# A Deep Learning Approach for Aspect Sentiment Triplet Extraction in Portuguese

José Meléndez Barros(✉) and Glauber De Bona

Universidade de São Paulo, São Paulo, SP 05508-010, Brazil
{jose.melendez,glauber.bona}@usp.br

**Abstract.** Aspect Sentiment Triplet Extraction (ASTE) is an Aspect-Based Sentiment Analysis subtask (ABSA). It aims to extract aspect-opinion pairs from a sentence and identify the sentiment polarity associated with them. For instance, given the sentence "Large rooms and great breakfast", ASTE outputs the triplet T = {(rooms, large, positive), (breakfast, great, positive)}. Although several approaches to ASBA have recently been proposed, those for Portuguese have been mostly limited to extracting only aspects without addressing ASTE tasks. This work aims to develop a framework based on Deep Learning to perform the Aspect Sentiment Triplet Extraction task in Portuguese. The framework uses BERT as a context-awareness sentence encoder, multiple parallel non-linear layers to get aspect and opinion representations, and a Graph Attention layer along with a Biaffine scorer to determine the sentiment dependency between each aspect-opinion pair. The comparison results show that our proposed framework significantly outperforms the baselines in Portuguese and is competitive with its counterparts in English.

**Keywords:** Deep learning · Natural Language Processing · Aspect Sentiment Triplet Extraction

## 1 Introduction

Sentiment analysis, also known as opinion mining, is a field of Natural Language Processing (NLP) that analyzes people's opinions, sentiments, and attitudes towards entities such as products, services, organizations, individuals, among others [21]. There are three granularity levels in sentiment analysis: document level, sentence level and aspect level. Aspect-Based Sentiment Analysis (ABSA) is the term used to denote a set of tasks that aim to resolve aspect level granularity problems [21,25], i.e., given a review, the model should select specific words corresponding to *aspect terms* and *opinion terms* and identify their *sentiment polarity*. In turn, ABSA subtasks may focus only on extracting/classifying one or more of these elements. In this work, we deal with Aspect Sentiment Triplet Extraction (ASTE), which is a relatively new subtask [24].


A. Britto and K. Valdivia Delgado (Eds.): BRACIS 2021, LNAI 13074, pp. 343–358, 2021.
https://doi.org/10.1007/978-3-030-91699-2_24

Consider for instance the sentence *"Large rooms and great breakfast"*. Aspects are word sequences describing attributes or features of the targets (e.g., rooms, breakfast), opinions are those expressions carrying subjective attitudes or (un)desirable characteristics (e.g., large, great) and sentiments are the polarities associated with aspect-opinion pair, which can be positive, negative or neutral. ASTE aims to extract not only the aspects and its sentiment, but also the corresponding opinion spans expressing the sentiment for each aspect, as the example in Fig. 1 illustrates.

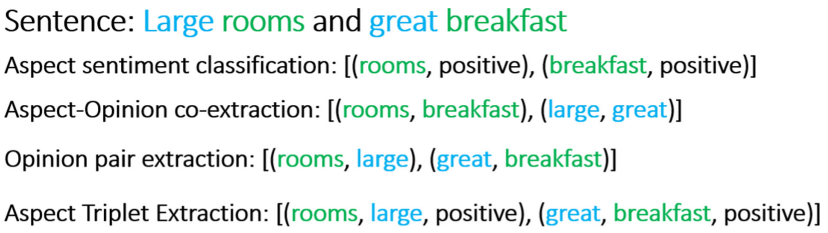


**Fig. 1.** Differences among ABSA subtasks.

ABSA tasks tackled in Portuguese include aspect term extraction [2,23], aspect sentiment classification [1,26] and aspect opinion co-extraction [4], but the proposed solutions employs methods that are considerably behind the state of the art in reference languages [25]. Besides, these works are focused on identifying aspects or opinions separately. Our work aims to extract aspects, opinions and sentiment polarity simultaneously. This not only provides richer context output, but also explains why a specific sentiment polarity was assigned.

We develop a framework based on deep learning to perform the Aspect Sentiment Triplet Extraction task in Portuguese. Our framework is based the state-of-the-art models for English, building on the work of Zhang et al. [32]. The main improvements proposed by our work are: we use BERT instead of Glove for sentence encoding, with word-vectors capturing a richer context; we add a Graph Attention layer to capture more complex relationships between two word vectors. The comparison results show that our proposed framework outperforms the baselines by more than 10%. Besides, the ablation study demonstrates the effectiveness of Graph Attention layers and BiLSTM layers for dimensionality reduction. The main contributions of this paper can be summarized as follows:

- We propose the first model to deal with ASTE tasks in Portuguese. To the best of our knowledge, all works in Portuguese-ABSA have been limited to Aspect Extraction.
- We provide a new dataset in Portuguese to work with ASTE tasks. Additionally, we adapt an existing dataset for this task (ReLi Corpus [12]). Other datasets [11,13,28] related to ABSA in Portuguese do not have the proper labeling for ASTE tasks.

## 2 Related Work

ABSA includes several subtasks that vary according to the domain of the target term. Xu et al. [31] developed a model based on BiLSTM networks, Attention and Conditional Random Fields (CRF), aiming to extract an opinion term given an aspect term as input. Their main contribution is the model's capability of capturing variable-length opinion spans. This model cannot manage overlapped aspects/opinions. He et al. [17] work on aspect-sentiment pair extraction task. They created a model using CNN layers and an attention mechanism. It incorporates two document-level classification tasks to be jointly trained with Aspect Extraction and aspect-level sentiment classification, allowing aspect-level tasks to benefit from document-level information. This model requires outputs intensively annotated to validate the results.

