JOSE DE JESUS MELENDEZ BARROS

A Deep Learning approach for Aspect Sentiment Triplet Extraction in Portuguese and Spanish

São Paulo
2022

JOSE DE JESUS MELENDEZ BARROS

A Deep Learning approach for Aspect Sentiment Triplet Extraction in Portuguese and Spanish

Versão Corrigida

Dissertation submitted to Escola Politécnica of the Universidade de São Paulo to obtain the degree of Master of Science

Concentration Area:
Computer Engineering

Advisor:
Glauber De Bona

São Paulo
2022

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, ______ de ____________________ de __________

Assinatura do autor: ________________________

Assinatura do orientador: ________________________

Catalogação-na-publicação

Melendez Barros, Jose de Jesus
A Deep Learning approach for Aspect Sentiment Triplet Extraction in Portuguese and Spanish / J. J. Melendez Barros -- versão corr. -- São Paulo, 2022.
51 p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Computação e Sistemas Digitais.

1.Inteligência Artificial 2.Aprendizado de Máquina 3.Processamento de Linguagem Natural I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Computação e Sistemas Digitais II.t.

Nome: Melendez Barros, Jose de Jesus

Título em inglês: A Deep Learning approach for Aspect Sentiment Triplet Extraction in Portuguese and Spanish

Título em português: Uma abordagem de Aprendizado Profundo para Extração de Trigêmeos de Sentimento de Aspecto em português e espanhol

Dissertação apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do título de Mestre em Engenharia Elétrica.

Aprovado em:

Banca Examinadora

Prof. Dr(a). ______________________________
Instituição: ______________________________
Julgamento: ______________________________

Prof. Dr(a). ______________________________
Instituição: ______________________________
Julgamento: ______________________________

Prof. Dr(a). ______________________________
Instituição: ______________________________
Julgamento: ______________________________

Dedicatória

A mis padres, que me enseñaron que no hay futuro sin educación y disciplina

ACKNOWLEDGMENTS

I would like to thank PCS, and all the people at Escola Politecnica, for supporting my academic development. I would like to especially thank my advisor Glauber De Bona who guided me through all the research and development process in this work.

Also, I thank my parents, who always gave me the necessary support in my academic life.

To the Ministerio de Ciencia y Tecnología de Colombia (MINCIENCIAS) for financing my studies.

To Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)

RESUMO

Aspect Sentiment Triplet Extraction (ASTE) é uma subtarefa de Aspect-Based Sentiment Analysis (ABSA), que visa extrair pares de opiniões e aspectos de uma frase e identificar a polaridade de sentimento associada a eles. Por exemplo, dada a sentença "Os quartos são amplos e ótimo café da manhã", ASTE gera o tripleto T = {(quartos, amplos, positive), (café de manhã, ótimo, positive)}. Embora várias abordagens tenham sido propostas recentemente, os trabalhos disponíveis em português e espanhol têm se limitado, em sua maioria, a extrair apenas aspectos, sem abordar tarefas ASTE. Este trabalho tem como objetivo desenvolver um framework baseado em Deep Learning para executar a tarefa de Aspect Sentiment Triplet Extraction em português e espanhol. O framework usa o BERT como um codificador de frase de consciência de contexto, várias camadas paralelas não lineares obtêm as representações dos aspectos e as opiniões e uma camada Graph Attention junto com um scorer não linear determina a dependência de sentimento entre cada par aspecto-opinião. Os resultados mostram que nosso framework proposto supera significativamente as linhas de base em português-espanhol e é competitiva com suas contrapartes em inglês.

Palavras-Chave – Aprendizagem profunda, Processamento de linguagem natural, Análise de sentimento baseada em aspectos, Extração Tripla de Sentimento de Aspecto.

ABSTRACT

Aspect Sentiment Triplet Extraction (ASTE) is an Aspect-Based Sentiment Analysis subtask (ABSA), which 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/Spanish 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 and Spanish. 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/Spanish and is competitive with its counterparts in English.

Keywords – Deep Learning, Natural Language Processing, Aspect-Based Sentiment Analysis, Aspect Sentiment Triplet Extraction.

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ABSA Aspect-Based Sentiment Analysis. 13, 15, 17, 34

AFOE Aspect-oriented Fine-grained Opinion Extraction. 18

ASTE Aspect Sentiment Triplet Extraction. 13–15, 17, 18

ATC Aspect/target Term Sentiment Classification. 17

ATE Aspect Term Extraction. 17

CRF Conditional Random Fields. 22, 24

E2E-ABSA End-to-End Aspect-Based Sentiment Analysis. 17, 18

GAL Graph Attention Layer. 35

GAM Graph Attention Module. 35

GAT Graph Attention Network. 19, 20

OPE Opinion Pair Extraction. 17, 18

OTE Opinion Term Extraction. 17

SVM Support Vector Machine. 24

CONTENTS

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 (LIU, 2012). 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 (LIU, 2012; ZHOU et al., 2019; PEREIRA, 2021), 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 (PENG et al., 2020).

ABSA tasks tackled in Portuguese include aspect term extraction (BALAGE, 2017; MACHADO; PARDO; RUIZ, 2017), aspect sentiment classification (AIRES et al., 2018; SAIAS; MOURÃO; OLIVEIRA, 2018) and aspect opinion co-extraction (CARDOSO; PEREIRA, 2020), but the proposed solutions employ methods that are considerably behind the state of the art in reference languages (PEREIRA, 2021). The picture is similar in Spanish, where most of the works focus on polarity detection and are limited to using context-independent word vectors (FastTex, Glove, word2vec) (ANGEL; NEGRÓN; ESPINOZA-VALDEZ, 2021). A few jobs have been developed on ABSA tasks, such as Aspect Term Extraction (HENRIQUEZ; SANCHEZ-TORRES, 2021;ARAQUE et al., 2015) and Aspect-Sentiment Pair Extraction (MIRANDA; BUELVAS, 2019; AKHTAR et al., 2019). Besides, these works are focused on identifying aspects or opinions separately. Our work aims to extract aspects, opinions and sentiment polarity simultaneously; this provides richer context output and explains why a specific sentiment polarity was assigned.

1.1 Our problem: Aspect Sentiment Triplet Extraction

Let S be a sentence *"Large rooms and great breakfast but location was horrible"*. Aspects are word sequences describing attributes or features of the targets (e.g., rooms, breakfast and location), opinions are those expressions carrying subjective attitudes or (un)desirable characteristics (e.g., large, great and horrible) and sentiments are the polarities associated with aspect-opinion pair, which can be positive, negative or neutral.

Considering the above, our Aspect Sentiment Triplet Extraction (ASTE) (PENG et al., 2020) 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 a n-gram aspect term, a 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.

The main difficulties in this task are: (i) to get an appropriate representation of a sentence, which highlights features relevant to ASTE; (ii) to handle long questions and superfluous information for aspect and opinion tagging; (iii) to delimit the start and end of an aspect/opinion n-gram term when n is large; (iv) to pair the aspect term with its corresponding opinion term; (v) to deal with overlapping relationships between aspects and opinions.

1.2 Objective

This dissertation aims to develop a framework based on deep learning to perform the aspect sentiment triplet extraction task in Portuguese and Spanish. Therefore, our model extracts aspect-opinion pairs and identify the sentiment polarity between them. Our framework is based the state-of-the-art models for English, building on the work of ZHANG et al.. 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 module to capture syntactic relationships between two word vectors.

