aimed to format the data correctly and proceed with the generation in its final format.

now put it all together in a csv with the following columns: text, aspect, polarity. repeat the text if there’s more than one aspect in the same review

Table 6
Input prompt 5: formatting the data

continue generating reviews, aspects and polarities like this in the CSV format

Table 7
Final input prompt: continuing the generation

The process turned out to be much more complex than initially anticipated due to ChatGPT’s inability to perform all these steps in a single prompt. Even when attempting to complete all five steps in one prompt, subsequent adjustments were consistently necessary to tailor the generation to our requirements. Continuous monitoring during the initial stages of generation was crucial to ensure the synthetic dataset’s usefulness.

Furthermore, following the execution of the general workflow, it became necessary to normalize the generated data. In several instances, the aspects generated for the reviews were not actually present in the text. This highlighted a significant drawback either in the generation method or the overall strategy: too many entries were eliminated during the ’cleaning’ process. As a result, achieving the initial goal of generating 5,000 lines promptly became unfeasible. Consequently, our team settled for 1,281 new entries, thereby increasing the size of the original training dataset from 4,828 to 6,109 entries.

After analyzing the generated data, our team concluded that creating approximately 400 lines in a single chat was the limit of the chatbot’s ’usefulness’ (with about 30% to 40% of the lines being removed later). Beyond this threshold, further generation became increasingly unreliable, with a significant portion of lines being discarded during the data cleaning process. Having these empirical estimates beforehand would have made the generation process more efficient and potentially enabled our team to reach the goal of generating five thousand additional lines.

The limitations of this method include the need to verify the quality and relevance of the generated data, as well as to check for biases in the synthetic texts. These constraints contradict the initial goal of automation expected from this approach, as ChatGPT's responses are shaped by the data it was trained on. It's possible that more domain-specific large language models might be better suited for this particular task.

Overall, unless significantly redesigned, the proposed strategy was not practical for meaningfully increasing the size of a dataset, especially for very large datasets. This is because the primary objective of easily automating the process of obtaining more relevant data was not achieved.

4.2. Task 1

Task 1 was not our team's focus.

4.3. Task 2

Our team analyzed the strategies outlined in [4] for extracting sentiment orientation. Consequently, we focused on utilizing BERT for sentiment extraction.

Table 1 presents the parameters used for the three evaluation tests conducted. The test scenarios consist of: (i) 'Based', where the model was trained and validated using only the dataset for ABSAPT 2024; (ii) 'ChatGPT', where we use only the LLM-generated dataset; and (iii) 'Extended', where both datasets are merged.

Starting from the aforementioned strategies, we focused on improving the accuracy and F1-score of sentiment orientation prediction. To that end, we trained the BERT model on an expanded version of the given training dataset, which included LLM-generated entries of example sentences fitting the dataset format. The larger amount of training data enabled us to enhance the model's accuracy.

Figure 1: Parameters used in the algorithm.

BERT	batch size: 32 epoch: 10 learning rate: 5e-5 Train/test 50/50

Figure 2: Evaluation with different datasets.

		Loss	Accuracy	F1-score
Base	Train	0.172	93.45	93.28
	Val.	0.689	77.80	76.97
ChatGPT	Train	0.126	94.82	94.83
	Val.	0.676	79.42	79.40
Extended	Train	0.155	92.32	92.27
	Val.	0.253	89.81	89.72

Table 2 presents the training and validation results for each of the respective scenarios. The training phase shows a difference of less than 2% between the three scenarios. However, the validation results indicate a reduction in loss, as well as an increase in accuracy and F1-score, when we consider the extended dataset that includes the ChatGPT-generated data.

5. Conclusion

In this paper, we describe TeamUFPR's participation in the ABSAPT 2024 (Aspect-Based Sentiment Analysis in Portuguese) competition. The first task involved developing a model to extract aspects from TripAdvisor reviews, while the second task focused on determining the sentiment expressed toward each aspect.

Our implementation focused mainly on the second task: classifying the extracted aspects from the text into one of three polarities—positive, negative, or neutral. To achieve this, our team proposed using the Hugging Face Transformers library, building upon the BERT-based approach from a previous ABSAPT entry [6].

We also experimented with using ChatGPT to significantly increase the training dataset size with the goal of improving the model's accuracy. However, our observations revealed the limitations of this approach. The synthetic data generated was not sufficient for our purposes, and most of it ended up being removed from the final dataset. Overall, while it is possible to achieve better results with this technique, the method for requesting data needs to be reviewed to better suit massive data generation.

Finally, we observed that our approach improved upon the previous results for the second task in the ABSAPT-2024 competition compared to the ABSAPT-2022 entry [6]. Specifically, the F1-score of the training results improved from 92.21% in the previous entry to 94.83% when using the ChatGPT-generated entries. However, the F1-score of the Extended model, which included both the base (given dataset) and ChatGPT-generated entries, fell to 92.27%, while its accuracy increased compared to the base model. Additionally, our focus was limited for the aspect extraction task, and we could have achieved better results had we explored more varied models.

For future work, it might be valuable to explore data generation solutions using other large language models to obtain larger datasets from smaller samples. Special attention should be paid to biases and ways to improve the quality of the data generated by these models. Furthermore, a comparison could be made between the accuracy of the Portuguese-language model and that of a similar English-language model.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES). The authors also thank the UFPR Informatics Department.

References

[1] A. D. Francisco, Aspect Term Extraction in Aspect-Based Sentiment Analysis, B.s. thesis, Federal Rural University of Pernambuco, 2019.

[2] L. Chiruzzo, S. M. Jiménez-Zafra, F. Rangel, Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org, 2024.

[3] G. A. Gomes, A. T. Bender, E. P. Lopes, L. A. de Freitas, U. B. Corrêa, ABSAPT 2024 at IberLEF: Overview of the Task on Aspect-Based Sentiment Analysis in Portuguese, Procesamiento del Lenguaje Natural 73 (2024).

[4] F. Leonel Vasconcelos da Silva, G. da Silva Xavier, H. Mota Mensenburg, L. Pereira dos Santos, R. Ferreira Rodrigues, R. Matsumura Araújo, U. Brisolara Corrêa, L. Astrogildo de Freitas, ABSAPT2022 at IberLEF: Overview of the Task on Aspect-Based Sentiment Analysis in Portuguese, Procesamiento del Lenguaje Natural 69 (2022).

[5] T. Heinrich, F. Ceschin, F. Marchi, TeamUFPR at IDPT 2021: Equalizing a Strategy Using Machine Learning for Two Types of Data in Detecting Irony, in: Iberian Languages Evaluation Forum, co-located with the Conference of the Spanish Society for Natural Language Processing, CEUR-WS.org, volume 2943, 2021, pp. 925–932.

[6] T. Heinrich, F. Marchi, TeamUFPR at ABSAPT 2022: Aspect Extraction with CRF and BERT, 2022.