TeamUFPR at ABSAPT 2022: Aspect Extraction with CRF and BERT

Tiago Heinrich[1,*,†], Felipe Marchi[2,†]

[1]Federal University of Paraná (UFPR), Curitiba, Brazil
[2]Santa Catarina State University (UDESC), Joinville, Brazil

Abstract

This paper describes the participation of the TeamUFPR at the Aspect-Based Sentiment Analysis in Portuguese (ABSAPT 2022), framed within the Iberian Languages Evaluation Forum (IberLEF 2022). Two tasks were proposed, aiming at evaluating aspect and sentiment. The first task consists of Aspect Term Extraction, where a set of aspects must be extracted from a group of texts. The second task focuses on Sentiment Orientation Extraction, where sentiment classification must be performed based on three classes. We considered the experiments carried out in the previous event to improve the strategies used for sentiment classification, and we explored an algorithm already discussed in the literature for aspect extraction. Our proposal focused on using the Conditional Random Fields (CRF) algorithm for Aspect identification and Bidirectional Encoder Representations from Transformers (BERT) for sentiment extraction.

Keywords

Sentimental Analysis, Natural Language Processing, Aspect Extraction, Machine Learning

1. Introduction

Strategies aimed at Natural Language Processing (NLP) have become popular in the last decade. These strategies aim to represent and extract features from text and sentences. This behavior is similar to the process of human understanding when reading a text, and It brings advantages when considering the volume of information on the Internet that cannot be evaluated on time by human beings [1].

The study of Aspect-Based Sentiment Analysis (ABSA) is a set of techniques aimed at identifying and extracting aspects of sentences. The aspect allows extracting information from sentences that can be used to better understand the sentiment in sentences. Aspects are also used to define the features and categories of the evaluated sentences.

ABSAPT 2022 is the first IberLEF task focused on aspect-based sentiment analysis in Portuguese [2]. The competition uses a dataset extracted from TripAdvisor. This dataset consists of

IberLEF 2022, September 2022, A Coruña, Spain.
*Corresponding author.
†These authors contributed equally.
theinrich@inf.ufpr.br (T. Heinrich); felipe.r.marchi@gmail.com (F. Marchi)
https://github.com/h31nr1ch (T. Heinrich); https://github.com/Markhyz (F. Marchi)
0000-0002-8017-1293 (T. Heinrich); 0000-0002-7711-3498 (F. Marchi)

user reviews. ABSAPT is one of the tasks offered at IberLEF 2022, in the sentiment, stance, and opinions section.

Our proposal for the competition consists of using two strategies, one aimed at extracting aspects and the other aimed at classifying and extracting sentiment. The choice of this strategy takes into account the team's learning when considering the previous tasks [3, 4].

Our team's experience is varied, working with machine learning problems for security and artificial intelligence for protein structure problems. Although we do not work directly with NLP problems, we are interested in learning strategies applied in this environment and if it is possible to use some prior knowledge in our solutions.

The remainder of this paper is structured as follows: Section 2 describes the ABSAPT 2022 task. Section 3 presents the related work. Section 4 explains our solution, discusses our findings, and Section 5 concludes the paper.

2. Task description

This year two tasks were proposed to be explored by the competitors. Both tasks focus on using a new proposed dataset, which addresses the problem of missing data in Portuguese for problems related to aspect extraction [2].

The dataset used by the competitors was extracted from TripAdvisor reviews written in Portuguese. Table 1 presents the distribution of data for training the models and the two sets of tests used to evaluate the efficiency of the models proposed by each team (one test set for each task).

Table 1
ABSAPT 2022 Dataset.

	Task	Samples
Train		3,111
Test	1	257
	2	686

Task 1 is aimed at aspect term extraction. Thus, each team must propose and test a strategy to extract one or more aspects of a sentence. Each review in the TripAdvisor dataset can generate a list of aspects that will be taken into account in the model evaluation process.

Task 2 consists of sentiment orientation extraction. Where each team should propose a strategy to extract polarity from TripAdvisor reviews. Three polarity classes are possible these being positive, negative, or neutral. For each review, polarity must be found in the models.

3. Related Works

[1] presents a study focused on the extraction of terms and aspects. The work focuses on using Conditional Random Fields (CRF) to extract aspects and achieves interesting results. The findings of the work would be the improvement of the performance of the algorithms through

the use of bigrams and the advantage of using POS tags. Finally, the work also presents a comparative study with strategies found in the literature.

Considering previous tasks like [4, 3]. Some learning's can be taken into account for new tasks. Despite classic machine learning strategies such as Random Forest (RF) and Multilayer Perceptron (MLP), they present satisfactory results for identifying sentiment in texts. Strategies using the Bidirectional Encoder Representations from Transformer (BERT) achieve superior results and flexibility when considering pre-trained models.

4. Applied approach

In this section, we describe the approaches used for each task. Due to the nature of the proposed tasks, two strategies were considered and applied. These strategies used the total set of data presented in Section 2, to train the algorithms.

4.1. Task 1: Aspect Term Extraction

The first task consists of extracting aspects of the dataset. As shown in Table 1, a single dataset was provided for training the algorithms. These data have a tag with the respective aspects found in each data sample (sentence).

Our approach for this task consisted in using Conditional Random Fields (CRF). The team chose this approach due to the popularity of the CRF in the literature [1]. CRF is a strategy applied to pattern recognition, allowing the prediction of labels by considering the context.