1.3 Contributions

Motivated by the above observations, we present an approach to deal with the extraction of triplets jointly for ASTE. The comparison results show that our proposed framework outperforms the baselines by more than 10% on average. Besides, the ablation study demonstrates the effectiveness of Graph Attention layers and BiLSTM layers for dimensionality reduction. The main expected contributions of this work can be summarized as follows:

- We propose the first model to deal with ASTE tasks in Portuguese and Spanish. To the best of our knowledge, all works in Portuguese/Spanish-ABSA have been mainly limited to Aspect Extraction.
- We provide two new datasets (Portuguese and Spanish) to work with ASTE tasks. Additionally, we adapt an existing Portuguese dataset for this task (ReLi Corpus, FREITAS et al., 2012). Other datasets Freitas e Vieira (2015), Farias et al. (2016), Vargas e Pardo (2018) related to ABSA in Portuguese do not have the proper labeling for ASTE tasks.
- Finally, we will make a repository available to serve as a basis for developing other works in Spanish/Portuguese-ABSA, to our knowledge, none of the research in Portuguese/Spanish cited in our work - except Farias et al. (2016) - made its source code available.

1.4 Publication

Most of the results presented here were published in 10th Brazilian Conference on Intelligent System (BRACIS 2021). The article, entitled “A Deep Learning approach for Aspect Sentiment Triplet Extraction in Portuguese” was presented at the conference that took place from November 29th to December 3th, 2021 (MELENDEZ; De Bona, 2021).

1.5 Organization of the work

The rest of this work is divided as follows: in Chapter 2 we briefly review some important concepts in Aspect-Based Sentiment Analysis and other theoretical aspects of this research. Chapter 3 discusses related works. Experiments and results are reported in Chapter 4 and 5 respectively.

2 BACKGROUND

Given an opinion document d, full sentiment analysis aims to discover all opinion quintuples $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$ in d (LIU, 2012), where:

- e_i is an entity, i.e., a product or any object (e.g., hotel)
- a_{ij} is a aspect of the entity (e.g., location)
- h_k is the opinion holder (e.g., guest)
- t_l is the time when the opinion is expressed by h_k (e.g., one month ago)
- s_{ijkl} is the sentiment on aspect a_{ij} of entity e_i (e.g., positive, negative or neutral)

2.1 Levels of sentiment analysis

According to (Balaji; Nagaraju; Haritha, 2017), sentiment analysis have been addressed at three levels:

- **Document level:** This consists of classifying whether a whole review is associated with positive or negative sentiment. This level assumes that the review is addressed to a single product.
- **Sentence level:** the task in this level identifies the polarity of each of the sentences that form the review. In this task, the classifier should be able to determine which sentences express subjective views and opinions.
- **Aspect level:** This performs finer-grained analysis than document and sentence level. This can identify the target(s) towards which an opinion(s) is(are) directed. For example, in the sentence "*Os quartos são amplos e ótimo café da manhã.*", aspect level output [(quartos, amplos), (cafe de manhã, ótimo)]

This work deals with sentiment analysis in the aspect level.

2.2 Aspect-based Sentiment Analysis Environment

The four main sub-tasks in ABSA are: Aspect category detection, Aspect Term Extraction (ATE), Aspect/target Term Sentiment Classification (ATC) and Opinion Term Extraction (OTE). Other subtasks that arise as a combination of the above ones are: Opinion Pair Extraction (OPE), Aspect and opinion co-extraction, End-to-End Aspect-Based Sentiment Analysis (E2E-ABSA) and Aspect Sentiment Triplet Extraction (ASTE). Let us consider the following overlapped and non-overlapped sentences to explain each of them:

(i) *Bathrooms are modern but a little dirty.*

(ii) *Large rooms and great breakfast.*

- Aspect category detection aims to group aspects of reviews into categories. Output >> {i: [infrastructure], ii:[infrastructure, food]}
- Aspect Term Extraction (ATE), also known as opinion target extraction, aims to identify the word sequences in the sentence, which describe attributes or features of the targets. Output >> {i: [bathrooms], ii:[rooms, breakfast]}
- Aspect/target Term Sentiment Classification (ATC), given a specific aspect, determines the sentiment associated with that (XU et al., 2020). Output >> Given "rooms" then [positive]
- Opinion Term Extraction (OTE), also known as opinion word extraction, aims to get those expressions carrying subjective attitudes. Output >> {i: [modern, dirty], ii:[large, great]}
- Opinion Pair Extraction (OPE)(In cases where the aspect is explicitly inputted into the model, the task is called target-oriented opinion words-extraction (FAN et al., 2019) - TOWE -) extracts aspect-opinion tuples. For each aspect term, it identifies the opinion that corresponds to it. It may or may not extract overlapped relations (WU et al., 2020). Output >> {i: [(bathrooms, modern), (bathrooms, dirty)], ii:[(rooms, large), (breakfast, great)]}
- Aspect and opinion co-extraction extract aspects and opinions, but unlike the previous one, these terms will not be paired. It extracts aspects and opinions in independent lists. It does not matter whether or not a sentence is overlapped. Output >> {i: [(bathrooms), (modern, dirty)], ii:[(rooms, breakfast), (large, great)]}

- End-to-End Aspect-Based Sentiment Analysis (E2E-ABSA), also known as Aspect-sentiment pair extraction, combines ATE and ATC. E2E-ABSA, like OPE, may or may not extract overlapped relations between aspects and opinions. Output >> {i: [(bathrooms, positive), (bathrooms, negative)], ii:[(rooms, positive), (breakfast, positive)]}

- Aspect Sentiment Triplet Extraction (ASTE) or Opinion triplet extraction gets aspect-opinion pairs and the sentiment associated with them (PENG et al., 2020). ASTE may or may not extract overlapped relations between aspects and opinions. Output >> {i: [(bathrooms, modern, positive), (bathrooms, dirty, negative)], ii:[(rooms, large, positive), (breakfast, great, positive)]}

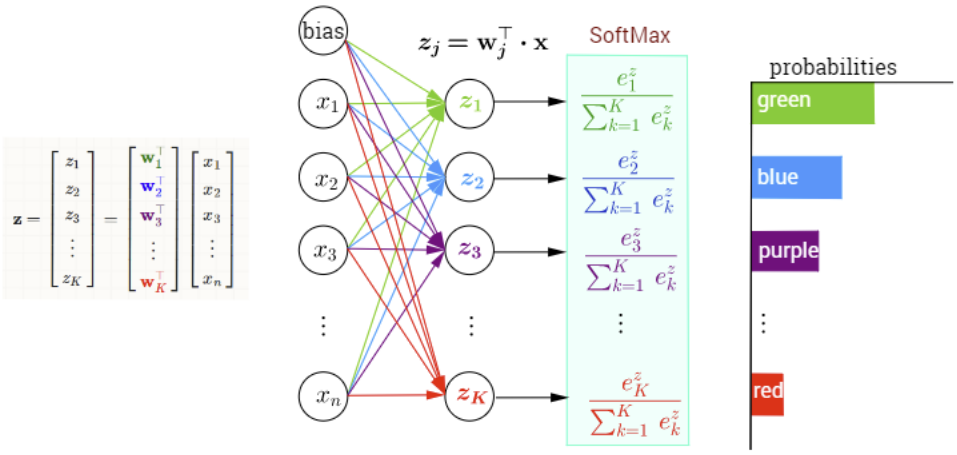
Another term used is Aspect-oriented Fine-grained Opinion Extraction (AFOE), which aims to extract opinion pairs (OPE) or opinion triplets (ASTE).

