Enhancing Aspect-Based Sentiment Analysis for Portuguese Using Instruction Tuning

Gabriel Pereira[1], **Luciano Barbosa**[1], **Johny Moreira**[2], **Tiago Melo**[3], **Altigran Silva**[2]

[1]Centro de Informática – Universidade Federal de Pernambuco (UFPE)

[2]Instituto de Computação – Universidade Federal do Amazonas (UFAM)

[3]Escola Superior de Tecnologia – Universidade do Estado do Amazonas (UEA)

{mgpp,luciano}@cin.ufpe.br

tmelo@uea.edu.br,{johny.moreira,alti}@icomp.ufam.edu.br

Abstract. *This study explores the application of instruction tuning in open-source small language models for Portuguese End-to-End Aspect-Based Sentiment Analysis (E2E-ABSA), focusing on restaurant reviews. Utilizing a diverse dataset from sources such as Google Reviews, TripAdvisor, Instagram, and iFood, the research evaluates the performance of PTT5 Base, a T5 model pretrained on Portuguese data, in comparison to multilingual models, namely FLAN-T5 Base and mT0 Small. The results show that the PTT5 Base has superior capabilities in E2E-ABSA, achieving an F1 Score of 0.60, Precision of 0.61, and Recall of 0.59. These findings emphasize the significance of language-specific pretraining in analyzing customer opinions for the ABSA task.*

1. Introduction

Traditional sentiment analysis focuses on determining whether a text expresses a positive, negative, or neutral sentiment. For example, a comment like "I loved the decoration" would typically be classified as positive. However, this approach falls short when users express opinions about multiple aspects within the same text. For instance, in the statement "I loved the decoration, but the pizza was awful," traditional sentiment analysis would struggle to capture the sentiment towards each individual aspect, leading to a less nuanced understanding of user feedback [Zhang et al. 2022].

To address this issue, Aspect-Based Sentiment Analysis (ABSA) has been developed. ABSA aims to identify aspects (e.g., "pizza"), aspect categories (e.g., "food"), opinion terms (e.g., "awful"), and sentiment polarities (e.g., "negative") [Zhang et al. 2022]. This allows for a more detailed and accurate analysis of user opinions, especially in complex reviews.

Most research in ABSA has focused on resource-rich languages, especially English [Zhang et al. 2022], leaving a gap in the development and application of ABSA techniques for less commonly studied languages like Portuguese. This limitation hinders the exploration of ABSA in specific domains where Portuguese is predominantly used, such as customer reviews in the Brazilian market.

Recent research, including the comprehensive survey by [Zhang et al. 2022], has identified several challenges in advancing ABSA. One critical issue is the limited scope

of existing ABSA datasets, with a significant portion of the literature relying on the SemEval Benchmark, which often fails to capture the diversity of user opinions as expressed across various platforms. User feedback can take many forms, from structured reviews to conversational dialogues in customer service interactions or question-answering forums. To build more robust ABSA systems, there is a clear need for datasets that encompass this broader spectrum of opinion sources, especially in languages other than English.

This research aims to address these gaps by assessing the efficacy of fine-tuning open-source small language models (SLMs) for the End-to-End Aspect-Based Sentiment Analysis (E2E-ABSA) task on Portuguese customer reviews. Large Language Models (LLMs) were not fine-tuned due to their high computational demands and deployment challenges, making SLMs a more practical and feasible choice. Recent research also shows that SLMs can outperform LLMs in specific tasks when fine-tuned with domain-specific data, further supporting their selection for this study [Hsieh et al. 2023]. Specifically, we employ instruction tuning, a form of supervised fine-tuning where language models are trained on datasets consisting of instruction-output pairs, to enhance the models' ability to follow human instructions effectively [Zhang et al. 2023].

To achieve this, we collected a diverse dataset of 5,000 customer reviews from multiple platforms, including Google Reviews, TripAdvisor, Instagram, and iFood, all written in Portuguese. This dataset captures a broad spectrum of user opinions from diverse sources, making it more reflective of the variability encountered in real-world applications within Portuguese-speaking markets. The dataset size was decided based on [Zhou et al. 2023] work showing that fine-tuning a large language model with just 1,000 curated examples can yield strong results. Given our use of smaller models, we increased the quantity to 5,000 to ensure robust performance.

