Procesamiento del Lenguaje Natural, Revista nº 73, septiembre de 2024, pp. 315-322 recibido 03-07-2024 revisado 23-07-2024 aceptado 25-07-2024

Overview of ABSAPT at IberLEF 2024: Overview of the Task on Aspect-Based Sentiment Analysis in Portuguese

Resumen de ABSAPT en IberLEF 2024: Resumen de la Tarea sobre Análisis de Sentimientos Basado en Aspectos en Portugués

Alexandre Thurow Bender,[*,1,2] **Gabriel A. Gomes**[*,1,2]**, Emerson P. Lopes**[1,2]**,**
Ricardo M. Araujo[1,2] **Larissa A. de Freitas**[1,2]**, Ulisses B. Corrêa**[1,2]
[*]These authors contributed equally.
[1]Center for Technological Development (CDTec)
[2]Artificial Intelligence Innovation Hub (H2IA)
Postgraduate Program in Computing
Federal University of Pelotas, RS, Brazil
{atbender, gagomes, eplopes, ricardo, larissa, ulisses}@inf.ufpel.edu.br

Abstract: This paper reports the results of a competition on Aspect-Based Sentiment Analysis in Portuguese at ABSAPT 2024. The ABCD Team participated in the Aspect Extraction sub-task, using a BIO tagging scheme and Transformer-based models, with the best performance from the BERTimbau Large model. The teams ABCD and UTFPR participated in the Aspect Sentiment Classification sub-task. The UTFPR Team achieved the best results by fine-tuning a BERT model and using data augmented by ChatGPT to expand the training set. The ABCD Team experimented with both generation and classification approaches, with the classification method performing best. This competition showed that Transformer models and data augmentation are interesting techniques for improving Aspect-Based Sentiment Analysis in Portuguese.
Keywords: Aspect-Based Sentiment Analysis, Portuguese, ABSAPT.

Resumen: Este artículo informa sobre los resultados de una competencia de Análisis de Sentimientos Basado en Aspectos en portugués en ABSAPT 2024. El equipo ABCD participó en la sub-tarea de Extracción de Aspectos, utilizando un esquema de etiquetado BIO y modelos basados en Transformers, con el mejor rendimiento obtenido por el modelo BERTimbau Large. Los equipos ABCD y UTFPR participaron en la sub-tarea de Clasificación de Sentimientos de Aspectos. El equipo UTFPR logró los mejores resultados afinando un modelo BERT y utilizando datos aumentados por ChatGPT para expandir el conjunto de entrenamiento. El equipo ABCD experimentó con enfoques tanto de generación como de clasificación, siendo el método de clasificación el que mejor rendimiento obtuvo. Esta competencia mostró que los modelos Transformers y la ampliación de datos son técnicas interesantes para mejorar la Análisis de Sentimientos Basado en Aspectos en portugués.
Palabras clave: Análisis de Sentimiento Basado en Aspectos, portugués, ABSAPT.

1 *Introduction*

Sentiment Analysis (SA) is a critical domain within Natural Language Processing (NLP) that focuses on deciphering the sentiments or opinions that individuals express regarding various entities. This analysis provides invaluable insights into consumer emotions and opinions about products, services, or concepts, aiding decision-making processes for businesses and governmental bodies (Liu, 2020).

SA typically explores sentiments at varying levels of detail, primarily categorized into three levels: document, sentence, and aspect levels (Freitas, 2015). Aspect-level analysis, in particular, offers a nuanced view by dissecting specific opinions related to various facets of a single entity or across multiple en-

ISSN 1135-5948 DOI 10.26342/2024-73-23

tities within the same context.

Building on the previous edition, the ABSAPT 2024 task continues to challenge participants to innovate and refine techniques that adeptly extract and classify aspect-based sentiments from textual data. The focus is on enhancing methods for aspect extraction and sentiment classification within the domain of hotel reviews. This work is part of the IberLEF (Chiruzzo, Jiménez-Zafra, y Rangel, 2024) shared evaluation campaign of natural language processing systems in Spanish and other Iberian languages, including Portuguese, Catalan, Basque, and Galician.