Aspect/target term sentiment classification is another subtask commonly addressed in ABSA, approached for instance by Cui and Maojie [5] and Han et al. [16]. Both works employed BiGRU/LSTM and attention mechanisms to attenuate irrelevant information, but neither of them can manage overlapped sentiments. Meanwhile, Fan et al. [10] focused their work on target-oriented opinion words-extraction (TOWE), which aims to extract aspect-opinion tuples given an aspect(s) as input. To achieve that, they proposed an Inward-Outward LSTM to get information from the left and the right contexts of the target. This work can handle overlapped aspect/opinion.

Aspect Sentiment Triplet Extraction was recently proposed by Peng et al. [24], and they put forward a two-stage framework to extract opinion triplets. In the first stage, they initially used a neural network based on BiLSTM and Graph Convolutional Networks to extract aspects-sentiments pairs and opinion terms separately. Then, to detect the relationship between aspect terms and opinion terms in the second stage, a LSTM network and pretrained word embeddings are employed. This pipeline approach might suffer from error propagation.

To handle error propagation problems in multi-stage frameworks, Wu et al. [30] proposed a tagging scheme model, which is implemented in three variants, using CNN, BiLSTM or BERT. Advantages of this approach include the following: only raw reviews are required as input; this model can extract all opinion factors of opinion pair extraction in one step, instead of pipelines; and it is easily extended to other pair/triplet extraction tasks from text. The most important limitation is again that outputs must be intensively annotated for the results to be validated.

Zhang et al. [32] developed a framework that can handle overlapping sentences, only requires raw review as input, and that is not multi-stage like the above one. The model structure consists of a word embedding (Glove) attached to a BiLSTM layer, which they use as a context sentence encoder. The BiLSTM output $h$ is passed to ReLU layers, which apply dimension-reduction to strip away irrelevant features. Finally, two independent softmax layers obtain distributions over the labels that denote an aspect-opinion term. A biaffine scorer [9] to determine if the sentiment dependency between each word pair is neutral, negative, positive, or null.

Unlike the works developed for English, in Portuguese, no available models were found in ASTE. The most remarkable works are focused on Aspect Extraction and Aspect-sentiment pair extraction. Cardoso and Pereira [4] used CRF and opLexicon to deal with aspect sentiment classification. Aires et al. [1] developed two models, one based on SVM and one employing an LSTM network. Saias et al. [26] opted for three methods: Maximum Entropy classifier, Sentiment lexicon SentiLex-PT and Rule-based methods, with the last two focusing on Aspect-sentiment pair extraction. Finally, Balage F. [2] used clustering techniques, word2vec embeddings and CRF to implement an unsupervised model.

## 3 Proposed Framework

Our framework[1] (BERT for Opinion Triplet Extraccion - BOTE) consists of 4 modules. The overall architecture is shown in Fig. 2. The sentence encoder module generates a set of word vectors, which encode semantic and context information. The aspect-opinion representation module extracts the aspect and opinion features from word vectors. Then, the aspect-opinion tagging module takes as input this feature vector to label the word as an aspect or an opinion or neither of them. Finally, the sentiment dependency between aspect and opinion vectors is detected by the dependency parsing module.

### 3.1 Problem Definition

Our Aspect Sentiment Triplet Extraction (ASTE) [24] problem is defined as, given a sentence $S = \{w_1, w_2, w_3, ..., w_n\}$ consisting of $n$ words, extracting all possible triplets $T = \{(a, o, p)_m\}_{m=1}^{|T|}$ from $S$, where $a$, $o$ and $p$ respectively denote an $n$-gram aspect term, an $n$-gram opinion term and a sentiment polarity; $a_m$ and $o_m$ can be represented as their start and end positions $(s_m, e_m)$ in $S$ and $p_m \in \{Positive, Negative, Neutral\}$. Note that this task involves dealing with One-to-One, One-to-Many and Many-to-Many relationships between aspects and opinions, as well as overlapped sentences.

### 3.2 Sentence Encoding

We adopt BERT [7] to encode sentences. Given a sentence $S = \{w_i\}_{i=1}^{|S|}$ we get a word-vectors set $V = \{v_i | v_i \in \mathbb{R}^{d_B}\}_{i=1}^{|S|}$ from the $n^{th}$ BERT hidden state, where $d_B$ denote the dimensionality of the BERT word-vector. The capability of BERT to generate different word embeddings depending on the context allows us to obtain feature vectors with better semantic information; for example, BERT vectors can deal with homonyms. Additionally, BERT can handle long reviews better than LSTM networks, thanks to attention mechanisms.

The BERT model works with WordPiece [7], therefore, from the sentence $S$, we obtain token vectors $t_i$ instead of word vectors $v_i$. A word can be formed by

[1] Source code available at https://github.com/josemelendezb/bote.

