Structured Sentiment Analysis in Brazilian Portuguese: An Exploratory Study Using BERTimbau

Andrew B. Campos[1], Ulisses B. Corrêa[1], Larissa A. de Freitas[1]
[1]Universidade Federal de Pelotas (UFPel), Pelotas, Brazil
{abdcampos, ulisses, larissa}@inf.ufpel.edu.br

Abstract

Structured Sentiment Analysis (SSA) aims to extract fine-grained opinion structures as tuples (holder, target, expression, polarity). While recent advances have improved SSA for English, Brazilian Portuguese lacks dedicated resources. This paper presents an exploratory study introducing a manually annotated dataset of hotel reviews in Brazilian Portuguese for SSA. We propose a baseline approach fine-tuning the BERTimbau model under a BIO tagging scheme to extract sentiment spans. Unlike traditional approaches that model relations explicitly, we assess the viability of span-level extraction as a first step for SSA in this language. Experimental results using a strict train/validation/test split show that our approach achieves a span-level F1-score of 48.41 for holder extraction and a macro F1-score of 61.52. We also discuss the linguistic challenges of holder extraction in Portuguese, specifically regarding implicit subjects (pro-drop), and provide a detailed error analysis. These results establish a preliminary baseline for future relation-aware models in Portuguese.

1 Introduction

Structured Sentiment Analysis (SSA) aims to extract tuples composed of a holder, an aspect, an opinion expression, and its polarity from opinionated text. Recent challenges proposed by SemEval 2022[1] have highlighted growing interest in this task.

Structured Sentiment Analysis is an important subfield of sentiment analysis that, according to (Barnes et al., 2021), addresses limitations of traditional sentiment analysis approaches by integrating multiple interrelated subtasks. Rather than focusing solely on extracting aspect terms or sentence-level polarity, SSA combines these tasks into a single task.

BERT is a bidirectional transformer model pre-trained on large amounts of unlabeled text using objectives that include masked language modeling and next-sentence prediction. It is the very base of models trained in Portuguese corpora like Albertina 100M PTBR(Rodrigues et al., 2023) and MBert(Devlin et al., 2018).

A solid model for this domain is BERTimbau[2] , which had 92,791 downloads as of December 2025, indicating its wide adoption and strong presence in Portuguese Natural Language Processing (NLP) research.

In this context, this work aims to extract opinion tuples from a Portuguese dataset using BERTimbau.

The reviews in our dataset were withdrawn from the ASQP-PT 2025 (Lopes et al., 2025) corpus. The dataset[3] annotation was done manually by identifying the holder in 762 samples of hotel reviews extracted originally from TripAdvisor.

The remainder of this paper is organized as follows. Section 2 presents the theoretical background, covering structured sentiment analysis, the BERTimbau model, the Optuna framework, and the BIO tagging scheme. Section 3 describes the methodology, including the dataset, preprocessing steps, and evaluation metrics. Subsequently, Section 4 discusses the experimental results. Finally, Section 5 shows the final remarks, limitations, and directions for future work.

2 Theoretical Background

This section covers the essential concepts relevant to the present work. It covers Structured Sentiment Analysis, the BERTimbau model, the Optuna framework, and BIO Tag.

[1]Available in: https://competitions.codalab.org/competitions/33556

[2]Available in: https://huggingface.co/neuralmind/bert-base-portuguese-cased

[3]Available in: https://github.com/AndrewwBC/PROPOR-2026-SSA-BERTIMBAU

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 927–932
April 13-16, 2026

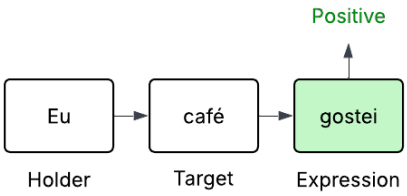

Figure 1: Example of a Structured Sentiment Analysis tuple.