Our research focuses on a compound ABSA task, where multiple aspects and their corresponding sentiments are analyzed simultaneously within each review. This approach not only addresses the complexity of multi-aspect sentiment analysis but also aligns with the need for more unified models in the field [Zhang et al. 2022].

To evaluate the effectiveness of this approach, we fine-tuned and compared three small language models: PTT5 Base [Carmo et al. 2020], FLAN-T5 Base [Chung et al. 2024], and mT0 Small [Muennighoff et al. 2023] that are widely used in academia and industry [Zhang et al. 2023].

Contributions: (a) We introduce the application of instruction tuning for Portuguese E2E-ABSA, leveraging a diverse dataset sourced from multiple platforms, including Google Reviews, TripAdvisor, Instagram, and iFood. (b) Our research demonstrates that the PTT5 Base model, a Portuguese-specific T5 variant, outperforms multilingual models such as FLAN-T5 Base and mT0 Small in the E2E-ABSA task, highlighting the value of language-specific pretraining. (c) We provide insights from a comprehensive evaluation of the models' performance that can guide future ABSA research.

2. Related Work

This research focuses on the E2E-ABSA task, aiming to jointly predict both aspects and their respective polarities—a Compound ABSA task. Previous work in ABSA has employed various modeling paradigms, such as Sequence-level Classification, Token-level

Classification, Machine Reading Comprehension (MRC), Sequence-to-Sequence modeling (Seq2Seq), and Pipeline [Zhang et al. 2022]. This study specifically adopts the Sequence-to-Sequence paradigm and incorporates instruction tuning, a technique proven effective in preparing language models for downstream tasks. Instruction tuning has demonstrated notable performance and generalization capabilities, as seen in various tasks, including the improvement of a language model for arithmetic tasks compared to GPT-4 [Mishra et al. 2022, Liu and Low 2023, Zhang et al. 2023].

For more challenging problems, such as Aspect Sentiment Quadruple Prediction (ASQP) [Zhang et al. 2022], a straightforward solution would involve developing four separate models, each dedicated to a specific subtask, and then combining them into a pipeline. However, this approach encounters the challenge of error propagation, as each model's accuracy is tied to the previous ones [Zhang et al. 2022]. The PARAPHRASE model [Zhang et al. 2021a] formulates the ASQP task as a Seq2Seq problem [Zhang et al. 2021a]. This model exhibited superior performance compared to the pipeline approach, which was previously mentioned. Similarly, [Varia et al. 2023] developed a Seq2Seq model that adeptly learns all ABSA subtasks through multitask learning, concurrently sampling input-output pairs from the tasks. This approach surpassed the prior state-of-the-art PARAPHRASE model in average F1 score across all datasets.

Those approaches, however, do not employ a task definition or examples of input-output pairs in the instruction. To deal with that, [Scaria et al. 2024] proposed a Seq2Seq approach that augments the input instructions with a task definition and includes input-output pairs containing examples of positive, negative, and neutral sentiments. The resulting model, InstructABSA, outperformed prior methods in Aspect Term Extraction (ATE), Aspect Term Sentiment Classification (ATSC), and E2E-ABSA while using only 20% of the training data compared to [Varia et al. 2023]. Our research builds on this approach by incorporating diverse sentiment examples to enhance model instructions. In addition, aiming to overcome a limitation in previous ABSA studies, which predominantly focused on rich-resource languages like English [Zhang et al. 2022], this research endeavors to extend recent methodologies to ABSA tasks in texts written in Portuguese.

Despite the significant improvements in performance due to instruction tuning methods, this approach encounters problems in specific scenarios. As highlighted by the study of [Kung and Peng 2023], these enhancements may result from the model learning superficial patterns such as output format and guessing. It was noted by [Scaria et al. 2024] that their model's performance decreased by approximately 10% when incorrect input-output mappings were incorporated in the instruction, a manifestation of a delusive example.