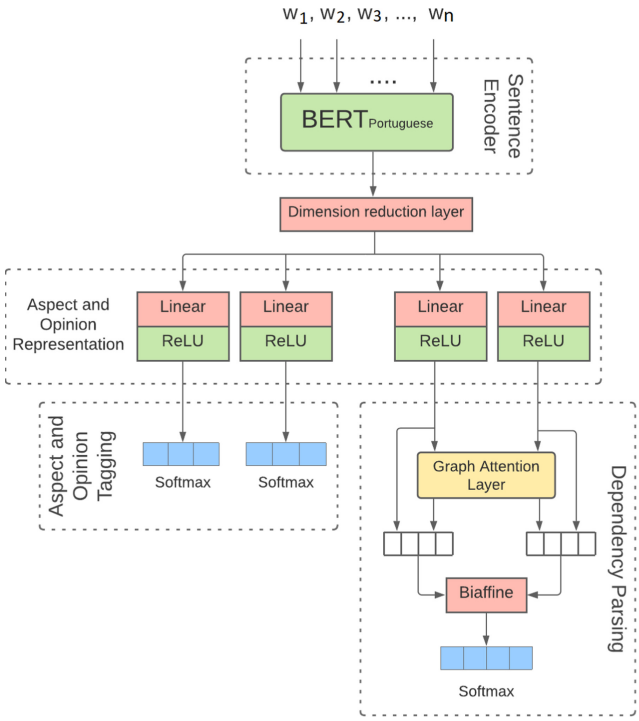


**Fig. 2.** Architectures of our model.

two or more tokens. To obtain a vector $v_i$ of a word $w_i$, we average the token vectors $(u_{i1}, u_{i2}, ..., u_{il})$ that form that word; formally, $v_i = \frac{1}{m} \sum_{l=1}^{m} u_{i,l}$ for all $v_i \in V$. A custom Average Pooling2D layer calculates the averages.

Finally, we apply a BiLSTM dimension-reducing layer to the word vectors. Since a BERT vector can have 758 or 1024 dimensions and machine learning does not understand causality, models map any input feature to the target variable, even if there is no causal relation. Many features demand a more complex model and increase the risk of overfitting. Dimensionality reduction removes multicollinearity and irrelevant features, making our model more straightforward, less data-hungry, and reducing the risk of overfitting.

### 3.3 Aspect and Opinion Representation

Following Zhang et al. [32], we apply linear layers and nonlinear functions on $V$ before the opinion and aspect tagging operation in order to attenuate the noise of irrelevant information. Formally, let $\mathbf{r}_i^{(ap)}$ and $\mathbf{r}_i^{(op)}$ ($\mathbf{r}_i^{(ap)\prime}$ and $\mathbf{r}_i^{(op)\prime}$) denote the aspect and opinion representations employed in aspect and opinion tagging (sentiment dependency parsing), with $\mathbf{r}_i^{(ap)}, \mathbf{r}_i^{(op)}, \mathbf{r}_i^{(ap)\prime}, \mathbf{r}_i^{(op)\prime} \in \mathbb{R}^{d_r}$, where $d_r$ is the dimensionality of the representation. The linear layers employ a ReLU family

function $g(.)$ and have parameters $\mathbf{W}_r^{(ap)}, \mathbf{W}_r^{(op)}, \mathbf{W}_r^{(ap)\prime}, \mathbf{W}_r^{(op)\prime} \in \mathbb{R}^{d_r \times d_B}$ and $\mathbf{b}_r^{(ap)}, \mathbf{b}_r^{(op)}, \mathbf{b}_r^{(ap)\prime}, \mathbf{b}_r^{(op)\prime} \in \mathbb{R}^{d_r}$:

$$\mathbf{r}_i^{(ap)} = g(\mathbf{W}_r^{(ap)} v_i + \mathbf{b}_r^{(ap)}) \tag{1}$$

$$\mathbf{r}_i^{(op)} = g(\mathbf{W}_r^{(op)} v_i + \mathbf{b}_r^{(op)}) \tag{2}$$

$$\mathbf{r}_i^{(ap)\prime} = g(\mathbf{W}_r^{(ap)\prime} v_i + \mathbf{b}_r^{(ap)\prime}) \tag{3}$$

$$\mathbf{r}_i^{(op)\prime} = g(\mathbf{W}_r^{(op)\prime} v_i + \mathbf{b}_r^{(op)\prime}) \tag{4}$$

### 3.4 Aspect and Opinion Tagging

Words are tagged by using two taggers built on a softmax layer as follows:

$$\mathbf{p}_i^{(l)} = softmax(\mathbf{W}_t^{(l)} \mathbf{r}_i^{(l)} + \mathbf{b}_t^{(l)}) \tag{5}$$

Above, $l \in \{ap, op\}$, $\mathbf{W}_t^{(l)} \in \mathbb{R}^{3 \times d_r}$ and $\mathbf{b}_t^{(l)} \in \mathbb{R}^3$ are trainable parameters. The softmax outputs two series of distributions over {B, O, I}, tagging aspect and opinion terms via the probability of each word $w_i$ being the Beginning/Inside/Outside of an aspect/opinion term. For example, in the sentence *"Large rooms and great breakfast. Room service was awesome."*, after decoding $\mathbf{p}_i^{(ap)}$ and $\mathbf{p}_i^{(op)}$ outputs, the result should be as in Table 1.

**Table 1.** Tagging result.

| | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
|---|---|---|---|---|---|---|---|---|---|---|
| | Large | rooms | and | great | breakfast | . | Room | service | was | awesome |
| $\mathbf{p}_i^{(ap)}$ | O | B | O | O | B | O | B | I | O | O |
| $\mathbf{p}_i^{(op)}$ | B | O | O | B | O | O | O | O | O | B |

### 3.5 Sentiment Dependency Parsing

In this module, the aim is to identify the sentiment polarity of every word pair $(w_i, w_j)$. Our model uses four tags (NEU, NEG, POS, NO-DEP) to denote these dependencies. The tags respectively indicate neutral, negative, positive, and non-existent sentiment dependency. There are $|S|^2$ possible word pairs in each sentence since, during the training process, the target triplets induce ordered pairs. To avoid redundant relations, following [3,32], sentiment dependency between an aspect and opinion term is assigned via the pair formed by their last word.

