Team PiLN at ABSAPT 2022: Lexical and BERT Strategies for Aspect-Based Sentiment Analysis in Portuguese

Francisco Assis Ricarte Neto[1], Rogério Figueredo de Sousa[1], Roney Lira de Sales Santos[2], Rafael Torres Anchiêta[1] and Raimundo Santos Moura[2]

[1]*Federal Institute of Piauí - Brazil*
[2]*Federal University of Piauí - Brazil*

Abstract

This paper describes the participation of the PiLN team in the IberLEF 2022 shared task on Aspect-Based Sentiment Analysis in Portuguese. The task aimed to develop methods for extracting aspect terms and identifying sentiment orientation. We approach these tasks in two different ways. For the aspect term extraction, we adopt a string-match strategy, using the aspects of an Ontology. For the sentiment orientation extraction, we fine-tuned a pre-trained BERT model of Portuguese. Also, we propose a weighting schema to deal with the dataset imbalance issues. Our simple approaches have achieved good results, reaching second place in both tasks.

Keywords

Sentiment Analysis, Aspect Extraction, Sentiment Orientation, Portuguese language

1. Introduction

According to Liu (2020) [1], Sentiment Analysis (SA) is the computational study of people's opinions, sentiments, emotions, and attitudes. Usually, opinions are expressed in comments (i.e., textual descriptions), through entities and their aspects or properties. SA is a recent area that includes research in Data Mining, Computational Linguistics, Information Retrieval, Artificial Intelligence, Natural Language Processing (NLP), among others.

The Scientific community has been investigated sentiment analysis at three levels, mainly [1]: i) **document level**, that the task is to classify whether a whole opinion document expresses a positive or negative sentiment; ii) **sentence level**, the task at this level goes to the sentences and determines whether each sentence expresses a positive, negative, or neutral opinion; and iii) **entity or aspect level**, that directly looks at the opinion itself. It is based on the idea that an opinion consists of *sentiment* (positive or negative) and a *target* (of opinion).

For the Portuguese language, it is difficult to find approaches that deal with the entity or aspect level. So, the ABSAPT 2022 - Aspect-Based Sentiment Analysis in Portuguese, was proposed the

IberLEF 2022, September, A Coruña, Spain.
farn@ifpi.edu.br (F. A. R. Neto); rogerio.sousa@ifpi.edu.br (R. F. d. Sousa); roneysantos@ufpi.edu.br (R. L. d. S. Santos); rta@ifpi.edu.br (R. T. Anchiêta); rsm@ufpi.edu.br (R. S. Moura)
0000-0003-4589-6157 (R. F. d. Sousa); 0000-0001-9562-0605 (R. L. d. S. Santos); 0000-0003-4209-9013 (R. T. Anchiêta); 0000-0002-1558-3830 (R. S. Moura)

first shared task dedicated to identifying aspects and extracting the polarity in Portuguese texts. This task is part of IberLEF 2022 (see https://sites.google.com/inf.ufpel.edu.br/absapt2022/home).

In general, the opinion mining problem can be structured into three generic tasks: i) feature extraction; ii) definition of semantic orientation; and iii) summarizing results.

This paper presents our strategies to deal with Aspect-Based Sentiment Analysis in Portuguese in the IberLEF 2022.

2. Aspect-Based Sentiment Analysis tasks

Aspect-Based Sentiment Analysis in Portuguese was organized into two shared tasks: aspect term extraction and sentiment orientation extraction. The first task comprehends the identification of aspects in the reviews, and the later task proposes to extract the sentiment orientation (polarity) of the review about a single aspect mentioned in it.

For these tasks, the organizers released a corpus with reviews of travelers about accommodation services companies. The training set contains $3,111$ samples from 847 reviews for the two tasks. The test set comprises 257 samples for the identification aspect task and 686 for the sentiment orientation extraction task. In what follows, we describe our strategies to deal with these tasks.

3. Developed strategies

We developed a straightforward approach using provided aspects from the HOntology [2] base with n-grams from reviews to extract aspects. HOntology is a freely available domain-specific ontology, that has five main top-level concepts: Accommodation, Facility, Room, Service/Staff, and Guest Type. It also includes other concepts: Design, Meal, Points of Interest, Price, Rating, and Staff. In general way, HOntology contains 282 concepts categorized into 16 top-level concepts, organized in a 5-level hierarchy, with all concepts and properties are defined in English, Portuguese, Spanish, and French. For the sentiment orientation extraction, we build a Bidirectional Encoder Representations from Transformers (BERT) [3] model using a sentence pair strategy to make the classification.