2.3 Classification: Softmax function and Cross-Entropy

In the context of classification problems, the softmax function is used to obtain a vector of probabilities, where each component is related to a category and the coefficients are associated with the probability of the category being the true value. Within a neural network, the Softmax function is commonly applied to the outputs of the last layer of the network (equation 4.5 and 4.6). The final calculation of the distribution is shown in Figure 1. When we work with a multi-class classification task, the target variable containing the class labels is first encoded to a one-hot vector, meaning that an integer is applied to each class label from 0 to N-1, where N is the number of category labels and the ones indicates the true value.

During training, the one-hot target vector is compared to the Softmax probability vector. The difference between them is used to optimize the prediction process by updating the connection weights. Usually, the difference between the two distributions is calculated using cross-entropy (loss function). The true probability p_i is the true label (one-hot vector), and the given distribution q_i is the predicted value of the current model (Softmax probability vector). Figure 2 exemplifies cross-entropy calculation. Given five categories {cat, dog, tiger, fish, bird}, the true value for the input "dog" is the one-hot vector [0 1 0 0 0]. Since the network weights are randomly initialized, it is unlikely that the result of the first iteration will match the true value. The greater the number of iterations, the softmax probability vector will converge to the true value, and therefore, the value of the cross-entropy will decrease.

Figure 1: Softmax Activation Function.



Source: http://rinterested.github.io/statistics/softmax.html

Figure 2: Cross-entropy.

Input: Dog		true value	Prediction t = 0	Prediction t = 1	....	Prediction t = n
Target	cat	0	0.1	0.16		0.05
	dog	1	0.1	0.3	....	0.9
	tiger	0	0.05	0.02	....	0.03
	fish	0	0.7	0.5	....	0.01
	bird	0	0.05	0.02		0.01

$$H_{dog}^{t=0} = -\sum_{k\in\{O,B,I\}} p(k)\log q(k) = -p(cat)\log q(cat) - p(dog)\log q(dog)$$

$$-p(tiger)\log q(tiger) - p(fish)\log q(fish)$$

$$-p(bird)\log q(bird) = 3.32$$

$$H_{dog}^{t=1} = 1.74$$

$$H_{dog}^{t=n} = 0.15$$

Source: Authors' own creation.

2.4 Graph Attention Networks

According to Dagar et al. (2020), "Graph Attention Network (GAT) is a novel neural network architecture that operates on graph-structured data, leveraging masked selfattentional layers to address the shortcomings of prior methods based on graph convolutions or their approximations". GAT computes the weights of the attention mechanism based on the connections that exist between the nodes. For example, in a semantic dependency graph, the network will use the connections between words to compute attention. Unlike graph networks based on convolutional architectures, GAT does not need to apply costly matrix operation, being computationally more efficient (VELICKOVIC et al., 2018).

In convolutional networks, the relationship between a pair of nodes (i, j) is calculated in Equation 2.1. Where j belongs to the neighborhood $\mathcal{N}$ of i (all nodes directly connected to i), and $c_{ij} = \sqrt{|\mathcal{N}(i)|}\sqrt{|\mathcal{N}(j)|}$ is a normalization constant. In this method, the coefficients that code the relationship between two nodes are strongly influenced by the number of neighbors of i and j. This static mechanism hardly captures the richness and complexity of language. The result is structure-dependent relationships, which may hurt its generalizability. In GAT, an attention mechanism replaces that static normalization. By doing that, GAT implicitly captures the weight so that more important nodes receive higher weight during neighborhood aggregation.

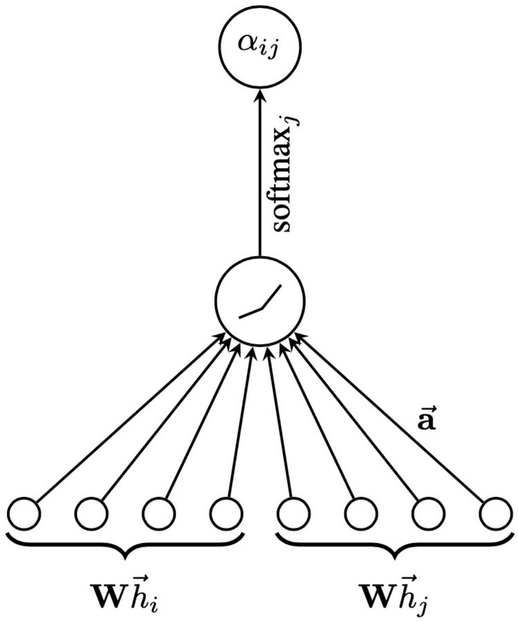
$$h_i^{(l+1)} = \sigma\left(\sum_{j\in\mathcal{N}(i)} \frac{1}{c_{ij}} \mathbf{W}^{(l)}\mathbf{h}_j^{(l)}\right) \tag{2.1}$$

Equations 2.2–2.6 show how the GAT mechanism works. According to Velickovic et al. (2018), equation 2.2 is a linear transformation to obtain sufficient expressiveness to transform the input features h_i into higher-level features. Equation 2.4 computes a pairwise un-normalized attention score between two neighbors that indicate the importance of node j's features to node i. First, we concatenate the vectors of the word pair (i, j). Nodes i and j represent two words related syntactically within the semantic dependency graph. Then, it computes the dot product between such concatenation and a learnable weight vector to calibrate the attention between the nodes. Lastly, a LeakyReLU function is applied to the result so that the model responds to non-linear variations. Figure 3 shows the attention mechanism initially proposed by Velickovic et al. (2018).

Equation 2.5 normalizes the coefficients by using the softmax function; this makes it easy to compare different nodes. Finally, in equation 2.6 the embeddings from neighbors are aggregated together, scaled by the attention scores. Some advantages of GAT over GCN are: GAT is computationally more efficient; GAT allows for assigning different importances to nodes of the same neighborhood; The graph is not required to be undirected;

GAT can use multi-head attention to enrich the model capacity (DAGAR et al., 2020).

Figure 3: Graph Attention Mechanism.



Source: Velickovic et al. (2018)

$$\mathbf{z}_i^{(l)} = \mathbf{W}^{(l)}\mathbf{h}_i^{(l)} \tag{2.2}$$

$$\mathbf{z}_j^{(l)} = \mathbf{W}^{(l)}\mathbf{h}_j^{(l)} \tag{2.3}$$

$$\mathbf{e}_{ij}^{(l)} = \text{LeakyReLU}(\vec{\mathbf{a}}^{(l)^T}(\mathbf{z}_i^{(l)}; \mathbf{z}_j^{(l)})) \tag{2.4}$$

$$\alpha_{ij}^{(l)} = \frac{\exp(\mathbf{e}_{ij}^{(l)})}{\sum_{k\in\mathcal{N}(i)} \exp(\mathbf{e}_{ik}^{(l)})} \tag{2.5}$$

$$h_i^{(l+1)} = \sigma\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}^{(l)}\mathbf{z}_j^{(l)}\right) \tag{2.6}$$

3 RELATED WORK

As we have previously discussed, ABSA includes several subtasks that vary according to the domain of the target term. (XU et al., 2020) developed a model based on BiLSTM networks, Attention and Multiple linear-chain CRF, aiming to extract an opinion term given an aspect term as input. Their main contribution is the model's capability of capturing opinion spans as a whole or variable-length opinion spans. One of the disadvantages is that it cannot manage overlapped aspects/opinions.

He et al. (2019) 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 e Maojie (2019) and Han et al. (2019). Both works employed BiGRU/LSTM and attention mechanisms to attenuate irrelevant information, but neither of them can manage overlapped sentiments. The main differences between them are: the former requires passing the target aspects as input and the latter can deal with unbalanced datasets.

Fan et al. (2019) 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. Unlike the above ones, this work can handle overlapped aspect /opinion. Meanwhile, Zhao et al. (2020) used BERT encoder along with MLP to attempt the same task, this model can extract all aspect-opinion pairs in one step instead of pipelines, like the previous one, it can manage overlapped terms, but only raw reviews are required as input.