Formally, we have a 3D-tensor $\mathbf{T} \in \mathbb{R}^{|S| \times |S| \times 4}$, where each element $t_{ijk}$ denotes the probability that the polarity in $(w_i, w_j)$ is $k$, with $k \in$ {NEU, NEG, POS, NO-DEP}. Finally, we get an asymmetric square matrix $\mathbf{D} \in \mathbb{R}^{|S| \times |S|}$ as shown in Fig. 3.

| | Large | rooms | and | great | breakfast |
|---|---|---|---|---|---|
| Large | NO-DEP | NO-DEP | NO-DEP | NO-DEP | NO-DEP |
| rooms | POS | NO-DEP | NO-DEP | NO-DEP | NO-DEP |
| and | NO-DEP | NO-DEP | NO-DEP | NO-DEP | NO-DEP |
| great | NO-DEP | NO-DEP | NO-DEP | NO-DEP | NO-DEP |
| breakfast | NO-DEP | NO-DEP | NO-DEP | POS | NO-DEP |

**Fig. 3.** A parsing example for sentiment dependency.

A Biaffine scorer [9] is used to obtain word-level sentiment dependencies. This mechanism is widely used and has shown success in various parsing-related tasks [19,20]. The score $s_{ijk}$ for a pair $(w_i, w_j)$ that have a dependency $k$ is computed using syntax-aware vectors $\mathbf{h}_i^{(ap)\prime}$ and $\mathbf{h}_j^{(op)\prime}$ obtained from a Graph Attention Layer (GAL). Using **;** to represent the concatenation operation and $\mathbf{W}^{(k)}$ and $\mathbf{b}^{(k)}$ to denote trainable weight and bias, the scores can be defined as:

$$s_{ijk} = [\mathbf{W}^{(k)}\mathbf{z}_i^{(ap)\prime} + \mathbf{b}^{(k)}]^\top \mathbf{z}_j^{(op)\prime} \tag{6}$$

$$\mathbf{z}_i^{(ap)\prime} = [\mathbf{r}_i^{(ap)\prime}; \mathbf{h}_i^{(ap)\prime}], \quad \mathbf{z_i} \in \mathbb{R}^{2d_r} \tag{7}$$

$$\mathbf{z}_j^{(op)\prime} = [\mathbf{r}_j^{(op)\prime}; \mathbf{h}_j^{(op)\prime}], \quad \mathbf{z_j} \in \mathbb{R}^{2d_r} \tag{8}$$

An example of how to calculate $s_{ijk}$ is shown in Fig. 4. The aspect-vector $\mathbf{z}_i$ representing the word $w_i$ is inputted into the linear layer $\mathbf{W}^{(k)}(.) + \mathbf{b}^{(k)}$ (Eq. 6). The purpose of this linear layer is to generate a vector $\mathbf{z}_i$ for each $k$-polarity. Therefore, by multiplying $\mathbf{z}_{ik}$ and $\mathbf{z}_j$, we obtain a score that measures the polarity between $w_i$ and $w_j$. In our example, the score vector of $(\mathbf{w}_1, \mathbf{w}_0) =$ ("rooms", "large") is $s_{10} = \langle s_{100}, s_{101}, s_{102}, s_{103} \rangle = \langle -0.32, -0.39, -0.55, 1.98 \rangle$. Once we apply the Softmax function, we obtain the probability vector $s_{13} = \langle 0.08, 0.07, 0.06, 0.79 \rangle$, indicating that the dependency between "quartos" and "rooms" is positive.

Although BERT contains implicit syntactic information, its ability to capture explicit syntactic features is limited [15]. A common problem in ABSA when feature vectors are generated from a sequence, without considering explicitly the knowledge about the language, is that they may incorrectly locate specific targets in sentences with multiple aspects and opinions [22].

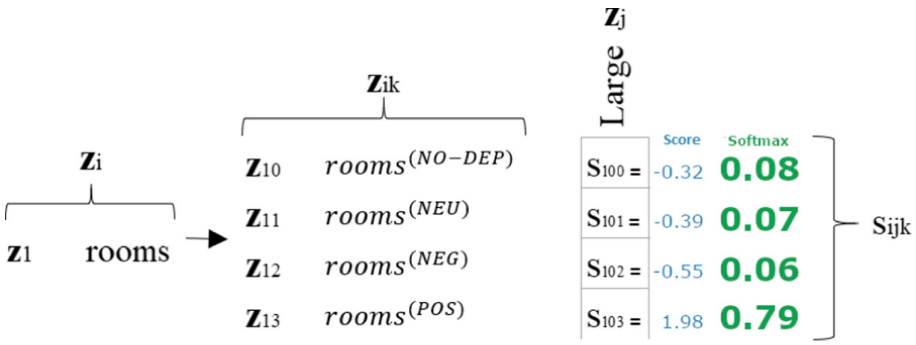


**Fig. 4.** A biaffine scoring example for sentiment dependency.

### 3.6 Syntax-Aware Vector

Unlike Zhang et al., we concatenate a syntax-aware vector $\vec{\mathbf{h}}$ with $\vec{\mathbf{r}}$ before inputting the features vector into the Biaffine scorer. As mentioned earlier, this vector $\vec{\mathbf{h}}$ is obtained from a Graph Attention Layer (GAL), which can to capture linguistic structures like dependency trees other than only sequential data [18]. The GAL receives the vector $\vec{\mathbf{r}}$ and an adjacency matrix that represents the syntactic dependency graph of the sentence[2]. GAL then embed the graph structure into the vector by performing masked attention – it computes self-attention between node $i$ and node $j$ if and only if $j$ is a neighbor of $i$.