In addition to the findings of [Pires et al. 2023], which propose that language models specialized for the target language, such as Portuguese, can achieve near-state-of-the-art performance with significant reductions in training costs, this study aims to build upon this knowledge. The investigation involves comparing the PTT5 Base, a T5 model pretrained on Portuguese data, and multilingual models such as FLAN-T5 Base and mT0 Small, evaluating their effectiveness in the Portuguese E2E-ABSA task.

In the recent Aspect-Based Sentiment Analysis in Portuguese (ABSAPT) 2022 competition [Gomes et al. 2023], the focus was on Aspect Term Extraction (ATE) and

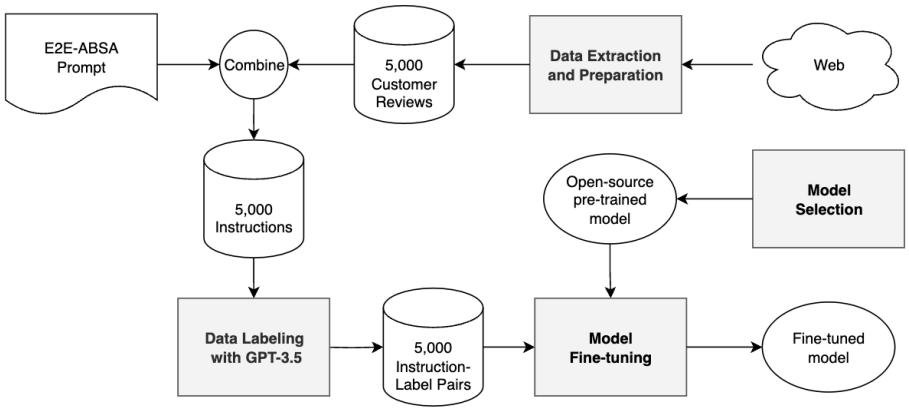

Figure 1. Methodology overview.

Sentiment Orientation Extraction (SOE) as distinct tasks rather than on their joint prediction. Our study, in contrast, emphasizes the integration of aspect and sentiment predictions in a Compound ABSA task, exploring a more holistic approach to sentiment analysis.

3. Methodology

This section outlines the methodology used in this research. We started by collecting and preparing a diverse dataset of customer reviews written in Portuguese from various platforms on the web, using tools like Apify. Next, we used GPT-3.5 to label the data with prompts designed for E2E-ABSA. We then selected and fine-tuned open-source language models using the labeled dataset. Figure 1 illustrates the entire process.

3.1. Data Extraction and Preparation

The dataset comprises 5,000 sentences from 4,872 restaurant reviews across diverse platforms, such as Google Reviews, TripAdvisor, Instagram, and iFood, ensuring a broad representation of user opinions. Figure 2 illustrates the distribution of the collected data. Figure 3 indicates the top keywords mentioned by users after applying filters to maintain only nouns and proper nouns, which are more closely related to topics.

After gathering the data, the subsequent step involved cleaning it. This cleaning process demanded eliminating unnecessary elements from the text, addressing situations where customers provide only a star rating without any accompanying textual review, and handling other related tasks as needed. The cleaned customer reviews were segmented into individual sentences because comments can contain multiple sentences. This strategy aims to reduce noise in the model input, acknowledging that a single comment may contain positive remarks about the food and negative feedback about the service, for instance.

3.2. Data Labeling with GPT-3.5

The process of supervised fine-tuning necessitates accurately labeled datasets. Labeling can be a formidable task, especially for languages with limited resources, due to its time-intensive and costly nature. The study leveraged the GPT-3.5 model for efficient dataset labeling to mitigate associated challenges. This approach aligns with the work by