This paper provides a comprehensive overview of the task set for the participants of this edition. Initially, we delve into some theoretical frameworks and advancements in Aspect-Based Sentiment Analysis (ABSA) in Section 2. Section 3 outlines the specific challenges and objectives of the task. Subsequent sections detail the dataset, annotation methodologies (Section 4), evaluation criteria (Section 5), and finally, the analysis of participant methodologies and their results (Section 6). The paper concludes with final remarks and potential future directions in Section 7.

2 Aspect-Based Sentiment Analysis

Aspect-Based Sentiment Analysis focuses on dissecting the nuanced opinions expressed toward various attributes or aspects of an entity, enabling a more detailed understanding of sentiments within the text. ABSA operates on a specific level of granularity where each opinion related to an aspect is analyzed independently. Effective execution of ABSA typically involves two primary tasks: Aspect Extraction (AE) and Aspect Sentiment Classification (ASC).

2.1 Aspect Extraction

The first task, Aspect Extraction, involves identifying the specific aspects mentioned in a text related to a given entity. This step sets the foundation for further sentiment analysis by pinpointing the exact features or properties being discussed.

For instance, consider the review excerpt: "The hotel had a fantastic breakfast!" Here, the AE task would entail recognizing "breakfast" as the key aspect being evaluated by the reviewer.

2.2 Aspect Sentiment Classification

Following the identification of aspects, Aspect Sentiment Classification defines the problem of determining the sentiment polarity — positive, negative, or neutral — associated with each identified aspect. This process is crucial for understanding the tone of opinions toward each aspect.

Using the earlier example: once "breakfast" is identified as an aspect, the sentiment associated with it is assessed. In this context, the sentiment would likely be classified as positive, indicated by the context "fantastic breakfast".

Together, these tasks allow researchers and practitioners to train models to obtain a refined view of the sentiments expressed in textual data, providing deeper insights into consumer attitudes and preferences specific to different aspects of a product or service.

3 Task Description

The extraction and analysis of public opinions are invaluable not only for individuals seeking guidance but also for organizations, both public and private, that rely on feedback to refine their services and products. While sentiment analysis at the document level is prevalent in studies of Portuguese texts, there is a notable gap in resources and methodologies for Aspect-Based Sentiment Analysis in the Portuguese language.

To address this gap, we introduce a task-centered around ABSA for reviews on TripAdvisor written in Portuguese. This task is divided into two specific sub-tasks: Aspect Extraction and Aspect Sentiment Classification. The first sub-task, AE, focuses on identifying different aspects mentioned within the reviews. The second sub-task, ASC, aims to determine the sentiment polarity—positive, negative, or neutral—associated with each identified aspect.

The scarcity of Portuguese-language corpora significantly hinders advancements in Natural Language Processing for this linguistic context. By introducing these sub-tasks, our goal is to foster the development of new tools and methodologies that will advance Portuguese NLP.

This initiative is inspired by previous ABSA challenges such as those featured in

SemEval (Pontiki et al., 2014; Pontiki et al., 2015; Pontiki et al., 2016) and EVALITA (De Mattei et al., 2020), which have significantly contributed to the evolution of ABSA in other languages. Through this task, we aim to catalyze similar progress for the Portuguese language, meeting an urgent need for enhanced computational resources and research in this area.

4 Corpora Description

The Corpus created for the ABSAPT 2024 competition is composed of data from the ABSAPT 2022 (da Silva et al., 2022) competition with additional annotations. The original data comprises the datasets from (Freitas, 2015) and (Corrêa, 2021).