Figure 5 shows the syntactic dependency of "The hotel has an awesome room service" and its adjacency matrix. In $\vec{\mathbf{r}}_6$, for example, self-attention is applied between $\vec{\mathbf{r}}_6$ and its neighbors (green region) instead of all tokens. In a dependency tree, the aspect-opinion pair will generally be related. By applying the mechanism of attention following the connections of the tree, more important nodes receive higher weight during neighborhood aggregation. Therefore, there will be more chances of identifying a sentiment dependency relationship in the selected pair. Although there are some useless dependency relationships for ASTE, aspect and opinion tagging outputs serve to counteract the noise produced by these irrelevant dependencies.

We take the syntactic dependency as an undirected graph. In directed graphs, self-attention is computed following the direction of the edge, however, edges do not always connect aspects and opinions directly, and if it does, the direction is not necessarily aspect → opinion. If we use a directed graph, we could lose information.

[2] We obtain the syntactic dependency graph of a sentence using Spacy parser.

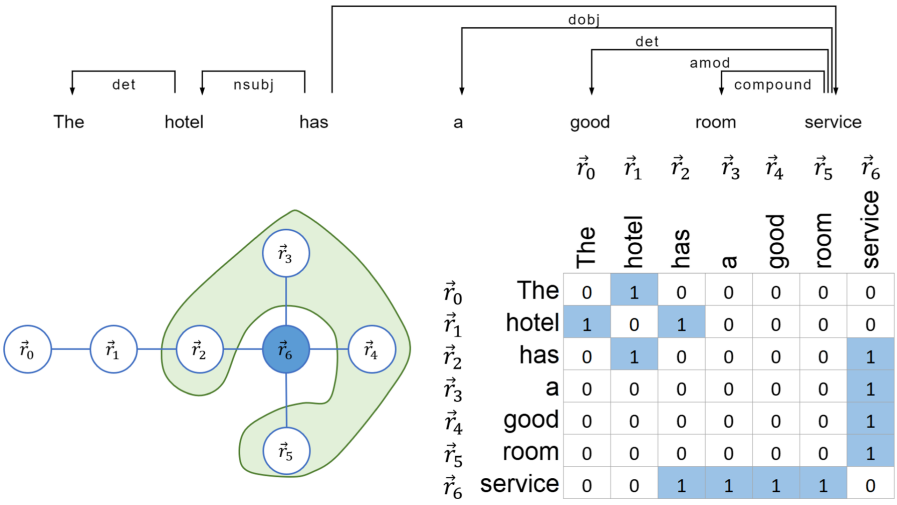


**Fig. 5.** A biaffine scoring example for sentiment dependency.

## 4 Experiments

### 4.1 Datasets

We conduct experiments in Portuguese and English datasets. Portuguese datasets include ReLi Corpus [12], which is composed of 1600 reviews from 14 different books with about 200 comments each. Besides, a new dataset was produced by us from hotel reviews collected from TripAdvisor (Table 2). Both datasets were manually annotated. These datasets are referenced in this work as ReLi and ReHol and they are available at https://dx.doi.org/10.21227/0ej1-br13.

Since there are no Portuguese models for ASTE tasks, English baselines were adapted to Portuguese-ASTE to compare with our model. Additionally, our model was trained in English to compare with the actual performance of these baselines. We used four English datasets[3] in the Laptop and Restaurant domain from Semeval 2014, 2015 and 2016 (Task 4, 12 and 5 respectively).

**Table 2.** Information regarding the composition of the ReHol dataset.

| | | | |
|---|---|---|---|
| # Reviews | 1514 | # Non-overlapped sentences | 941 |
| # Triplets | 3198 | # Overlapped sentences | 573 |
| # Positive triplets | 1845 | Average words per sentence | 16 |
| # Negative triplets | 1242 | Minimum sentence size (words) | 8 |
| # Neutral triplets | 111 | Maximum sentence size (words) | 38 |

[3] Datasets are available by Zhang et al. at https://github.com/GeneZC/OTE-MTL.

### 4.2 Baselines

Existing Portuguese-ABSA solutions do not provide enough information to reproduce their models/experiments. The lack of information includes: the source code is not available, there is insufficient information on the hyperparameters used, or the datasets are not available. The baselines were adapted to Portuguese by replacing the English pre-trained word embedding with a Portuguese pre-trained word embedding[4]. No further changes were made.

- **OTE-MTL** [32] is a model that encodes the sentences using Glove and BiLSTM networks, then, it applies dimension-reducing linear layers and non-linear functions on the hidden states to identify aspects and opinion and finally, the biaffine scorer [9] is used to get aspect-opinion pairs along with the associated sentiment with each other.
- **CMLA-MTL** [29] is an aspect-opinion co-extraction system, which is based on multi-layer attentions. It was adapted by Zhang et al. for ASTE task.