2.1 Structured Sentiment Analysis

Structured Sentiment Analysis aims to extract tuples *(h, t, e, p)* where:

- *h* – is the **holder** who expresses the opinion towards a **target**;
- *t* – is the target of the opinion;
- *e* – is the **expression** of the opinion, associated with a **polarity**;
- *p* – denotes the polarity of the sentiment expressed by *h*.

An example is shown in Figure 1, where the holder is "Eu," the target is "café," the expression is "gostei," and the polarity is positive. A more complex example, which was extracted from our dataset: "O hotel em geral é bom, o que achamos ruim foi o tamanho do quarto e do banheiro."

- **Tupla 1:** (Holder: "achamos", Target: "hotel", Expression: "bom", Polarity: positiva)
- **Tupla 2:** (Holder: "achamos", Target: "tamanho do quarto", Expression: "ruim", Polarity: negativa)
- **Tupla 3:** (Holder: "achamos", Target: "tamanho do banheiro", Expression: "ruim", Polarity: negativa)

It is challenging to pinpoint the exact origin of the various tasks involved in sentiment analysis. One can trace the origin of Structured Sentiment Analysis to early works by Kim and Hovy (2004), where **holder** appears as the **person** or **organization** that holds an opinion, and Wiebe et al. (2005), where **holder** is the **person** or **entity** that expresses something.

More recently, and in contrast to earlier formulations such as Al-Mars et al. (2017), Barnes et al. (2021) proposed SSA as a structured prediction task based on dependency graph parsing. Despite methodological differences across these works, the underlying task remains the same.

2.2 BERTimbau Model

BERTimbau is a pre-trained Bidirectional Encoder Representations from Transformers (BERT) model for Brazilian Portuguese, which is why we chose it. The base architecture contains 115M parameters, 12 transformer layers, and a hidden size of 768. In addition, the large architecture contains 335M parameters, 24 transformer layers, and a hidden size of 1,024, increasing both precision and computer requirements (Souza et al., 2020). Due to its lower computer requirements, the base architecture was used in this work.

2.3 Hyperparameters Optimization with Optuna

Hyperparameters, such as batch size, learning rate, epochs, and weights are the central stone to achieve a good model performance. Finding a good set of values that fits a specific task can be time-consuming. To find the best combination of hyperparameter values, we use Optuna for hyperparameter optimization.

Optuna is an open-source framework that provides an automated search to yield the best possible model performance. It plays a vital role in the hyperparameter fine-tuning process by efficiently exploring the search space (Akiba et al., 2019).

2.4 BIO Tag

BIO (Beginning, Inside, Outside) Tagging is a strategy for labeling tokenized phrases to identify classes that the model should predict. Given a word, we need to determine that word as a **holder**, **target**, or **expression**. Note that, to maintain the SSA tuple, we extract both expression and polarity from the same word. An example of text that is mapped to BIO tagging is shown in the following items.

Sentence #1: O quarto é agradável.
BIO format: O, B-TARGET, O, B-EXP-POS.

Sentence #2: Meu marido e eu gostamos do café.
BIO format: B-HOLDER, I-HOLDER, O, B-HOLDER, B-EXP-POS, O, B-TARGET.

As previously shown, the prefixes **B**, **I**, and **O** indicate the **beginning**, **inside**, and **outside** of an entity, respectively. The prefix **I** denotes that a

token belongs to the same entity initiated by a token tagged with prefix **B**. As a consequence of this structure, the first occurrence of an element is always tagged with the prefix **B**, such as “quarto” in Sentence #1 or “café” in Sentence #2.

In this sense, “Meu” in Sentence #2 begins (**B-HOLDER**) the holder entity, which is continued by “marido” (**I-HOLDER**). To extract the polarity of a sentiment expression, we tag the expression with its corresponding polarity. Therefore, “agradável” in Sentence #1 is simultaneously labeled as a sentiment expression and assigned a positive polarity.