The chosen corpus for the competition comprises traveler reviews collected from the TripAdvisor website, specifically focusing on hotels located in New York, Las Vegas, Paris, and Porto Alegre. The entire dataset is written in Portuguese. Following the ABSAPT 2022 competition, new data annotation was conducted to expand the corpus, resulting in a total of 2,382 additional annotations. The training and test data separation was carried out ensuring that each unique text appears only in the training set or the test set to prevent data leakage. We allocated 30% of the unique texts to the test set and the remainder to the training set.

Furthermore, it is important to highlight that for the test data, only newly annotated data were used. This decision was made because the data used in the previous competition had already been made public, making their use for competitor evaluation unfeasible. The remaining data were used in the training set.

The test data were divided for the two tasks, following the same logic as the previous division, where annotations were separated by unique reviews, splitting the data 50/50.

Table 4 presents the size of the dataset allocated for training, and table 4 presents the size of the data allocated for test of Task 2, showing the number of annotations, unique reviews, and unique aspects. It is important to mention that the annotations are tuples of review and aspect; thus, a review can be repeated in the annotations due to having more than one aspect mentioned in the text.

Attribute	Amount
Annotations	4828
Reviews	1320
Aspects	82

Table 1: Size of the Training Corpus.

Attribute	Amount
Annotations	1176
Reviews	282
Aspects	80

Table 2: Size of the Test Corpus for Task 2.

The new version of the annotation set was annotated following the same pattern as the previous one (Freitas, 2015), using concepts created by the Accommodation Services Domain Ontology, HOntology (Chaves, Freitas, y Vieira, 2012) to define the aspects. HOntology has 282 concepts categorized into hierarchies with a maximum depth level of 5.

In Figure 1, we present the sentiment polarity distribution for the training set, and in Figure 2, we present the same information for the test set for Task 2. We can observe that the dataset is imbalanced, especially for the training set, where the number of positive polarity data is greater than the sum of negative and neutral polarity data. For the test set, we have a larger amount of negative polarity data, although still less than the positive amount, with the neutral class having the smallest quantity. It is important to identify this imbalance to adopt the best error metrics for evaluating the competitors, indicating the need to use metrics sensitive to imbalanced data.

In Figures 3 and 4, we observe the sentiment polarity distribution for each aspect in percentage for the training set and Task 2 test set, respectively. We selected the top 40 most mentioned aspects for the training set, and for the test set, we selected all aspects that had more than 10 mentions.

5 Evaluation Measures

The aspect extraction subtask can be formulated as an information retrieval problem in which participants extract (retrieve) aspects (documents). There are well-defined metrics to evaluate document retrieval tasks, such as recall and precision. Recall is used to mea-

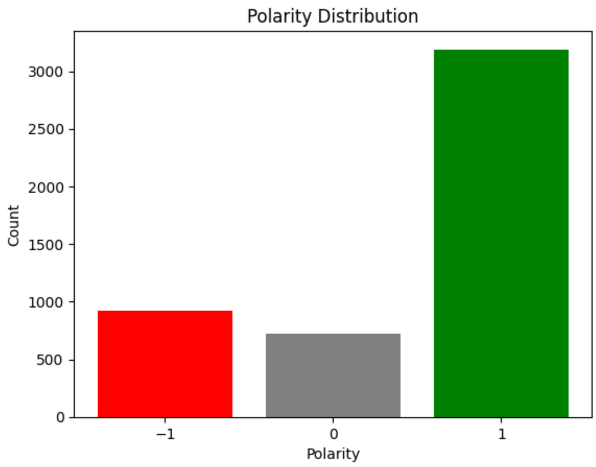

Figure 1: Sentiment Polarity Distribution Plot for the Training Set.

.