### 4.3 Experiment Settings

For the BERT encoder, we use the pre-trained cased BERT developed by NeuralMind [27]. To compare our model with the baselines in English, we use uncased BERT base model developed by Google Research[5]. The maximum sequence length is 512 with a batch size of 32. We train our model for a maximal of 60 epochs using Adam optimizer. The learning rate is set to 1e−5. BiLSTM dimension reducing vector is 300-dimensional. Aspect and opinion representation vectors and syntax-aware vectors are 150-dimensional with a single hidden layer. Dropout is applied to avoid overfitting and the drop rate is 0.3. The development set is used for early stopping.

### 4.4 Evaluation

Following the previous works [24,30,32], F1-Score, recall, and precision measurements were used to evaluate the results and compare our model to the baselines. A triplet is correct if and only if the aspect span, the opinion span, the parsing and the sentiment polarity are all correct. The $5 \times 2$-nested cross-validation technique [8] was used to generate sufficient data from the selected metrics. Lilliefors, Anderson – Darling, and Shapiro – Wilk tests were used to validate the assumption of normality, and the Mauchly's test for the assumption of Sphericity [6]. Because cross-validation process generates repeated measures, appropriate tests should be used to deal with within-subjects. Based on the evaluation of the assumptions, non-parametric tests were selected. Comparing our model with the baselines, we used Friedman's test to identify if there was a statistically significant difference between the models [6]. Finally, we applied Hommel post-hoc procedure to identify if our model was better than the baselines [14]. The significance level used in all tests was 0.05.

[4] http://nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc.
[5] https://github.com/google-research/bert.

## 5 Results and Analysis

### 5.1 Comparison with Baselines

The results are shown in Table 3. Our framework consistently achieves the best scores, both for the aspect and opinion extraction task and triplet extraction task in Portuguese. The average difference between our model and the Portuguese baselines in aspect extraction is 6.68%, opinion extraction is 7.87% and triplet extraction is 12.58%. In aspect extraction in English, the results are not conclusive. However, in opinion and triplet extraction in English, the framework outperforms the baselines significantly in three of the four datasets. In Rest 14, there is no statistically significant difference between BOTE and OTE-MTL. Detailed information about precision, recall, and how the errors affect the application can be found at https://dx.doi.org/10.21227/0ej1-br13.

**Table 3.** Results F1-score (%). * indicates difference is not statistically significant between flagged models. Bold indicates statistically significant superior performance.

| Task | Model | ReLi | ReHol | Rest14 | Rest15 | Rest16 | Lap14 |
|---|---|---|---|---|---|---|---|
| Aspect Ext. (AT) | OTE-MTL | 71.48 | 78.70 | 76.11* | 77.95* | 72.95 | 71.99 |
| | CMLA-MTL | 72.14 | 74.31 | 72.80 | 73.55 | 71.22 | 68.72 |
| | BOTE (ours) | **79.49** | **82.20** | 76.07* | 77.61* | **77.77** | 72.29* |
| Opinion Ext. (OT) | OTE-MTL | 68.61 | 72.93 | 77.91* | 74.27 | 76.19 | 65.62 |
| | CMLA-MTL | 62.16 | 69.45 | 74.46 | 74.72 | 74.69 | 63.82 |
| | BOTE (ours) | **70.52** | **81.10** | 77.98* | **77.69** | **78.88** | **68.66** |
| Triplet Ext. (TR) | OTE-MTL | 49.92 | 60.91 | 57.26* | 52.46 | 53.01 | 39.67 |
| | CMLA-MTL | 40.61 | 47.03 | 45.57 | 39.70 | 42.45 | 32.05 |
| | BOTE (ours) | **57.49** | **66.65** | 56.49* | **55.12** | **55.79** | **44.04** |

Additionally, experiments were carried out to test the models in domains for which they were not trained. The tests were performed in both Portuguese and English. The models in Portuguese were trained in the Books domain and tested in the Hotels domain. The English models were trained in the restaurant domain and tested in the laptop domain. The results are displayed in Table 4. BOTE shows significantly better performance than baselines in all three tasks, although lower when compared with BOTE results in the same domain. This indicates that, comparing with the baselines, our model responds better to new information, relationships and objects that were shown to it during training.

Finally, several tests were run to evaluate the performance of the proposed model based on the number of triples per sentence (Table 5). The results indicate that the greater the number of triplets in a sentence, the higher the hit rate (except between 3-triplets and 4-triplets in ReHol). Since the greater the number of triplets, the more complex are the relationships between aspects and opinion

**Table 4.** F1-score (%) transfer learning between different domains.

| Model | Train: rest14, Test: lap14 | | | Train: reli, Test: rehol | | |
|---|---|---|---|---|---|---|
| | AT | OT | TR | AT | OT | TR |
| OTE-MTL | 37.41 | 55.67 | 23.56 | 34.82 | 48.57 | 17.15 |
| CMLA-MTL | 33.38 | 54.44 | 15.27 | 20.18 | 43.93 | 4.37 |
| BOTE | **67.24** | **66.10** | **33.91** | **54.26** | **64.99** | **37.78** |

**Table 5.** BOTE F1-score (%) according to number of triplets per sentence.

| | 1-triplet | 2-triplets | 3-triplet | 4-triplets |
|---|---|---|---|---|
| Rest14 | 36.36 | 50.60 | 56.14 | 65.82 |
| Rest16 | 41.86 | 50.60 | 60.41 | 68.18 |
| ReHol | 33.33 | 56.34 | 75.62 | 68.03 |

terms, further experiments are needed to identify the underlying cause of this behavior.