3.1 Aspect sentiment triplet extraction Task

Recently, several works have been developed to extract triplets from a sentence (aspect, opinion and sentiment polarity), also known as ASTE (ZHANG et al., 2020). This task was proposed by Peng et al. (2020), 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. (2020) 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 review is required as input; this model can extract all opinion factors of OPE 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. (2020) 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 four independent ReLU layers, these layers apply dimension-reduction to strip away irrelevant features. Two of the layers yield aspect $\mathbf{r}_i^{(ap)}$ and opinion $\mathbf{r}_i^{(op)}$ representations, which are prepared for tagging. The outputs of the remaining layers ($\mathbf{r}_i^{(ap)\prime}$ and $\mathbf{r}_i^{(op)\prime}$) are used for sentiment parsing. The $\mathbf{r}_i^{(ap)}$ and $\mathbf{r}_i^{(op)}$ representations are passed to two independent softmax layers whose objective is to obtain distributions over $\{B, I, O\}$ tags. The B tag indicates that the token is the beginning of a aspect-opinion term, I tag indicates that the token is inside a aspect-opinion term (when the term consists of more than one token) and O tag indicates that a token belongs to no aspect-opinion term. Regarding $\mathbf{r}_i^{(ap)\prime}$ and $\mathbf{r}_i^{(op)\prime}$ vectors, these are passed to a biaffine scorer (DOZAT; MANNING, 2016) to determine if the sentiment dependency between each word pair is neutral, negative, positive, or null. This study used more recent baselines compared to the previously mentioned works and can be considered the state-of-the-art at the time of this review.

3.2 Aspect-based Sentiment Analysis in Portuguese

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 e Pereira (2020) used CRF and opLexicon to deal with aspect sentiment classification. Aires et al. (2018) developed two models, one based on SVM and one employing an LSTM network. Saias, Mourão e Oliveira (2018) 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 (2017) used clustering techniques, word2vec embeddings and CRF to implement an unsupervised model.

3.3 Aspect-based Sentiment Analysis in Spanish

As in Portuguese, to the best of our knowledge, no model has been developed to solve ASTE in Spanish. Most of the sentiment analysis works in Spanish focus on sentiment classification at the document or sentence level (ANGEL; NEGRÓN; ESPINOZA-VALDEZ, 2021). Henriquez e Sanchez-Torres (2021) developed a semantic model for aspect extraction in Spanish. The system uses ontology, semantic similarity and co-occurrence matrix to detect explicit and implicit aspects. Miranda e Buelvas (2019) worked on a semantic model that integrates NLP, ontologies, semantic similarity and unsupervised automatic learning. This model comprises an aspect extractor that extracts the explicit and implicit aspects and a sentiment identifier that determines the polarity associated with an aspect. They aim to resolve Aspect-Sentiment Pair Extraction. These models do not use word vectors but are mainly based on co-occurrences and Pointwise Mutual Information (PMI); therefore, their ability to deal with context is limited.

Akhtar et al. (2019) proposed a language-agnostic deep neural network architecture for aspect-based sentiment analysis in six languages, including Spanish. The proposed approach is based on Word2Vec and Bidirectional LSTM networks and tackles two ABSA tasks, Aspect Term Extraction and Aspect sentiment classification. In these tasks, they obtained, respectively, 73% F1-score and 87.1% Accuracy. Finally, Araque et al. (2015) designed and developed an adaptable system based on Support Vector Machine, extracting aspect terms and classifying the detected aspects in polarity values.

4 METHODS

This work emphasizes on the aspect of reproducibility and statistical evaluation of the results. The practices considered for this purpose include, but are not limited to:

- Explicitly identify limitations or technical assumptions.
- All source code and datasets used in this work were included in a public repository with its respective DOI.
- All statistical assumptions were validated before applying a particular test.
- Random process settings are described to allow replication of results.
- The computing infrastructure used for running experiments are specified.
- The development and documentation of the code are according to the best Python practices.
- This work states the number and range of values tried per hyperparameter during its development.

4.1 Data management plan

This plan is based on the Data Management Plan guidelines[1] provided by Universidade de São Paulo (USP)

4.1.1 Data Creation and Collection

We conducted experiments in Brazilian Portuguese, Spanish and English datasets. The data used includes reviews of products and services. The data is made up of 8511 sentences distributed in 7 datasets in CSV format. A dataset was produced during this research from hotel reviews (hereinafter referred to as ReHol dataset). Portuguese reviews

[1] https://www.aguia.usp.br/apoio-pesquisador/dados-pesquisa/plano-gestao-dados-2

were collected from TripAdvisor and triplets were manually annotated by us following the pattern shown in Figure 4. Portuguese ReLi dataset was made available by (FREITAS et al., 2012); however, we re-formatted the triplet labels to facilitate the data ingestion process during model training. Spanish RestES dataset was also partially produced during this research. Reviews were extracted from Twitter and made available by (DUBIAU; ALE, 2013). Triplets were manually annotated by us. These datasets are available on IEEEDataport[2].

We used four English datasets[3] in the Laptop (Lap14) and Restaurant (Rest14) domain from Semeval 2014 Task4 (PONTIKI et al., 2014), Semeval 2015 Task 12 (Rest15) (PONTIKI et al., 2015) and Semeval 2016 Task 5 (Rest16) (PONTIKI et al., 2016). A detailed description all of the datasets is shown in Table 1.

Table 1: Dataset details.

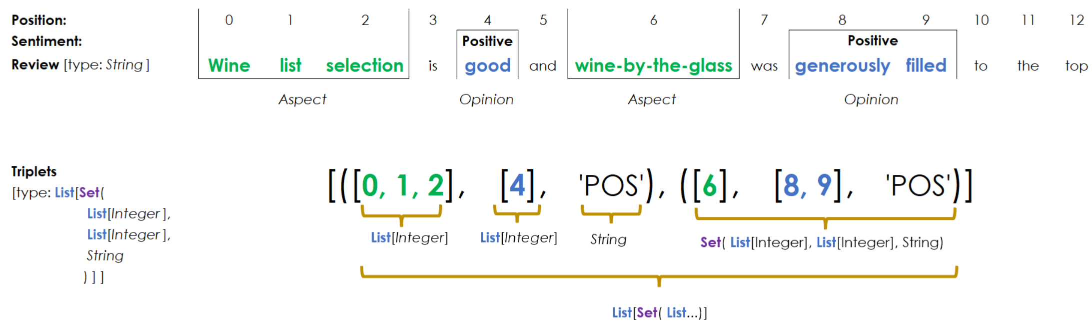
Nomenclature	ReHol	ReLi	RestES	Rest14	Rest15	Rest16	Lap14
Produced by this research	Yes	Partially	Partially	No	No	No	No
Language	PT	PT	ES	EN	EN	EN	EN
Domain	Hotel	Book	Restaurant	Restaurant	Restaurant	Restaurant	Laptop
# Triplets	3199	1448	3085	3906	1747	2247	2349
# Positive triplets	1846	1170	2707	2867	1285	1674	1350
# Negative triplets	1242	278	334	753	401	483	774
# Neutral triplets	111	0	44	286	61	90	225
# Non-overlapped sentences	941	767	454	1427	811	1055	1040
# Overlapped sentences	573	242	404	456	264	338	413
Avg. words per sentence	16	22	23	17	15	15	18
Min. sentence size (words)	8	2	7	3	2	2	3
Max. sentence size (words)	38	193	63	79	68	78	83
Q1 sentence size (words)	11	9	14	10	9	8	11
Q3 sentence size (words)	19	28	29	22	19	19	23

4.1.2 Metadata: Attribute Description

The datasets contain two fields: "Review" is a string field type, where a user expresses an opinion about some product/service. "Triplet" is a field that mixes integers and strings. This field indicates the positions of aspect and opinion terms in a sentence. Figure 4 shows an example of Review and Triplet attributes. Aspect $List[Integer]$ indicates the positions of the tokens that belong to an aspect term. Opinion $List[Integer]$ indicates the positions of the tokens that belong to an opinion term. Sentiment $String$ indicates sentiment polarity (POS: positive, NEG: negative, NEU: neutral).