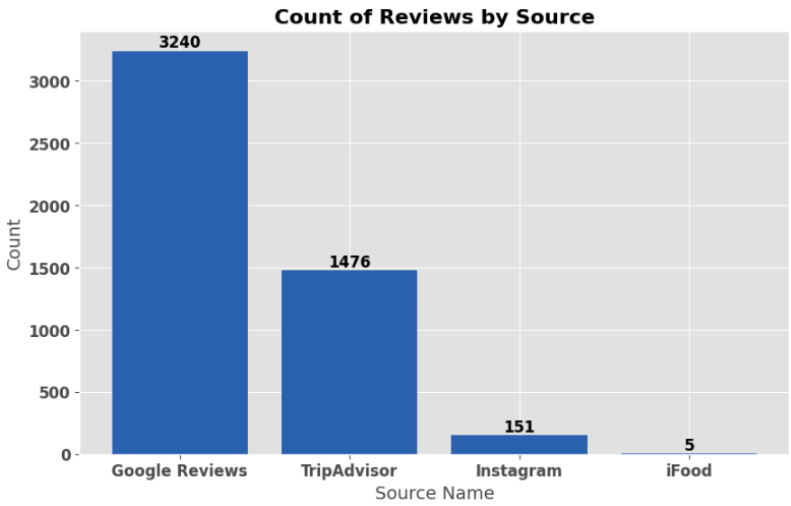

Figure 2. Distribution of User Reviews Across Various Platforms. The bar graph displays the frequency of reviews collected from multiple sources, with Google Reviews contributing the majority, followed by TripAdvisor. Reviews from Instagram and iFood are significantly fewer, reflecting the data collection process rather than the popularity or user preference for these platforms.

[Wang et al. 2021], which shows that GPT-3 can generate labels with a performance level on par with human-labeled data. Additionally, the work by [Tan et al. 2024] provides a comprehensive survey on the use of large language models for data annotation, further supporting the effectiveness of this approach in achieving high-quality annotations.

Prompt engineering was essential in this process. The research entailed developing and iteratively testing a series of prompts to guide GPT-3.5. These prompts described the E2E-ABSA task, accompanied by examples of human-labeled sentences and their expected outputs. Designed to minimize ambiguity and direct the model's focus towards explicitly mentioned aspects within the dataset, this methodology—inspired by and expanding upon the work of [Scaria et al. 2024]—ensured the proper identification of each aspect and sentiment extracted from the sentences.

GPT-3.5 and the more expensive GPT-4 were compared to select the most suitable model for this study. Ultimately, GPT-3.5 was chosen based on its satisfactory performance in meeting the task's quality criteria. This paper does not cover the detailed cost-effectiveness analysis between the two models. The decision prioritized a balance between cost and performance, focusing on qualitative output evaluation.

Figures 4 and 5 provide an analytical view of the data, revealing dominant sentiment polarities and frequently extracted aspects. These figures underscore patterns in customer experiences, with service quality emerging as a pivotal theme.

3.3. Model Selection

Model selection focused on two main criteria: performance in prior studies and efficiency, emphasizing smaller sizes to ensure CPU compatibility, making them both cost-effective and accessible for real-world applications. The PTT5 Base model [Carmo et al. 2020], with 223 million parameters, was chosen for its specialized training on the brWaC corpus

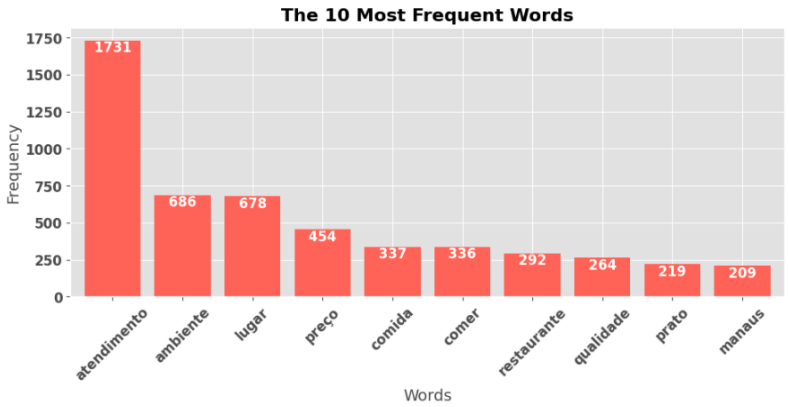

Figure 3. Prevalence of Key Terms in Customer Reviews. The bar chart quantifies the occurrences of the ten most frequent words in the dataset, with "atendimento" (service) leading, followed by "ambiente" (ambiance) and "lugar" (place).

[Wagner Filho et al. 2018], a large Brazilian Portuguese web dataset, making it particularly proficient in processing Portuguese texts.