3.1. Aspect Term Extraction

This approach was conducted in two stages (see Figure 1). In the first one, we built a list of aspects terms from the provided set to the task[1], a subset of HOntology. We considered all terms from the set, including subaspects and subsubaspects.

The second stage starts with the pre-processing of reviews, removing punctuation and accents of the words. In sequence, with the size of the largest aspect from the list of aspects (6 terms), we transformed all reviews into 6-gram sequences using the Natural Language Toolkit (NLTK) [4]. Finally, we extracted the aspects from matches obtained by directly comparing the 6-grams sequences with aspects from the list. Due to the simplicity of our approach, we did not conduct

[1] Available at http://famahotel.tk/git/subsubfeatures.pdf

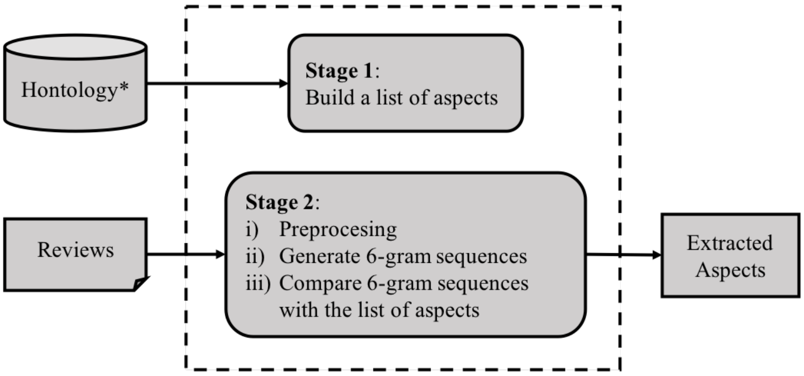

Figure 1: Aspect Term Extraction: Overview

any experiments on this task. Although the strategy is simple, it ranked $2nd$, achieving 0.65 on the test set when used the balanced accuracy metric.

3.2. Sentiment Orientation Extraction

We tackled this task using the approach proposed by Hoang et al. [5]. They take advantage of BERT's [3] ability to handle sentence pairs. The review is placed in the first sentence while the aspect is in the second sentence. The second sentence was formatted following a similar structure to the one proposed in SemEval-2016: "ENTITY#ASPECT" [6]. The authors formatted it to have a sentence-like structure, so, for example, the pair "FOOD#STYLE_OPTIONS" gets parsed into "food, style options". Figure 2 presents the general scheme of the adopted model for this task.

Similarly, we had to change to the suggested format for the second sentence since the task does not require identifying entities but only aspects. We remove the part referring to the entity and use only the aspect, and if it is multi-word, we separate it with underscore.

4. Experiments and Results

To conduct the experimentation process, we split the training set into 80% for training and 20% for testing. Furthermore, considering the distribution of labels in the dataset, we propose a weighting of the labels trying to address its imbalance. We weight each label depending on how frequently they appear in the training set. The higher the frequency of a label, the lower the weight of the given label. The best weights chosen are $[3, 0.49, 2.5]$ for Negative, Positive, and Neutral.

We used the Bertimbau [7, 8] as a pre-trained model for BERT. This model was fine-tuned and the pre-trained parameters were not frozen during fine-tuning. The reviews were tokenized using the default tokenizer of the Bertimbau model. We applied a single-layer feed-forward network to the classification output vector (CLS - 768 dimensions) to classify the instances. The main hyperparameters are presented in Table 1. These hyperparameters were empirically chosen.