[2]https://dx.doi.org/10.21227/0ej1-br13
[3]Datasets are available by Zhang et. al. at https://github.com/GeneZC/OTE-MTL

Figure 4: Attribute Description



Source: Authors' own creation.

4.1.3 Ethical, copyright and legal issues

ReLi, RestES, Rest14, Lap14, Rest15 and Rest16 datasets are in the public domain and do not have any use restriction whenever the author is mentioned and correctly referenced. These datasets also do not contain sensitive personal information.

Regarding ReHol dataset, these reviews were collected from TripAdvisor. TripAdvisor did not raise any objection to use these data since we clearly fall under academic fair use. The research is only using a small amount of data, and it is a "transformative" use since we are only using a sentence from each review. The data extracted correspond exclusively to the 1514 sentences. The reviews are only about hotels. The sentences are from the Brazilian Portuguese language. The project is strictly academic.

4.1.4 Storage and Backup

During the project's development, the backup frequency was made according to the level of activity in the project. During high activity periods, the backup was done whenever a new model was trained. During reduced activity periods, the backup was updated every fourteen days. Manual backup was done whenever an important milestone was reached.

The author was responsible for configuring the backup frequency and ensuring that they were done correctly. Backups were tested every month to ensure that saved information was recoverable. We have distributed multiple remote copies of digital files to keep data safe. The main repositories that store these copies are IEEEDataport[4] and GitHub[5].

[4] https://dx.doi.org/10.21227/0ej1-br13

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

4.2 Baselines

Existing Portuguese/Spanish-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/Spanish by replacing the English pre-trained word embedding with Portuguese and Spanish pre-trained word embeddings [6] [7]. No further changes were made. Additionally, our model was trained in English to compare with the actual performance of these baselines.

- **OTE-MTL** (ZHANG et al., 2020) 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 (DOZAT; MANNING, 2016) is used to get aspect-opinion pairs along with the associated sentiment with each other.
- **CMLA-MTL** (WANG et al., 2017) 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 Proposed framework

As mentioned in Section 1.2, we propose a deep learning framework for opinion triplet extraction based on the work of Zhang et al. (2020). The main differences between them and our work are: Instead of Glove + BiLSTM for sentence encoding, we use Portuguese BERT developed by NeuralMind (SOUZA; NOGUEIRA; LOTUFO, 2020), Spanish BERT (CANETE et al., 2020) and BERT multilingual developed by Google Research (DEVLIN et al., 2018). We use a Graph Attention Module with syntactic dependencies as an adjacency matrix, which enables our method to encode explicit syntactic relations between aspects and opinions. Figure 5 brings an overview of the model. Sections 4.3.1 and 4.3.5 expand on the reasons and advantages of these modifications. Figures 6 and 7 show ZHANG et al. model and ours, respectively.

[6] http://nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc
[7] https://github.com/dccuchile/spanish-word-embeddings

Figure 5: Model overview

Let us consider the sentence "room service is great" to generate the triplets

Words	Room	service	is	great
Indexes	0	1	2	3

Sentence Encoder

Let us assume that the word vectors have four dimensions. The sentence encoder generates a vector for each word that encodes semantic and contex information.

	d_0	d_1	d_2	d_3
room	1.1	0.4	1.1	0.3
service	0.7	1.4	0.2	0.2
is	0.6	0.9	1.1	0.4
great	1.3	0.4	0.8	1.5

Let us assume that the even indices of the word vector components encode semantic information alluding to the concept of aspect. As for the odd indexes, it encodes information about the concept of opinion.

Aspect and Opinion Representation

When an aspect word passes through the aspect representation node, the weight of the d0 and d2 components will be enhanced, while d1 and d3 will be attenuated. Analogously it will happen if a word that is an opinion goes through the opinion representation node.*

Aspect Representation Node

	d_0	d_1	d_2	d_3
room	**1.7**	0.2	**2.8**	0.1
service	**1.9**	0.1	**2.7**	0.1
is	0.5	0.4	0.3	0.4
great	0.4	0.3	0.5	0.5

aspect features

Opinion Representation Node

	d_0	d_1	d_2	d_3
room	0.5	0.4	0.1	0.3
service	0.4	0.5	0.5	0.4
is	0.1	0.2	0.3	0.1
great	0.2	**1.8**	0.1	**2.9**

opinion features

Aspect and Opinion Tagging

To represent an aspect or opinion term, we use {B, I, O} tags. The B tag indicates that the token is the beginning of a aspect-opinion term, I tag indicates that the token is inside a aspect-opinion term (when the term consists of more than one token) and O tag indicates that a token belongs to no aspect-opinion term

Aspect tagging head generates a probability vector for each word, where the highest probability indicates which tag corresponds to the word. For example, the word "room" will produce a probability vector, where **B** tag will have the highest probability; therefore, it would be the starting word of an aspect. When passing through the ***opinion tagging head***, this same word would generate a vector such that **O** tag will have the highest probability, indicating that the word has no relation to an opinion term.

Aspect Tagging Head

	B	I	O	
room	**0.7**	~~0.2~~	~~0.1~~	→ B
service	0.3	**0.6**	~~0.1~~	→ I
is	0.1	0.1	**0.8**	→ O
great	0.2	0.2	**0.6**	→ O

Opinion Tagging Head

	B	I	O	
room	0.1	0.1	**0.8**	→ O
service	0.2	0.1	**0.7**	→ O
is	0.2	0.3	**0.5**	→ O
great	**0.6**	~~0.3~~	~~0.1~~	→ B

Dependency Parsing

Determine the type of sentiment dependence between an aspect and an opinion term. Analogously to aspect/opinion tagging, the dependency parsing head generates a vector of probabilities on {NEU, NEG, POS, NO-DEP} tags.

	room service	is	great
room service	NO-DEP	NO-DEP	**POS**
is	NO-DEP	NO-DEP	NO-DEP
great	**POS**	NO-DEP	NO-DEP