3 Related Works

In recent years, SSA has been explored through various approaches. (Barnes et al., 2021) proposed a neural graph parsing model based on Bidirectional Long Short-Term Memory (BiLSTM), treating SSA as a dependency graph parsing task and incorporating syntactic information, which led to improved performance.

More recently, (Masaling and Suhartono, 2024) employed RoBERTa and XLM-RoBERTa for SSA. Their approach extracts opinion **holders**, **targets**, and **expressions** using BIO-based sequence labeling, and infers polarities from the extracted opinion expressions. The proposed models outperformed the graph parsing baseline in span-level F1-scores for **holder**, **target**, **aspect**, and **expression**.

More recent work has explored joint extraction strategies for Structured Sentiment Quadruple Extraction (SSQE). Li et al. (Li et al., 2026) proposed a Boundary-Sensitive Token Pair Labeling framework (BTPL) that reformulates SSQE as a unified entity–relation extraction task. Instead of relying on BIO-based sequence labeling, their method labels boundary token pairs to simultaneously identify opinion **holders**, **targets**, **expressions**, **polarities**, and their relations.

By integrating syntactic dependency information via a Graph Convolutional Network (GCN) and employing Conditional Layer Normalization (CLN), their approach effectively handles long-span roles and overlapping sentiment structures, achieving state-of-the-art results on multiple benchmarks.

4 Methodology

The pipeline of this work is defined as follows:

- A manually annotated dataset was used to identify opinion holders in hotel reviews extracted from TripAdvisor[4] and written in Brazilian Portuguese. The dataset was originally proposed for a shared task[5], in which the original annotation schema included an aspect category instead.
- The next step was model selection. Since the dataset is written in Brazilian Portuguese, we adopted a pretrained language model suitable for this language.
- We preprocessed the dataset by segmenting each hotel review into tokens. The position of each token in the text was used to generate a BIO tagging scheme that assigns a label to each word in the original text.
- The model was optimized using macro span F1-score as the selection criterion during fine-tuning. In this context, macro span F1-score corresponds to the average F1-score of each tuple element. The final results were reported from the trial that achieved the highest macro span F1-score.

The following sections present a detailed description of those steps.

4.1 Dataset Annotation

To enable the extraction of opinion tuples *(h, t, e, p)*, a single annotator performed the annotations to meet the SSA tuple extraction format already mentioned. An annotation guide was followed to help identify entities in the text, mostly the **holder**.

The challenge of identifying the holder lies in its explicit or implicit manifestation, as demonstrated by (Wiebe et al., 2005) with the notion of nested sources. To address this issue, we defined two main approaches. The holder, "I" may explicitly, as in "I really enjoy the coffee.". The holder may implicitly, as in "Liked the coffee", "liked", where the **holder**, we chose to annotate the verb in its inflected form. Otherwise, in absence of even implicit holder, we anotated as null. This annotation choice addresses a specific linguistic feature of Portuguese: it is a pro-drop language where the subject (holder) is frequently omitted but morphologically marked in the verb’s desinence (e.g., "Gostei" implies "Eu"). While standard SSA conventions often annotate implicit holders as a special "Null" token or the

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

[5] https://sites.google.com/inf.ufpel.edu.br/asqp-pt-2025/home

"Author" node, explicitly tagging the inflected verb allows the span-based model to ground the holder in the textual surface, serving as a proxy for the implicit entity. We acknowledge this diverges from English-centric SSA guidelines but argue it provides a necessary adaptation for span-extraction models in Romance languages. In cases like "The coffee was good.", we annotate as null.

The resulting dataset comprises 3,438 SSA tuples and 64,068 tokens, including repetitions, across 762 samples distributed among the annotated reviews. A detailed breakdown of the elements is as follows: 2,877 null entries and 561 filled entries of **holder annotation**; 2,196 positive (POS), 1,181 negative (NEG), and 61 neutral (NEU) sentiment labels; and a total of 3,438 aspect and 3,438 sentiment annotations.