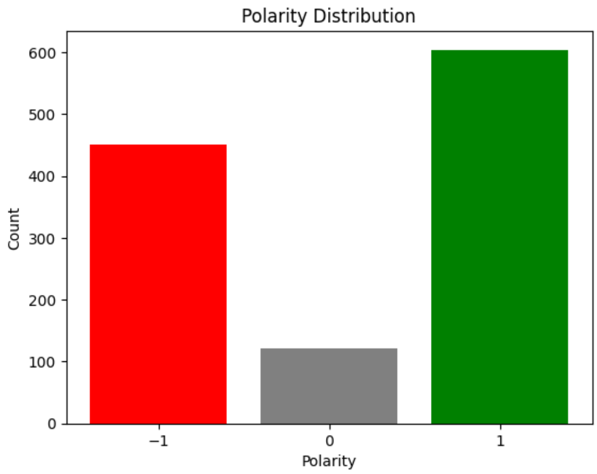

Figure 2: Sentiment Polarity Distribution Plot for the Test Set of Task 2.

sure the fraction of relevant documents that are successfully retrieved; in our particular task, this relates to the amount of aspects in the ground-truth annotation that were correctly extracted.

$$\text{Recall} = \frac{|\{\text{Relevant}\} \cap \{\text{Retrieved}\}|}{|\{\text{Relevant}\}|} \tag{1}$$

Recall is a useful metric, but it may yield unreliable results in the scenario where models predict every single aspect for all examples, as this will trivially maximize the recall metric. Its counterpart is the precision metric, which measures the fraction of the retrieved documents that are relevant. This penalizes the retrieval of aspects that are not present in the ground truth.

$$\text{Precision} = \frac{|\{\text{Relevant}\} \cap \{\text{Retrieved}\}|}{|\{\text{Retrieved}\}|} \tag{2}$$

These two metrics complement each other effectively. However, to have a single unified metric, the F-measure was introduced. It is the harmonic mean of precision and recall.

$$F = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

The aspect sentiment classification subtask is a ternary classification problem between the classes positive, neutral, and negative. Being ASC a notoriously imbalanced problem, we opt to use the balanced accuracy metric (BACC), which provides a robust approach to calculating accuracy whenever problems are not balanced. It is based on the classification recall and specificity formulas. While recall measures the true positive ratio, specificity measures the true negative ratio.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{5}$$

$$\text{BACC} = \frac{\text{Recall} + \text{Specificity}}{2} \tag{6}$$

6 *Participants Systems and Discussion of the Results*

Two teams participated in this task. In the Aspect Extraction sub-task, only one participant team submitted the results, and both teams submitted their results for the Aspect Sentiment Classification sub-task.

6.1 Aspect Extraction

ABCD Team was the only participant in the AE sub-task. The results are shown on table 6.1. Their methodology consisted of modeling the task as a token-classification problem, where each token is assigned a label in the BIO tagging scheme, in which the first token of each aspect receives the "B" label, the tokens that belong to an aspect, but are not the first, receive the "I" label, and all tokens that do not belong to any aspect are labeled as "O". Then, Transformers based models were fine-tuned to predict the correct label for each token in a given example.