### 5.2 Ablation Study

We examine the effectiveness of two components of our BOTE model, namely the reduction layer and graph attention layer (GAL). Table 6 presents the ablation results on Reli and ReHol datasets. BOTE-NoReduction does not use any type of reduction, BOTE-Linear uses a Linear layer, BOTE-ReLu uses a ReLU layer and BOTE-NoGraph uses a BiLSTM layer but does not use graph attention. The original BOTE uses a BiLSTM layer to reduce the size of the word vectors and graph attention layer to parse aspect and opinion terms. Regarding Aspect Extraction (AT) and Opinion Extraction (OT) tasks, there is no statistically significant difference between linear, nonlinear and no reduction. However, we can observe a significant improvement when the model uses a BiLSTM layer to reduce the dimensionality of the word vectors. Since GAL is not involved in AT and OT, it is not expected to be any significant difference between BOTE-NoGraph and BOTE. Regarding Triplet Extraction (TR), nonlinear and BiLSTM reduction show a significant improvement in the sentiment parsing process compared to no reduction. Likewise, when we use GAL we obtain better results compared to the other variants.

**Table 6.** Ablation study F1-score (%).

| Model | Reli | | | ReHol | | |
|---|---|---|---|---|---|---|
| | AT | OT | TR | AT | OT | TR |
| BOTE-NoReduction | 75.64 | 63.56 | 46.36 | 78.69 | 74.32 | 56.67 |
| BOTE-Linear | 74.83 | 62.97 | 47.84 | 78.24 | 74.13 | 58.22 |
| BOTE-ReLu | 74.14 | 63.17 | 50.15 | 78.14 | 73.41 | 57.66 |
| BOTE-NoGraph | 78.42 | 69.50 | 55.31 | 81.06 | 80.67 | 63.21 |
| BOTE | **79.49** | **70.52** | **57.49** | **82.20** | **81.10** | **66.65** |

## 6 Conclusions and Future Work

In this work, we proposed a Machine Learning framework to deal with Aspect Sentiment Triplet Extraction in Portuguese. Our model used a BERT model to obtain context-aware word vectors, a Graph Attention Layer to exploit the syntactic information contained in the sentence, and a Biaffine scorer as a sentiment parser. The model achieved an average F1-score of 62.07% in the datasets in ASTE Portuguese. When the model was trained in one domain and tested in another, it achieved an F1-score of 37.78% compared to 17.15% of the best baseline performance. Besides, the results showed the higher the number of triplets, the more accurate the model. The ablation study proved that performance is better when a dimensionality reducer is applied to BERT word vectors.

The experimental results verify the effectiveness of our framework compared to the baselines. To the best of our knowledge, we have developed the most fine-grained model to deal with aspect sentiment tasks in Portuguese. The proposed model even outperforms its English counterparts in some cases. Ablation study validates the effectiveness of applying dimensionality reduction in BERT word vectors, especially when using a BiLSTM layer. Besides, syntactic information improves sentiment parsing. Finally, our model shows a greater capacity of transfer learning across different domains, responding better to new information.

Future works include developing an opinion dependency tree where the relationship between words is based not on syntax, but on sentiment, as syntax trees usually ignore many connections between aspects and opinion words. Other possible research directions could aim at improving the classification ability of the model in multiple domains or determining the reason why the higher the number of triplets per sentence, the better is the model performance.

**Acknowledgments.** This work was supported by CAPES and Ministerio de Ciencia from Colombia.

## References

1. Aires, J.P., Padilha, C., Quevedo, C., Meneguzzi, F.: A deep learning approach to classify aspect-level sentiment using small datasets. In: 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2018)

2. Balage Filho, P.P.: Aspect extraction in sentiment analysis for Portuguese language. Ph.D. thesis, Universidade de Sao Paulo (2017)
3. Bekoulis, G., Deleu, J., Demeester, T., Develder, C.: Joint entity recognition and relation extraction as a multi-head selection problem. Expert Syst. Appl. **114**, 34–45 (2018)
4. Cardoso, B., Pereira, D.: Evaluating an aspect extraction method for opinion mining in the Portuguese language. In: Anais do VIII Symposium on Knowledge Discovery, Mining and Learning, pp. 137–144. SBC (2020)
5. Cui, Z., Maojie, Z.: Aspect level sentiment classification based on double attention mechanism. In: Proceedings of the 2019 2nd International Conference on E-Business, Information Management and Computer Science. EBIMCS 2019. Association for Computing Machinery, New York (2019). https://doi.org/10.1145/3377817.3377834
6. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7**(1), 1–30 (2006). http://jmlr.org/papers/v7/demsar06a.html
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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, Minneapolis, Minnesota, pp. 4171–4186. Association for Computational Linguistics, June 2019. https://doi.org/10.18653/v1/N19-1423
8. Dietterich, T.G.: Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput. **10**(7), 1895–1923 (1998)
9. Dozat, T., Manning, C.D.: Deep biaffine attention for neural dependency parsing. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 24–26 April 2017, Conference Track Proceedings. OpenReview.net (2017). https://openreview.net/forum?id=Hk95PK9le
10. Fan, Z., Wu, Z., Dai, X.Y., Huang, S., Chen, J.: Target-oriented opinion words extraction with target-fused neural sequence labeling. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, Minnesota, pp. 2509–2518. Association for Computational Linguistics, June 2019. https://doi.org/10.18653/v1/N19-1259
11. Farias, D.S., Matsuno, I.P., Marcacini, R.M., Rezende, S.O.: Opinion-meter: a framework for aspect-based sentiment analysis. In: Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web, pp. 351–354. ACM (2016). https://doi.org/10.1145/2976796.2988214
12. Freitas, C., Motta, E., Milidiú, R., César, J.: Vampiro que brilha... rá! desafios na anotaçao de opiniao em um corpus de resenhas de livros. Encontro de Linguística de Corpus **11**, 22 (2012)
13. Freitas, L.A., Vieira, R.: Exploring resources for sentiment analysis in Portuguese language. In: Brazilian Conference on Intelligent Systems, pp. 152–156. IEEE (2015)
14. García, S., Herrera, F.: An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. J. Mach. Learn. Res. **9**(89), 2677–2694 (2008). http://jmlr.org/papers/v9/garcia08a.html
15. Goldberg, Y.: Assessing Bert's syntactic abilities. arXiv (2019). http://arxiv.org/abs/1901.05287
16. Han, H., Li, X., Zhi, S., Wang, H.: Multi-attention network for aspect sentiment analysis. In: Proceedings of the 2019 8th International Conference on Software and Computer Applications, ICSCA 2019, pp. 22–26. Association for Computing Machinery, New York (2019). https://doi.org/10.1145/3316615.3316673