Expected Result

	0 room	1 service	2 is	3 great
aspect tag	**B**	**I**	O	O
opinion tag	O	O	O	**B**

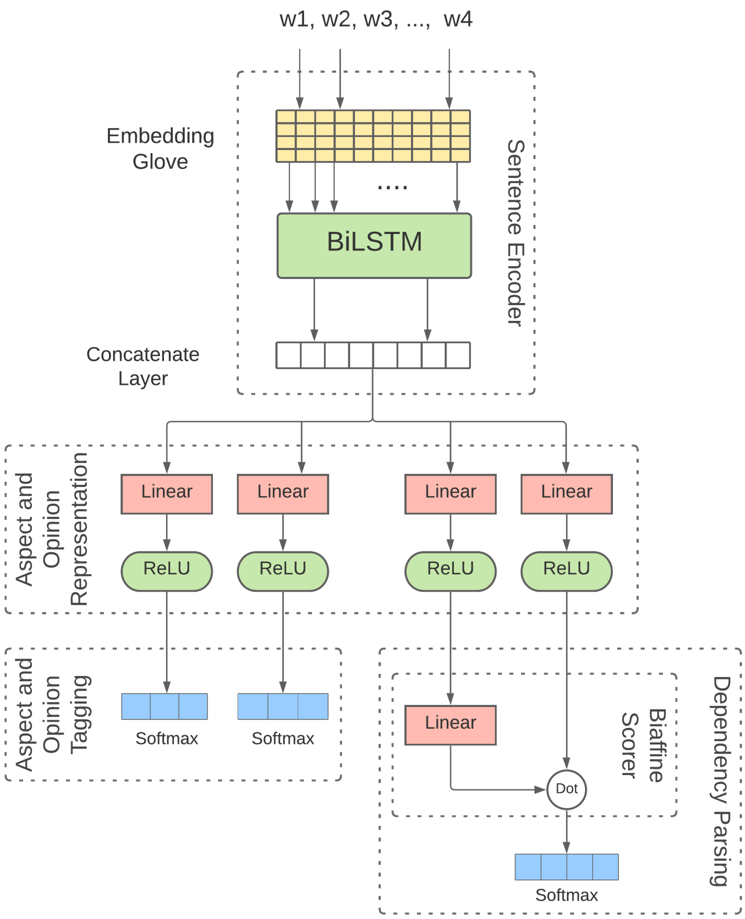
triplet {([0, 1], [3, 3], POS)}

* These considerations are for illustration. In practice, the relationship between the coefficients of the vectors is more diffuse due to the black box condition of the neural networks.

Source: Authors' own creation.

Our framework[8] (BERT for Opinion Triplet Extraction - BOTE) can handle One-to-One, One-to-Many and Many-to-Many relationships between aspects and opinions, as well as overlapped sentences. It consists of 4 modules. The overall architecture is shown in Figure 7. 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.

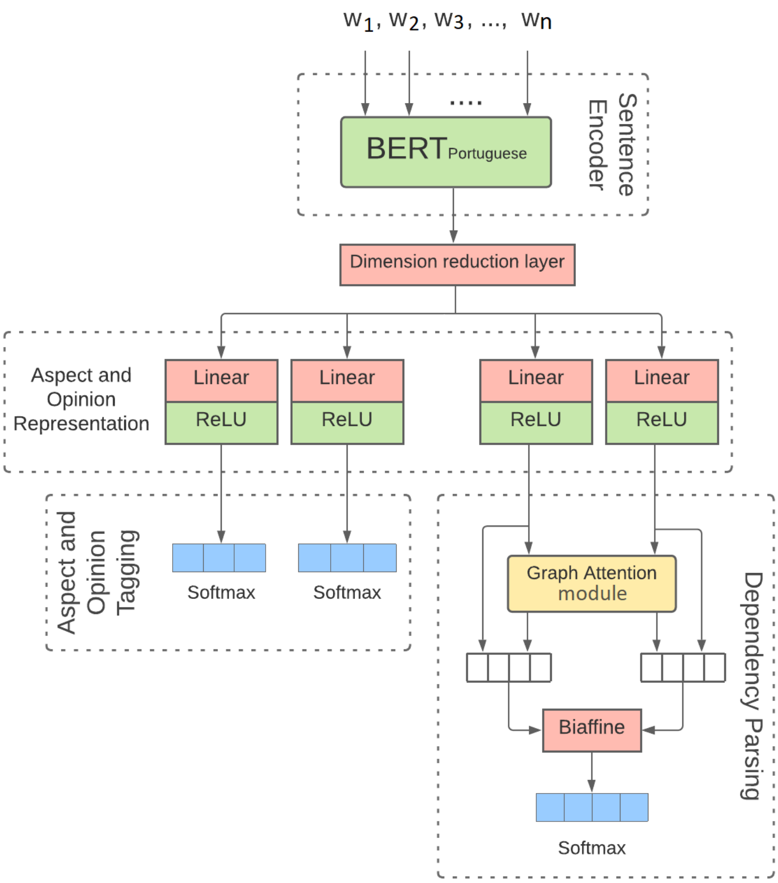
Figure 6: Reference model ZHANG et al., 2020. Interpretation of outputs can be found in Sections 4.3.3 and 4.3.4



Source: Authors' own creation.

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

Figure 7: Architectures of our model. Interpretation of outputs can be found in Sections 4.3.3 and 4.3.4



Source: Authors' own creation.

4.3.1 Sentence encoding

We adopt BERT (DEVLIN et al., 2019) to encode sentences. Given a sentence $S = \{w_i\}_{i=1}^{|S|}$ we get an word-vectors set $V = \{v_i | v_i \in \mathbb{R}^{d_B}\}_{i=1}^{|S|}$ from n^{th} BERT hidden state, where d_B denote the dimensionality of 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.

The BERT model works with WordPiece (DEVLIN et al., 2019), therefore, from the sentence S, we obtain token vectors t_i instead of word vectors v_i. A word can be formed by 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 is responsible for calculating 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 multi-collinearity and irrelevant features, making our model more straightforward, less data-hungry, and reducing the risk of overfitting.

4.3.2 Aspect and opinion representation

Following (ZHANG et al., 2020), 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 $\mathbf{b}_r^{(ap)}, \mathbf{b}_r^{(op)}, \mathbf{b}_r^{(ap)\prime}, \mathbf{b}_r^{(op)\prime} \in \mathbb{R}^{d_r}$:

$$\mathbf{r}_i^{(ap)} = g(\mathbf{W}_r^{(ap)} v_i + \mathbf{b}_r^{(ap)}) \tag{4.1}$$

$$\mathbf{r}_i^{(op)} = g(\mathbf{W}_r^{(op)} v_i + \mathbf{b}_r^{(op)}) \tag{4.2}$$

$$\mathbf{r}_i^{(ap)\prime} = g(\mathbf{W}_r^{(ap)\prime} v_i + \mathbf{b}_r^{(ap)\prime}) \tag{4.3}$$

$$\mathbf{r}_i^{(op)\prime} = g(\mathbf{W}_r^{(op)\prime} v_i + \mathbf{b}_r^{(op)\prime}) \tag{4.4}$$

4.3.3 Opinion and aspect tagging

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

$$\mathbf{p}_i^{(ap)} = softmax(\mathbf{W}_t^{(ap)} \mathbf{r}_i^{(ap)} + \mathbf{b}_t^{(ap)}) \tag{4.5}$$

$$\mathbf{p}_i^{(op)} = softmax(\mathbf{W}_t^{(op)} \mathbf{r}_i^{(op)} + \mathbf{b}_t^{(op)}) \tag{4.6}$$

Above, $\mathbf{W}_t^{(ap)}, \mathbf{W}_t^{(op)} \in \mathbb{R}^{3\times d_r}$ and $\mathbf{b}_t^{(ap)}, \mathbf{b}_t^{(op)} \in \mathbb{R}^3$ 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 Begining/Inside/Outside of an aspect/opinion term. For example, in the sentence *"Large rooms and great breakfast. Room service was awesome."*, after decoding [9] $\mathbf{p}_i^{(ap)}$ and $\mathbf{p}_i^{(op)}$ outputs, the result should be as in Table 2.

Table 2: Tagging result.

	0 Large	**1** rooms	**2** and	**3** great	**4** breakfast	**5** .	**6** Room	**7** service	**8** was	**9** 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

4.3.4 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, i.e., the pair (w_i, w_j) does not have the same relation as the pair (w_j, w_i). To avoid redundant relations, following Miwa e Sasaki (2014), Bekoulis et al. (2018) and Zhang et al. (2020), 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 Figure 8.