4.2 Preprocessing data

To train BERTimbau, we preprocessed each hotel review. We: (i) split each hotel review; (ii) created a numerically labeled representation of each text by mapping the words already defined as an element of the tuple *(h, t, e, p)* by their location defined on the dataset; (iii) labeled with a BIO tag to create a true reference.

4.3 Model

We fine-tuned the BERTimbau base model on our annotated dataset of hotel reviews.

Hyperparameter optimization was performed using Optuna with 20 trials defined, exploring learning rates ranging from 1×10^{-5} to 5×10^{-4} on a logarithmic scale, batch sizes of 4, 8, and 16, weight decay values between 0.0 and 0.1, training epochs ranging from 3 to 10, and warmup ratios between 0.0 and 0.2. The optimal hyperparameters achieved are shown on table 1.

Hyperparameter	Value
Learning Rate	3.44e-05
Batch Size	4
Weight Decay	0.0759
Epochs	9
Warmup Ratio	0.0351

Table 1: Optimal Hyperparameters Found via Optuna

This configuration achieved the optimal balance for span-level predictions, representing the average F1-score across all entities.

4.4 Experimental Setup

To mitigate data leakage, the dataset was randomly partitioned at the review level into training (70%), validation (15%), and test (15%) sets, ensuring that sentences from the same review did not appear across different splits. A 5-fold cross-validation approach was employed to optimize validation performance. The validation set was used during training and hyperparameter optimization with Optuna. The test set was used exclusively for the final evaluation reported in Section 5.

For tokenization, we used the standard BERTimbau tokenizer. To deal with subword tokenization, we used a tokenization function provided by the Hugging Face library. BIO labels were aligned by assigning the original label to the first subword token and masking the remaining subwords of the same word during loss computation, so that they do not affect training.

4.5 Evaluation

Both token-level and span-level metrics were used to evaluate the performance of the BERTimbau base model on this task. We employed `seqeval`[6], an open-source library commonly used for sequence labeling evaluation, to compute the evaluation metrics. Precision, Recall, and F1-score were reported to the span-level assessment, while Accuracy was additionally included in the token-level analysis.

More specifically, the method classification_report was used to compute the span-level metrics. In the span-level measure, a true positive is a prediction when all tokens are correctly predicted for each element of the tuple.

Note that we decided to consider both positive and negative expressions because the polarity is directly correlated with the expression.

5 Analysis of Results

In this section, we present the results obtained using the best checkpoint identified by Optuna. Span-level metrics are reported in the table 2.

[6] https://github.com/chakki-works/seqeval

Component	Precision	Recall	F1-Score
Holder	48.30	49.47	**48.41**
Target	70.48	75.07	**72.65**
Expression	62.47	64.71	**63.51**
Macro Avg	–	–	**61.52**

Table 2: Span-level results with 5-fold Cross-Validation

Despite achieving an F1 score of 48.41 for **holder**, our results are comparable to those reported in the existing literature. For instance, Masaling and Suhartono (2024) reported an F1-score of 51.0 using a RoBERTa-based model, as well as a graph-parsing F1 score of 43.8 on the MPQA dataset.

Similarly, Li et al. (2026) obtained span-level F1-score for **holder** ranging from 47.00 in their worst result to 62.50 in their best result. In contrast, our approach achieved a F1-score of **72.65** for **aspect** terms, which is comparable to the best result reported by Li et al. (2026), who achieved an F1-score of 75.50.

A high divergence was observed in expression detection, likely influenced by the imbalance in class distributions. In particular, the substantially lower F1-score for the neutral polarity deserves attention. In percentage terms, neutral expressions account for approximately 1.77% of the dataset, while positive expressions represent about 64% and negative expressions 34.34%. The results for polarity extraction are presented in Table 3.