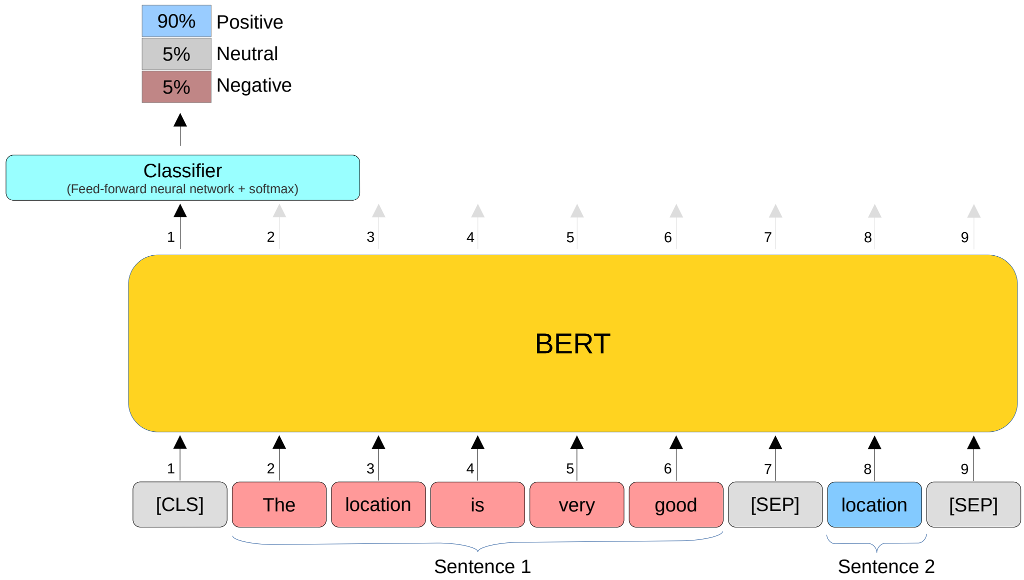

Figure 2: Sentiment Orientation Extraction: Overview.

Table 1
Main hyperparameters for fine-tuning the BERT model.

Hyperparameter	Value
Epoch	2
Learning rate	$5e-5$
Optimizer	AdamW [9]
Bach size	8
Max sequence length	384
Seed	2

Table 2 presents the results of our model in the training set and Table 3 shows the results in the testing set.

Table 2
Results of our approach in the training set.

Label	Precision	Recall	F1-score
NEG	0.69	0.83	0.75
NEU	0.65	0.63	0.64
POS	0.93	0.90	0.91

As in the aspect term extraction, our approach ranked $2nd$ in the sentiment orientation extraction task, achieving competitive results in the test set. Our methods and trained models are available at https://github.com/lplnufpi/absapt.

Table 3
Results in the testing set.

Approach	BACC	F1	Precision	Recall
Fine-tuned BERT	0.78	0.77	0.76	0.78

5. Final Remarks

This paper presented our strategies for IberLEF 2022 shared task on Aspect-Based Sentiment Analysis in Portuguese. For the aspect term extraction, we developed a method based on string-match, and for the sentiment orientation task, we fine-tuned a pre-trained BERT model of Portuguese. Moreover, we propose a weighting strategy to handle the dataset imbalance issues. We achieved promising results, with our approaches ranked $2nd$ in both tasks. For future work, we intend to investigate BERT-based methods for aspect extraction.

References

[1] B. Liu, Sentiment Analysis: Mining Opinions, Sentiments, and Emotions, 2nd edition ed., Cambridge University Press, 2020. ISBN 978-1108486378.

[2] M. Chaves, L. Freitas, R. Vieira, Hontology: A multilingual ontology for the accommodation sector in the tourism industry, in: Proceedings of the 4th International Conference on Knowledge Engineering and Ontology Development, Barcelona, Spain, 2012, pp. 149–154.

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186.

[4] S. Bird, E. Klein, E. Loper, Natural language processing with Python: analyzing text with the natural language toolkit, " O'Reilly Media, Inc.", 2009.

[5] M. Hoang, O. A. Bihorac, J. Rouces, Aspect-based sentiment analysis using BERT, in: Proceedings of the 22nd Nordic Conference on Computational Linguistics, Linköping University Electronic Press, Turku, Finland, 2019, pp. 187–196.

[6] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. AL-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq, V. Hoste, M. Apidianaki, X. Tannier, N. Loukachevitch, E. Kotelnikov, N. Bel, S. M. Jiménez-Zafra, G. Eryiğit, SemEval-2016 task 5: Aspect based sentiment analysis, in: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), Association for Computational Linguistics, San Diego, California, 2016, pp. 19–30.

[7] F. Souza, R. Nogueira, R. Lotufo, Portuguese named entity recognition using bert-crf, arXiv preprint arXiv:1909.10649 (2019).

[8] F. Souza, R. Nogueira, R. Lotufo, Bertimbau: Pretrained bert models for brazilian portuguese, in: R. Cerri, R. C. Prati (Eds.), Intelligent Systems, Springer International Publishing, Cham, 2020, pp. 403–417.

[9] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations, New Orleans, EUA, 2019.