To apply the CRF, the dataset had to be adapted with a Tokenization and POS Tagging (part-of-speech tagging) technique. The POS Tagging step aims to add specific labels for each token found in the text. Due to the limited number of samples and the few options for samples of POS taggers in Portuguese, a model already trained was used for the POS Tag process[1]. It is important to note that this process was added to complete the aspect tag present in the training dataset. This process solved the overfitting problem observed in previous tests.

We do not apply any Lemmatization strategy. The only complementary pre-processing performed was Stemming and an adaptation of the tokens to consider the context. Figure 1 describes the pipeline architecture.

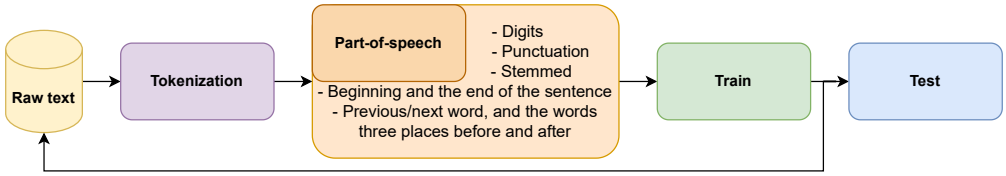

Figure 1: CRF Pipeline Architecture.

As a result of the training, we observed that the trained model could take advantage of POS Tagging for most of the classes added in this pre-processing step. However, the aspect

[1]https://github.com/inoueMashuu/POS-tagger-portuguese-nltk

identification was hampered by some Tags already defined in the pre-trained model used in the POS Tag stage. Due to the team's time constraints this problem was not solved, and harmed the model's performance in the competition.

Despite the result obtained, CRF has the potential to be a suitable approach for aspect extraction. The algorithm ends up hampered by the limitations of Portuguese models that would be used in the POS Tag stage, but a better adaptation of this stage will contribute to a performance gain for the CRF.

4.2. Task 2: Sentiment Orientation Extraction

Our team considered the strategies explored in [3] for the sentiment orientation extraction task. In this way, we approach the problem with a focus on using Bidirectional Encoder Representations from Transformer (BERT) for sentiment extraction.

We performed tests with a generic BERT model, but due to the dataset size we decided to explore a pre-trained[2] model aimed at the language addressed in the task. BERTimbau was the model chosen due to the variety in the training set and options already offered by pre-trained models in Portuguese [5, 6].

The result achieved by our model ends up reflecting some of the decisions made by the team. The tie with three other teams in the third position was interesting when considering all the metrics used in the model evaluation process.

The model proposed by our team ended up hampered by the training time. Due to the time limitation of the team, we limited the number of epochs to train the model. In a more detailed evaluation of the model, we observed that the proposed model could have taken more advantage if a greater number of epochs had been explored. The threshold used to define the learning between each epoch was very high, causing the training to stop at an early stage.

Despite this problem, BERT stands out as a suitable algorithm for identifying and extracting feelings. The algorithm can also take advantage of a wide range of options for problems related to the Portuguese language, which contributes to the development and application of BERT to solve problems related to natural language processing in Portuguese.

Table 2 shows the parameters employed by the models submitted for each task. The best results (f1-score) found during training are also displayed.

Table 2
Parameters used in the algorithms.

CRF	**BERT**
algorithm: arow	batch size: 32
max iterations: 300	epoch: 10
transitions: True	learning rate: 5e-5
	dropout: False
Best Train result	
F1-score: 96.40%	F1-score: 92.21%

[2] https://github.com/neuralmind-ai/portuguese-bert

5. Conclusion

In this paper, we describe the participation of the TeamUFPR at the ABSAPT 2022 Task on aspect-based sentiment analysis in Portuguese. Task 1 consists of developing a model to extract aspects and task 2 consists of extracting sentiment. The tasks must be developed using a training base extracted from TripAdvisor, and two different test bases. We proposed to use a strategy for each task.

For aspect extraction, we used Conditional Random Fields (CRF) which is an algorithm aimed at pattern identification, with the advantage of extracting aspects since it considers the context of the sentence. In the sentiment extraction task, we chose to use Bidirectional Encoder Representations from Transformer (BERT), an algorithm proposed by Google that allows the extraction of sentiments.

At the end of the competition, we observed that our approach towards extracting aspects ended up being limited, due to the limitation of the models used by the team. In the sentiment extraction task, we identified that our result could have been improved if a larger set of epochs had been used in the training stage.

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 Computer Science department.

References

[1] A. D. Francisco, Aspect Term Extraction in Aspect-Based Sentiment Analysis, B.S. thesis, Brasil, 2019.

[2] 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 69 (2022).

[3] U. B. Corrêa, L. Coelho, L. Santos, L. A. d. Freitas, Overview of the idpt task on irony detection in portuguese at iberlef 2021 (2021).

[4] 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: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021). CEUR-WS.org, volume 2943, CEUR, 2021, pp. 925–932.

[5] F. Souza, R. Nogueira, R. Lotufo, Portuguese named entity recognition using bert-crf, arXiv preprint arXiv:1909.10649 (2019). URL: http://arxiv.org/abs/1909.10649.

[6] F. Souza, R. Nogueira, R. Lotufo, 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), 2020.

A. Online Resources

The source code is available in:

- GitHub https://github.com/h31nr1ch/TeamUFPR-ABSAPT2022.