The FLAN-T5 Base model [Chung et al. 2024], with 248 million parameters, has been fine-tuned on over 1000 tasks across multiple languages, including Portuguese. This extensive instruction fine-tuning makes it highly adaptable, achieving strong performance even compared to larger models.

The mT0 Small model [Muennighoff et al. 2023], with 300 million parameters, was evaluated for its multilingual capabilities. mT0 is based on the mT5 architecture, which is pretrained on a corpus derived from mC4, covering 101 languages, and is further enhanced through multitask prompted fine-tuning (MTF), which improves its performance across various languages.

3.4. Model Fine-tuning

The next phase involved supervised fine-tuning on the selected models, employing the instruction tuning paradigm—a methodology akin to the approach introduced by [Scaria et al. 2024]. The dataset was divided into training, validation, and testing subsets with proportions of 70%, 15%, and 15%, respectively. This allocation provided 3,500 examples for training and 750 examples each for validation and testing. The instruction tuning paradigm has demonstrated effectiveness across a diverse range of downstream tasks, as evidenced by several studies [Liu et al. 2023, Mishra et al. 2022, Yin et al. 2022, Zhang et al. 2023].

4. Experimental Evaluation

The evaluation approach in this study emphasizes the effectiveness of models in E2E-ABSA.

4.1. Setup

To ensure a fair comparison, both the model outputs and the labels are converted to lowercase. This normalization process mitigates discrepancies arising from case sensitivity, thus maintaining a focus on the semantic accuracy of the outputs.

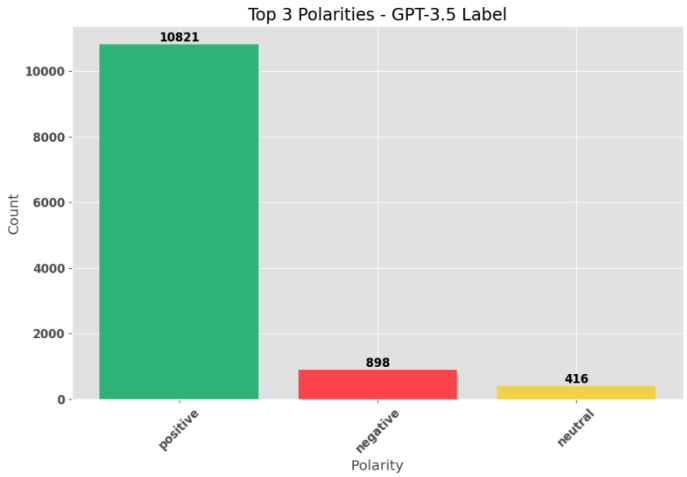

Figure 4. Distribution of Polarity in GPT-3.5 Label Annotations. The bar chart represents the frequency of the top three polarities as labeled by the GPT-3.5 model. The positive polarity is predominant, with 10,821 instances, underscoring a generally positive sentiment in the dataset. This is followed by negative and neutral polarities, with 898 and 416 instances, respectively, illustrating a significantly lower occurrence of these sentiments.

The evaluation stringently adheres to the exact match criterion. This means that for a model's output to be considered correct, it must precisely match the ground truth label in both the aspect and its associated sentiment. For example, if the GPT-3.5 label is "<*lugar,positivo*>" and the model output is "<*lugar agradável,positivo*>", it would not be considered a correct match under exact match criteria. However, this could be deemed correct under a partial match criterion due to the similarity in sentiment and context.

The key metrics for this study's evaluation are F1 Score, Precision, and Recall, all calculated based on the principle of an exact match between the model outputs and the ground truth labels, following established methodologies from previous works [Zhang et al. 2021b, Scaria et al. 2024].

4.2. Results and Analysis

The comparative performance of the PTT5 Base, FLAN-T5 Base, and mT0 Small models in ABSA tasks, presented in Table 1, considers their respective sizes and specific training. The PTT5 model, with 223 million parameters, benefited from specialized training on Portuguese data, which contributed to its top performance across all metrics. This highlights the advantages of language-specific pretraining in enhancing a model's capability to handle language nuances.