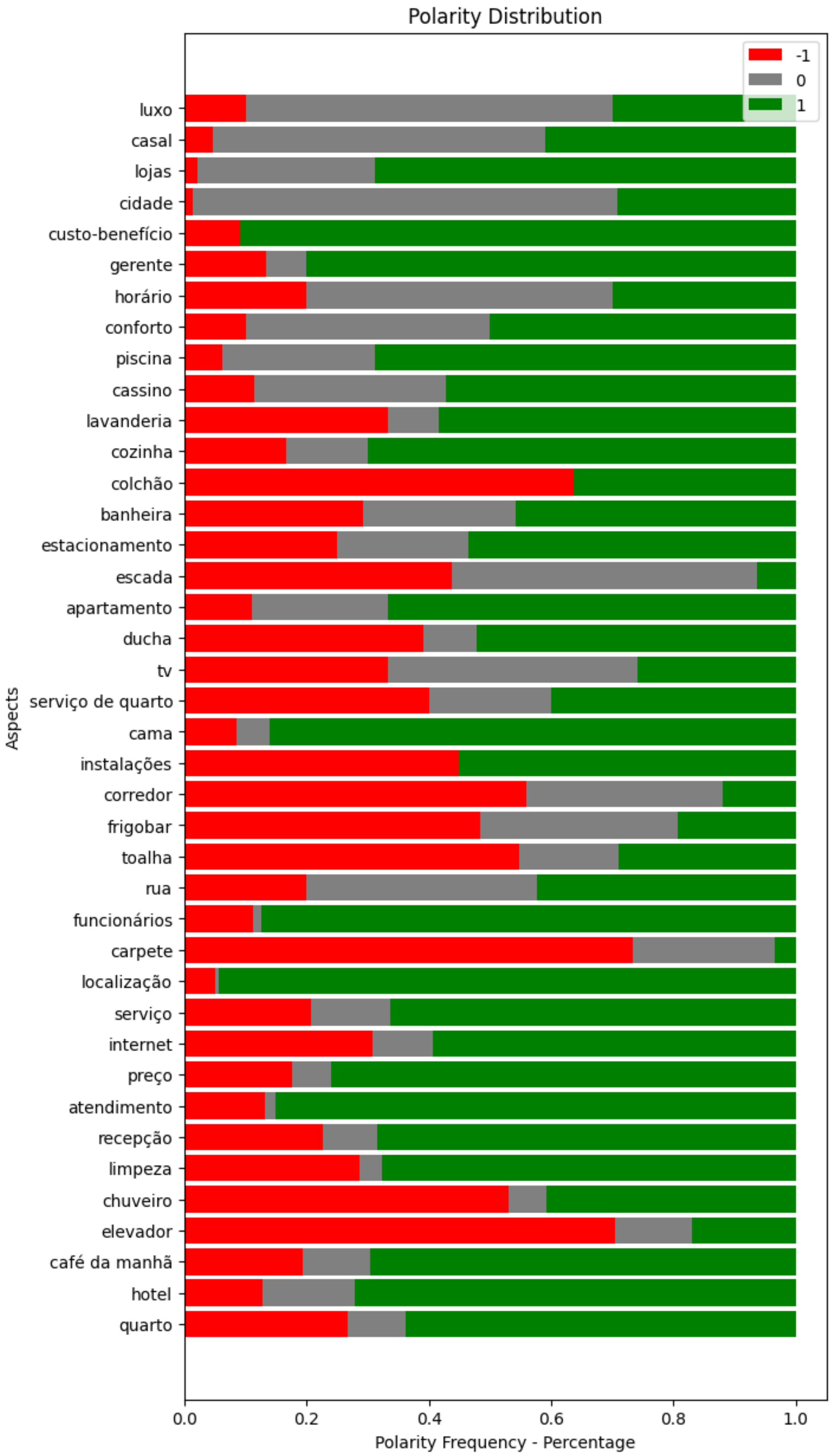

Figure 3: Aspect Sentiment Polarity Distribution Plot for the Training set.