A Biaffine scorer (DOZAT; MANNING, 2016) is used to obtain word-level sentiment dependencies. This mechanism is widely used and has shown success in various parsing-related tasks (LI et al., 2019a; LI et al., 2019b). 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 Module (GAM). 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:

[9] understand by decoding apply argmax function to obtain the index with greater probability and then obtain the tag assigned to this index

$$s_{ijk} = [\mathbf{W}^{(k)} \mathbf{z}_i^{(ap)\prime} + \mathbf{b}^{(k)}]^\top \mathbf{z}_j^{(op)\prime} \tag{4.7}$$

$$\mathbf{z}_i^{(ap)\prime} = [\mathbf{r}_i^{(ap)\prime}; \mathbf{h}_i^{(ap)\prime}], \quad \mathbf{z_i} \in \mathbb{R}^{2d_r} \tag{4.8}$$

$$\mathbf{z}_j^{(op)\prime} = [\mathbf{r}_j^{(op)\prime}; \mathbf{h}_j^{(op)\prime}], \quad \mathbf{z_j} \in \mathbb{R}^{2d_r} \tag{4.9}$$

Figure 8: A parsing example for sentiment dependency.

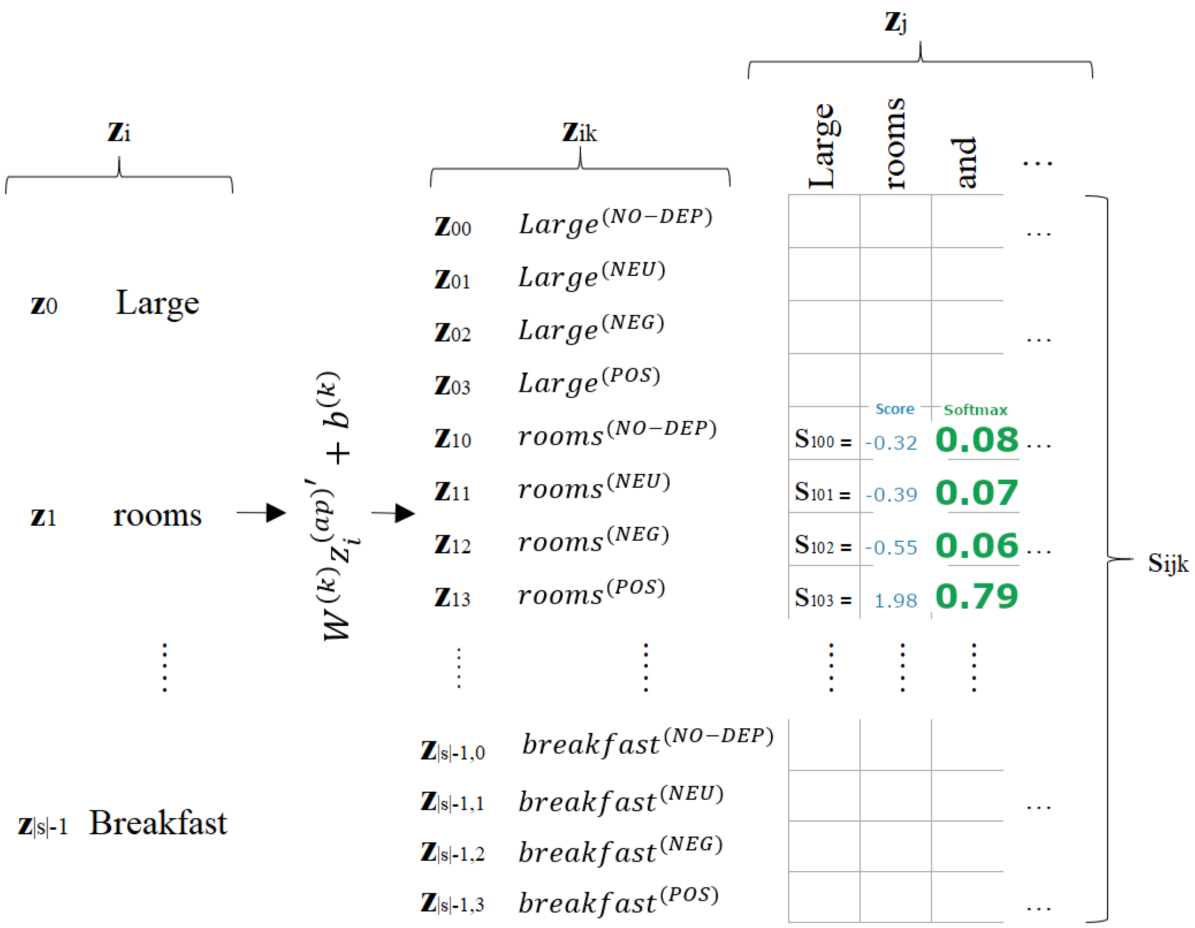
	Large	rooms	and	great	breakfast
Large	NO-DEP	NO-DEP	NO-DEP	NO-DEP	NO-DEP
rooms	POS	NO-DEP	NO-DEP	NO-DEP	NO-DEP
and	NO-DEP	NO-DEP	NO-DEP	NO-DEP	NO-DEP
great	NO-DEP	NO-DEP	NO-DEP	NO-DEP	NO-DEP
breakfast	NO-DEP	NO-DEP	NO-DEP	POS	NO-DEP

Source: Authors' own creation.

An example of how to calculate s_{ijk} is shown in Figure 9. The aspect-vector $\mathbf{z}_i$ representing the word w_i is inputted into the linear layer $\mathbf{W}^{(k)}(.) + \mathbf{b}^{(k)}$ (Equation 4.7). 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 k-dependency between w_i and w_j. In our example, the score vector of $(\mathbf{w}_1, \mathbf{w}_3)$ = ("rooms", "large") is $s_{13} = \langle s_{130}, s_{131}, s_{132}, s_{133} \rangle = \langle -0.32, -0.39, -0.55, 1.98 \rangle$. 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 "rooms" and "large" is positive.

Although BERT contains implicit syntactic information, its ability to capture explicit syntactic features is limited (GOLDBERG, 2019; LIU et al., 2020). 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 (LU; DU; NIE, 2020; WANG et al., 2020).

Figure 9: A biaffine scoring example for sentiment dependency.



Source: Authors' own creation.