Table 1. Performance comparison of PTT5 Base, FLAN-T5 Base, and mT0 Small.

Model	F1	Precision	Recall
PTT5 Base	**60.30**	**61.55**	**59.10**
FLAN-T5 Base	58.86	58.67	59.06
mT0 Small	57.22	57.99	56.47

The FLAN-T5 Base, with 248 million parameters, showed good performance, coming in second after the PTT5 Base. The slightly larger size of the FLAN-T5 Base,

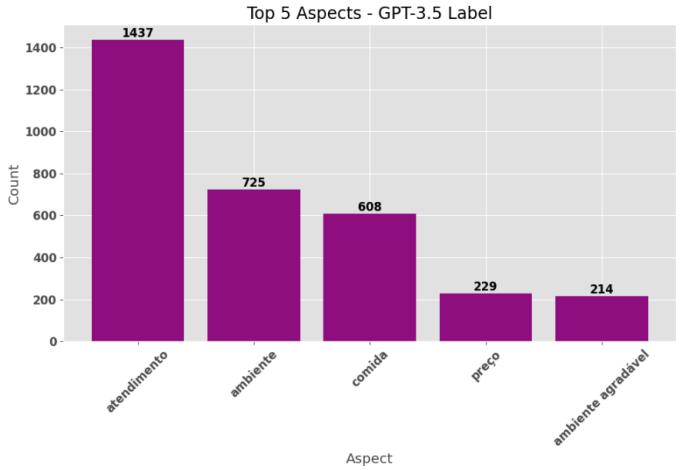

Figure 5. Bar chart showcasing the distribution of the top five aspects labeled by GPT-3.5. ”atendimento” (service) leads with 1437 mentions, followed by ”ambiente” (ambiance) at 725, and ”comida” (food) at 608.

coupled with its multilingual training, likely contributed to its broad adaptability and robust performance across different linguistic contexts.

The mT0 Small model, despite having the largest parameter count of 300 million, ranked third in performance. Several factors could explain this result. First, the larger model size does not necessarily translate to better performance in language-specific tasks, especially when compared to models like PTT5 Base, which are fine-tuned for a particular language. Second, the mT0 model's multilingual capabilities, while advantageous for general-purpose use across different languages, might not provide the same level of specialization required for the nuanced understanding of Portuguese as in the case of PTT5 Base.

Table 2. Comparison between the GPT-3.5 label and model outputs from PTT5 Base, FLAN-T5 Base, and mT0 Small for a review concerning the organization and pricing of a restaurant.

Sentence	ótimo lugar, suco muito bom, pena que no dia 07/12/2019 estava um pouco desorganizado e havia pouca informação sobre os preços.
GPT-3.5 Label	<lugar,positivo>, <suco,positivo>, <desorganizado,negativo>, <informação sobre os preços,negativo>
PTT5 Base Output	<lugar,positivo>, <suco,positivo>, <organizado,negativo>, <informação sobre os preços,negativo>
FLAN-T5 Base Output	<lugar,positivo>, <suco,positivo>, <informação sobre os preços,negativo>
mT0 Small Output	<lugar,positivo>, <suco,positivo>, <informação sobre os preços,negativo>

Upon analyzing the output examples shown in Table 2, it is evident that all three models — PTT5 Base, FLAN-T5 Base, and mT0 Small — adhered to the task format as instructed. However, the models diverged in their identification of aspects. The FLAN-T5 Base and mT0 Small models generated identical results, correctly identifying ”lugar” (place) and ”suco” (juice) with positive sentiments and marking ”informação sobre os preços” (information about the prices) negatively. However, they missed out on identi-

fying ”desorganizado” (unorganized) as having a negative sentiment. In contrast, PTT5 Base incorporated ”organizado” (organized), an aspect not explicitly present in the text, and assigned it a negative sentiment. This ability to infer aspects based on broader context demonstrates the model's advanced language understanding capabilities.

However, this case also underscores the important role of instruction in guiding model responses. The discrepancy between the model's output and GPT-3.5's explicit labeling of ”desorganizado” (disorganized) indicates a deviation influenced by PTT5 Base's interpretation of the instructions. This discrepancy highlights a potential area for refinement in instruction tuning to ensure that models can accurately distinguish between explicit and implied content within texts.