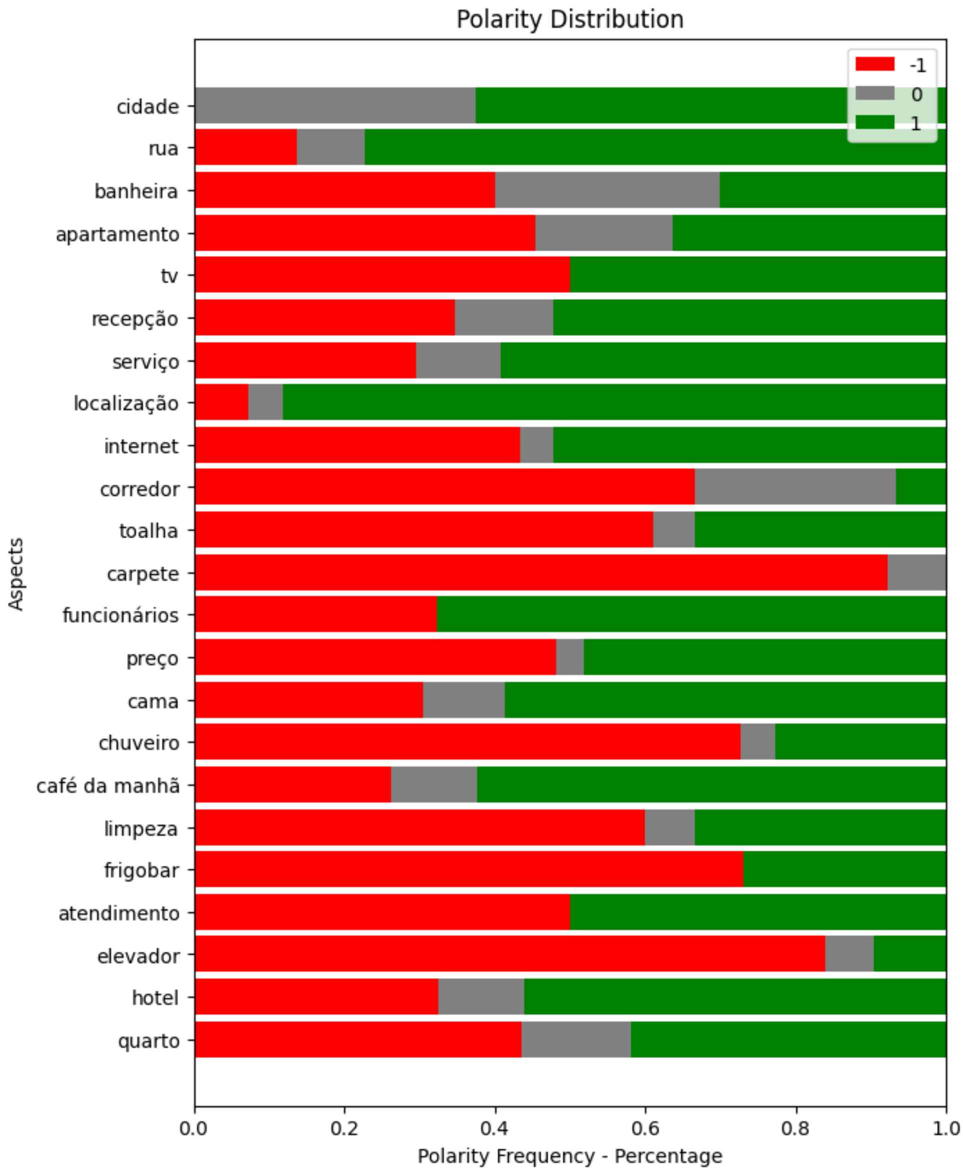

Figure 4: Aspect Sentiment Polarity Distribution Plot for the Test set of Task 2.

The selected pre-trained models included models trained for Brazilian Portuguese and also multilingual models. The best result was obtained by the BERTimbau Large model, a model that is trained exclusively for the Portuguese.

Team	Prec	Recall	F1
ABCD	0.855	0.730	0.637

Table 3: Results of the AE sub-task.

6.2 Aspect Sentiment Classification

Both teams submitted their results in the ASC sub-task, with the UFPR Team obtaining the best results, as shown in table 6.2.

The UFPR Team obtained the best results, using an approach based on fine-tuning a BERT based model for a sentence pair classification task, using an augmented dataset, that was automatically generated by ChatGPT. The data augmentation process consisted of prompting ChatGPT for the generation of new reviews, with their expected annotation (of aspects and their polarities), and excluding examples with badly generated aspects. Using this approach, 1281 new exam-

Team	BAcc	Prec	Recall	F1
UFPR	0.652	0.656	0.652	0.653
ABCD	0.571	0.569	0.571	0.568

Table 4: Results of the ASC sub-task.

ples were generated, augmenting the training dataset from 4828 to 6109 examples.