17. He, R., Lee, W.S., Ng, H.T., Dahlmeier, D.: An interactive multi-task learning network for end-to-end aspect-based sentiment analysis. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, pp. 504–515. Association for Computational Linguistics, July 2019
18. Huang, B., Carley, K.: Syntax-aware aspect level sentiment classification with graph attention networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Hong Kong, China, pp. 5469–5477. Association for Computational Linguistics, November 2019. https://doi.org/10.18653/v1/D19-1549
19. Li, Y., Li, Z., Zhang, M., Wang, R., Li, S., Si, L.: Self-attentive biaffine dependency parsing. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI, pp. 5067–5073. ijcai.org (2019). https://doi.org/10.24963/ijcai.2019/704
20. Li, Z., et al.: Dependency or span, end-to-end uniform semantic role labeling. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 6730–6737 (2019). https://doi.org/10.1609/aaai.v33i01.33016730
21. Liu, B.: Sentiment analysis and opinion mining. Synthesis Lect. Hum. Lang. Technol. **5**(1), 1–167 (2012)
22. Lu, Z., Du, P., Nie, J.-Y., et al.: VGCN-BERT: augmenting BERT with graph embedding for text classification. In: Jose, J.M. (ed.) ECIR 2020. LNCS, vol. 12035, pp. 369–382. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-45439-5_25
23. Machado, M.T., Pardo, T.A.S., Ruiz, E.E.S.: Analysis of unsupervised aspect term identification methods for Portuguese reviews. In: Anais do XIV Encontro Nacional de Inteligncia Artificial e Computacional (ENIAC), SBC, pp. 239–249 (2017)
24. Peng, H., Xu, L., Bing, L., Huang, F., Lu, W., Si, L.: Knowing what, how and why: a near complete solution for aspect-based sentiment analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 05, pp. 8600–8607 (2020). https://doi.org/10.1609/aaai.v34i05.6383
25. Pereira, D.A.: A survey of sentiment analysis in the Portuguese language. Artif. Intell. Rev. **54**(2), 1087–1115 (2020). https://doi.org/10.1007/s10462-020-09870-1
26. Saias, J., Mourão, M., Oliveira, E.: Detailing sentiment analysis to consider entity aspects: an approach for Portuguese short texts. Trans. Mach. Learn. Artif. Intell. **6**, 26–35 (2018)
27. Souza, F., Nogueira, R., Lotufo, R.: BERTimbau: pretrained BERT models for Brazilian Portuguese. In: Cerri, R., Prati, R.C. (eds.) BRACIS 2020. LNCS (LNAI), vol. 12319, pp. 403–417. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-61377-8_28
28. Vargas, F.A., Pardo, T.A.S.: Hierarchical clustering of aspects for opinion mining: a corpus study. In: Finatto, M.J.B., Rebechi, R.R., Sarmento, S., Bocorny, A.E.P. (eds.) Linguística de Corpus: Perspectivas, pp. 69–91 (2018)
29. Wang, W., Pan, S.J., Dahlmeier, D., Xiao, X.: Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31 (2017)
30. Wu, Z., Ying, C., Zhao, F., Fan, Z., Dai, X., Xia, R.: Grid tagging scheme for aspect-oriented fine-grained opinion extraction. In: Findings of the Association for Computational Linguistics, pp. 2576–2585. Association for Computational Linguistics, November 2020. https://doi.org/10.18653/v1/2020.findings-emnlp.234

31. Xu, L., Bing, L., Lu, W., Huang, F.: Aspect sentiment classification with aspect-specific opinion spans. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 3561–3567. Association for Computational Linguistics, November 2020. https://doi.org/10.18653/v1/2020.emnlp-main.288
32. Zhang, C., Li, Q., Song, D., Wang, B.: A multi-task learning framework for opinion triplet extraction. In: Findings of the Association for Computational Linguistics, pp. 819–828. Association for Computational Linguistics, November 2020. https://doi.org/10.18653/v1/2020.findings-emnlp.72