Table 3. Comparison between the GPT-3.5 label and model outputs from PTT5 Base, FLAN-T5 Base, and mT0 Small for a review concerning the size of a restaurant and waiting times.

Sentence	Apesar de ser um dos principais restaurantes de Manaus, o salão é pequeno e o tempo de espera por uma mesa pode passar de uma hora.
GPT-3.5 Label	<restaurante,positivo>, <salão pequeno,negativo>, <tempo de espera por uma mesa,negativo>
PTT5 Base Output	<restaurante,positivo>, <salão pequeno,negativo>, <tempo de espera por uma mesa,negativo>
FLAN-T5 Base Output	<restaurante,neutro>, <salão pequeno,negativo>, <tempo de espera por uma mesa,negativo>
mT0 Small Output	<salão pequeno,negativo>, <tempo de espera por uma mesa,negativo>

In Table 3, the PTT5 Base model aligns precisely with the GPT-3.5 label, effectively capturing positive and negative sentiments in the described restaurant scenario. In contrast, the FLAN-T5 Base model deviates slightly, labeling the ”restaurante” (restaurant) as neutral, suggesting a more subdued interpretation of the restaurant's quality. The mT0 Small model, however, omits the restaurant aspect, focusing only on negative elements, thus indicating a partial grasp of the sentence's sentiment. This comparison highlights these models' varying degrees of comprehension and sentiment analysis accuracy.

5. Limitations

While this research offers valuable insights into instruction tuning for Portuguese E2E-ABSA, it does have some limitations that should be acknowledged.

The study's focus on restaurant reviews means the findings may not encompass the broader complexities encountered in other domains. While the dataset used is extensive, it might not fully capture the rich linguistic diversity of Portuguese, potentially limiting the generalization of the results to other contexts.

Furthermore, this study concentrated on the overall E2E-ABSA task without separately analyzing the models' performance on more straightforward yet important tasks like Aspect Term Extraction (ATE) and Sentiment Orientation Extraction (SOE). Notably, these tasks form the core of the Aspect-Based Sentiment Analysis in Portuguese (ABSAPT) 2022 challenge, highlighting their significance in sentiment analysis research. A comparative evaluation of the models' capabilities in ATE and SOE tasks, using data

from the ABSAPT challenge, could have provided additional insights into each model's specific strengths and weaknesses.

Addressing these limitations in future research will enhance the understanding of small language models' capabilities in diverse sentiment analysis tasks, especially in the context of Portuguese, a language less frequently studied in ABSA research.

6. Conclusion

This study evaluates instruction tuning for small language models in the context of End-to-End Aspect-Based Sentiment Analysis (E2E-ABSA) in Portuguese. It highlights the capabilities and limitations of these models, notably the PTT5 Base, FLAN-T5 Base, and mT0 Small, in processing complex sentiment data. The PTT5 Base model, with its specialized training in Portuguese, showed promising results, achieving higher scores in F1, Precision, and Recall, indicating the potential benefits of language-specific training for ABSA tasks.

A notable observation is the tendency of models, especially PTT5 Base, to infer aspects not explicitly mentioned in the text. While this demonstrates advanced language understanding, it also suggests a divergence from the specific instructions. This study's result indicates room for further refinement in instructions provided to the models.

This research suggests that these models could be used for practical industry applications by incorporating simple filters to refine their output. These filters would enhance the relevance and accuracy of the results by focusing on aspects explicitly mentioned in the text, thus improving the practicality and reliability of the models in real-world scenarios.

7. Acknowledgments

This research is partially supported by the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES-PROEX) - Funding Code 001, by the Amazonas State Research Support Foundation (FAPEAM) through the POS-GRAD 2024/2025 project and the NeuralBond Project (UNIVERSAL 2023 Proc. 01.02.016301.04300/2023-04), and by the National Council for Scientific and Technological Development (CNPq) under the IAIA Project (406417/2022-9) and an individual grant to Altigran da Silva (307248/2019-4).