ABCD Team evaluated both generation and classification approaches, using fine-tuned Transformers models in them. The generation based approach consisted in fine-tuning models for generation of the "name" of the labels (negative, neutral, or positive), instead of the numeric class. The fine-tuning procedure was also divided into a single task approach, in which the model was tasked with generating only the polarity, and a multi task approach, that included the task instruction in the context for generation, and then the same model could be fine-tuned for both AE and ASC tasks.

In the classification approach, which obtained the best results for the evaluation set, the approach consisted in fine-tuning a pre-trained encoder model for a sequence pair classification task, in which the pair contains the review, followed by the aspect that is going to be classified.

7 Final Remarks

In this paper, we reviewed the methodologies and results of the participating teams in ABSAPT 2024, which included the Aspect Extraction and Aspect Sentiment Classification sub-tasks. The ABCD Team participated in the AE sub-task employing a token-classification approach and fine-tuning various transformers-based models. Their best performance came from the BERTimbau Large model. For the ASC sub-task, the UFPR Team led with a BERT-based model fine-tuned for a sentence pair classification task, utilizing data augmentation by generating new examples with ChatGPT.

The ABCD Team explored both generation and classification approaches for the ASC sub-task, with their classification approach yielding the best results. However, their overall performance was lower than that of the UFPR Team. The study shows the effectiveness of pre-trained transformer models and data augmentation in enhancing aspect-based sentiment analysis. Future work could include further evaluation of these techniques and the exploration of alternative approaches to ABSA for different low-resource languages to improve sentiment analysis systems.

Acknowledgments

Thank you to all annotators for their essential work. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research. This work was partly financed by the following Brazilian research agencies: CAPES and CNPq.

References

Chaves, M., L. Freitas, y R. Vieira. 2012. Hontology: a multilingual ontology for the accommodation sector in the tourism industry.

Chiruzzo, L., S. M. Jiménez-Zafra, y F. Rangel. 2024. Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages. En *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*.

Corrêa, U. B. 2021. Análise de sentimento baseada em aspectos usando aprendizado profundo: uma proposta aplicada à língua portuguesa.

da Silva, F. L., G. d. S. Xavier, H. M. Mensenburg, R. F. Rodrigues, L. P. dos Santos, R. M. Araújo, U. B. Corrêa, y L. A. de Freitas. 2022. Absapt 2022 at iberlef: Overview of the task on aspect-based sentiment analysis in portuguese. *Procesamiento del Lenguaje Natural*, 69:199–205.

De Mattei, L., G. De Martino, A. Iovine, A. Miaschi, M. Polignano, G. Rambelli, y others. 2020. Ate absita@ evalita2020: Overview of the aspect term extraction and aspect-based sentiment analysis task. En *CEUR WORKSHOP PROCEEDINGS*, volumen 2765, páginas 67–74. CEUR-WS.

Freitas, L. A. d. 2015. *Feature-level sentiment analysis applied to brazilian por-*

tuguese reviews. Ph.D. tesis. Faculdade de Informática.

Liu, B. 2020. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge university press.

Pontiki, M., D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. Al-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq, y others. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. En *ProWorkshop on Semantic Evaluation (SemEval-2016)*, páginas 19–30. Association for Computational Linguistics.

Pontiki, M., D. Galanis, H. Papageorgiou, S. Manandhar, y I. Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. En P. Nakov T. Zesch D. Cer, y D. Jurgens, editores, *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, páginas 486–495, Denver, Colorado, Junio. Association for Computational Linguistics.

Pontiki, M., D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, y S. Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. En P. Nakov y T. Zesch, editores, *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, páginas 27–35, Dublin, Ireland, Agosto. Association for Computational Linguistics.