Polarity	F1-Score
Positive	0.6737 ± 0.0118
Negative	0.4961 ± 0.0323
Neutral	0.0879 ± 0.0781
Macro Avg	**0.4192**

Table 3: Expression Polarity Classification Results

Barnes et al. reported average polarity F1-scores of 38.5 on the MPQA dataset and 44.5 on the DSunis dataset as their worst results. In comparison, although our macro-average F1-score is within a comparable range, these findings highlight a weakness in our polarity extraction approach, which was substantially penalized by the neutral polarity due to its low frequency and high variance.

6 Final Remarks, Limitations and Future Works

The lack of previous studies using Brazilian Portuguese datasets for this task creates a non-comparable scenario. Nevertheless, our results are close to those reported in the existing literature when compared with the relevant works previously mentioned. In particular, the holder F1-score, when compared with the results of Li et al. (2026) and Zhou et al. (2021), demonstrates that our model achieves comparable performance within the same range reported in those studies.

Our study presents a first step towards SSA in Portuguese, but several limitations must be addressed in future work. First, our current approach relies on BIO tagging to identify spans but does not explicitly model the dependency relations between them (e.g., linking a specific Holder to a specific Target). As noted by Barnes et al. (2021), full SSA requires predicting valid tuples. In this work, we assume a heuristic where entities within the same sentence boundary are candidates for linking, which suffices for simple sentences but fails in complex, multi-opinion reviews. Future work will implement relation-aware architectures, such as dependency graph parsing or generative sequence-to-sequence models.

Second, the dataset was annotated by a single expert. While this ensures internal consistency, we lack inter-annotator agreement (IAA) metrics to guarantee reproducibility. We plan to expand the annotation team to validate the gold standard.

Finally, although the Neutral class was retained in our experiments, its extreme class imbalance (less than 2% of samples) severely affected performance, leading to low F1-scores and high variance. Future work will explore data augmentation strategies to better model neutral expressions.

Future work should include comparisons of our dataset across different models to highlight performance similarities and differences. Furthermore, collecting new data or applying data augmentation techniques is crucial to mitigate class imbalance among entities in the dataset, thereby improving model performance. In this context, a revised annotation methodology should be adopted to enable the application of the Cohen's kappa measure[7] . Moreover, translating English datasets to Brazilian Portuguese in order to improve the range of possibilities. As a final remark, the application

[7] https://numiqo.com/tutorial/cohens-kappa

of this method to other models, such as the previously mentioned Albertina or MBert, should be considered to explore different model behaviors.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Diala Al-Mars, Lilja Øvrelid, and Erik Velldal. 2017. Structured sentiment analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*, pages 150–158, Copenhagen, Denmark.

Jeremy Barnes, Robin Kurtz, Stephan Oepen, Lilja Øvrelid, and Erik Velldal. 2021. Structured sentiment analysis as dependency graph parsing. In *arXiv preprint arXiv:2105.14504*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 1367–1373, Geneva, Switzerland.

You Li, Shaocong Zhang, Yuming Lin, Yongdong Lin, and Liang Chang. 2026. Extracting structured sentiment quadruples by labeling boundary token pairs for aspect-based sentiment analysis. *Expert Systems with Applications*, 296:129017.

Emerson P Lopes, Gabriel A Gomes, Alexandre Thurow Bender, Ricardo M Araujo, Larissa A de Freitas, and Ulisses B Corrêa. 2025. Overview of asqp-pt at iberlef 2025: Overview of the task on aspect-sentiment quadruple prediction in portuguese. *Procesamiento del Lenguaje Natural*, 75.

Nikita Ananda Putri Masaling and Derwin Suhartono. 2024. Utilizing roberta and xlm-roberta pre-trained model for structured sentiment analysis. *International Journal of Informatics and Communication Technology*, 13(3):410–421.

João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. 2023. Advancing neural encoding of portuguese with transformer albertina pt-*. *Preprint*, arXiv:2305.06721.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2–3):164–210.

Chengjie Zhou, Bobo Li, Hao Fei, Fei Li, Chong Teng, and Donghong Ji. 2021. Revisiting structured sentiment analysis as latent dependency graph parsing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, Online. Association for Computational Linguistics.