References

Carmo, D., Piau, M., Campiotti, I., Nogueira, R., and Lotufo, R. (2020). Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. *arXiv preprint arXiv:2008.09144*.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., and Wei, J. (2024). Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Gomes, J. R. S., Garcia, E. A. S., Junior, A. F. B., Rodrigues, R. C., Silva, D. F. C., Maia, D. F., da Silva, N. F. F., Soares, A. d. S., et al. (2023). Deep learning brasil at absapt 2022: Portuguese transformer ensemble approaches. *arXiv preprint arXiv:2311.05051*.

Hsieh, C.-Y., Li, C.-L., Yeh, C.-k., Nakhost, H., Fujii, Y., Ratner, A., Krishna, R., Lee, C.-Y., and Pfister, T. (2023). Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, Toronto, Canada. Association for Computational Linguistics.

Kung, P.-N. and Peng, N. (2023). Do models really learn to follow instructions? an empirical study of instruction tuning. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1317–1328, Toronto, Canada. Association for Computational Linguistics.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Liu, T. and Low, B. K. H. (2023). Goat: Fine-tuned llama outperforms gpt-4 on arithmetic tasks. *arXiv preprint arXiv:2305.14201*.

Mishra, S., Khashabi, D., Baral, C., and Hajishirzi, H. (2022). Cross-task generalization via natural language crowdsourcing instructions. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Le Scao, T., Bari, M. S., Shen, S., Yong, Z. X., Schoelkopf, H., Tang, X., Radev, D., Aji, A. F., Almubarak, K., Albanie, S., Alyafeai, Z., Webson, A., Raff, E., and Raffel, C. (2023). Crosslingual generalization through multitask finetuning. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Pires, R., Abonizio, H., Almeida, T. S., and Nogueira, R. (2023). Sabiá : Portuguese large language models. In *Intelligent Systems*, pages 226–240. Springer Nature Switzerland.

Scaria, K., Gupta, H., Goyal, S., Sawant, S., Mishra, S., and Baral, C. (2024). InstructABSA: Instruction learning for aspect based sentiment analysis. In Duh, K., Gomez, H., and Bethard, S., editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 720–736, Mexico City, Mexico. Association for Computational Linguistics.

Tan, Z., Beigi, A., Wang, S., Guo, R., Bhattacharjee, A., Jiang, B., Karami, M., Li, J., Cheng, L., and Liu, H. (2024). Large language models for data annotation: A survey. *arXiv preprint arXiv:2402.13446*.

Varia, S., Wang, S., Halder, K., Vacareanu, R., Ballesteros, M., Benajiba, Y., Anna John, N., Anubhai, R., Muresan, S., and Roth, D. (2023). Instruction tuning for few-shot aspect-based sentiment analysis. In Barnes, J., De Clercq, O., and Klinger, R., editors, *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 19–27, Toronto, Canada. Association for Computational Linguistics.

Wagner Filho, J. A., Wilkens, R., Idiart, M., and Villavicencio, A. (2018). The brWaC corpus: A new open resource for Brazilian Portuguese. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Wang, S., Liu, Y., Xu, Y., Zhu, C., and Zeng, M. (2021). Want to reduce labeling cost? GPT-3 can help. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yin, W., Li, J., and Xiong, C. (2022). ConTinTin: Continual learning from task instructions. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3062–3072, Dublin, Ireland. Association for Computational Linguistics.

Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., Li, J., Hu, R., Zhang, T., Wu, F., et al. (2023). Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Zhang, W., Deng, Y., Li, X., Yuan, Y., Bing, L., and Lam, W. (2021a). Aspect sentiment quad prediction as paraphrase generation. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9209–9219, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhang, W., Li, X., Deng, Y., Bing, L., and Lam, W. (2021b). Towards generative aspect-based sentiment analysis. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 504–510, Online. Association for Computational Linguistics.

Zhang, W., Li, X., Deng, Y., Bing, L., and Lam, W. (2022). A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Trans. on Knowl. and Data Eng.*, 35(11):11019–11038.

Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., YU, L., Zhang, S., Ghosh, G., Lewis, M., Zettlemoyer, L., and Levy, O. (2023). Lima: Less is more for alignment. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 55006–55021. Curran Associates, Inc.