4.3.5 Syntax-aware vectors $\mathbf{h}_i^{(ap)\prime}$ and $\mathbf{h}_j^{(op)\prime}$

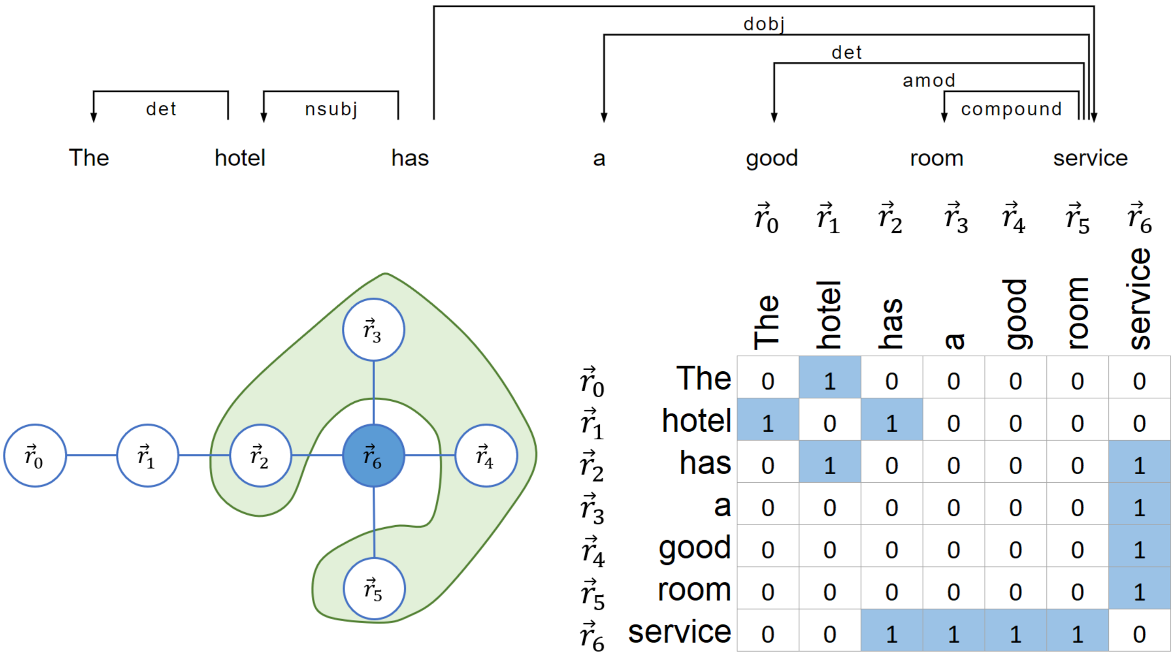
Unlike ZHANG et al., we concatenate a syntax-aware vector $\vec{\mathbf{h}}$ with $\vec{\mathbf{r}}$ before inputting the features vector into the Biaffine scorer. As mentioned earlier, this vector $\vec{\mathbf{h}}$ is obtained from a Graph Attention Module (GAM) consists of two Graph Attention Layer (GAL), which can to capture linguistic structures like dependency trees other than only sequential data (HUANG; CARLEY, 2019). The GAM receives the vector $\vec{\mathbf{r}}$ and an adjacency matrix that represents the syntactic dependency graph of the sentence[10]. GAM 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.

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

Figure 10 shows the syntactic dependency of “The hotel has a good 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 $\rightarrow$ opinion. If we use a directed graph, we could lose information.

Figure 10: Syntax-aware vector overview.



Source: Authors’ own creation.

4.4 Hyperparameter settings

For the BERT encoder, we use the pre-trained cased Portuguese BERT developed by NeuralMind (SOUZA; NOGUEIRA; LOTUFO, 2020), pre-trained cased Spanish BERT (CANETE et al., 2020) and uncased BERT multilingual developed by Google Research (DE-

VLIN et al., 2018). To compare our model with the baselines in English, we use uncased BERT base model developed by Google Research[11]. 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 10^{-3}, and weight decay (L2 penalty) is 10^{-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.

4.5 Performance comparison process

Initially, the adjacency matrices are generated for each sentence. These matrices represent the dependency graph. Groups are created for cross-validation. Given an X dataset, it is divided into four parts (4 folds). Two groups are used as a training set (50% of the data points). The other two groups are used as a validation set (25%) and a test set (25%) and vice versa. Then the training and validation-test groups are alternated. Therefore, from each dataset, we get four partitions.

Subsequently, the models (ours and baselines) are trained and tested in the four partitions of each dataset. The process is then repeated five times with different initializations on each partition. Each model in each dataset generates 20 performance measures for each metric used (Figure 11). These 20 measurements are used in the hypothesis tests to determine if our model exceeds the baselines.

Following related works (PENG et al., 2020; WU et al., 2020; ZHANG et al., 2020;) 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 5×2-nested cross-validation technique (DIETTERICH, 1998) 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 (DEMŠAR, 2006). 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

[11]https://github.com/google-research/bert

Friedman test to identify if there was a statistically significant difference between the models (DEMŠAR, 2006). Finally, we applied Hommel post-hoc procedure to identify if our model was better than the baselines (GARCíA; HERRERA, 2008). The significance level used in all tests was 0.05.

Figure 11: Cross-validation partition generation.

Model	Datasets		Partitions				Total runs per dataset
			Train	Val	Test	Runs	
Model X	Dataset_1	Splitted in 4 parts	P_{11}, P_{12}	P_{13}	P_{14}	5 times	20
			P_{11}, P_{12}	P_{14}	P_{13}	5 times	
			P_{13}, P_{14}	P_{11}	P_{12}	5 times	
			P_{13}, P_{14}	P_{12}	P_{11}	5 times	
	Dataset_2		P_{21}, P_{22}	P_{23}	P_{24}	5 times	20
			P_{21}, P_{22}	P_{24}	P_{23}	5 times	
			P_{23}, P_{24}	P_{21}	P_{22}	5 times	
			P_{23}, P_{24}	P_{22}	P_{21}	5 times	
	⋮		⋮	⋮	⋮	⋮	⋮
	Dataset_n		P_{n1}, P_{n2}	P_{n3}	P_{n4}	5 times	20
			P_{n1}, P_{n2}	P_{n4}	P_{n3}	5 times	
			P_{n3}, P_{n4}	P_{n1}	P_{n2}	5 times	
			P_{n3}, P_{n4}	P_{n2}	P_{n1}	5 times	

Source: Authors' own creation.

5 RESULTS AND ANALYSIS

The models implemented have been developed in Pytorch, using the Google Colab PRO platform with a Tesla P100-PCIE-16GB GPU. The BERT encoders used were imported through Hugging Face libraries.

5.1 Comparison with Baselines

The results are shown in Table 3. Our framework (monolingual and multilingual version) consistently achieves the best scores, both for the aspect and opinion extraction task and triplet extraction task in Portuguese and Spanish. The average difference in percentage points between our model and the Portuguese baselines in aspect extraction is 6.68 percentage points%, 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. In Spanish, our model also showed significantly higher performance; the average difference in triplet extraction was 15.64% compared to baselines. In triplet extraction, the monolingual BOTE models showed to be more competent than the multilingual ones.

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.

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

Task	Model	Portuguese ReLi	Portuguese ReHol	Spanish RestES	English Rest14	English Rest15	English Rest16	English Lap14
Aspect Ext.(AT)	OTE-MTL	71.48	78.70	84.21	76.11*	77.95*	72.95	71.99
	CMLA-MTL	72.14	74.31	83.40	72.80	73.55	71.22	68.72
	BOTE (ours)	**79.49**	**82.20**	87.04	76.07*	77.61*	**77.77**	72.29*
	BOTE-MLing (ours)	79.20	79.44	**87.54**	76.90	77.69*	76.93	71.44
Opinion Ext.(OT)	OTE-MTL	68.61	72.93	75.98	77.91*	74.27	76.19	65.62
	CMLA-MTL	62.16	69.45	72.97	74.46	74.72	74.69	63.82
	BOTE (ours)	**70.52**	**81.10**	81.19	77.98*	**77.69**	**78.88**	**68.66**
	BOTE-MLing (ours)	65.21	77.52	**81.73**	78.03*	76.50	78.66	67.47
Triplet Ext.(TR)	OTE-MTL	49.92	60.91	62.33	57.26*	52.46	53.01	39.67
	CMLA-MTL	40.61	47.03	47.03	45.57	39.70	42.45	32.05
	BOTE (ours)	**57.49**	**66.65**	**70.32**	56.49*	**55.12**	**55.79**	**44.04**
	BOTE-MLing (ours)	51.77	58.71	68.98	53.81	50.06	55.05	43.72

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

Model	Train: rest14, Test: lap14 AT	OT	TR	Train: reli, Test: rehol 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**

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

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
RestES	57.14	73.17	79.99	68.26

5.2 Ablation Study

In the ablation study, we remove some relevant components of the model to test their impact on the accuracy of the results. We examined the effectiveness of two components of our BOTE model, namely the reduction layer and Graph Attention Module (GAM). Table 6 presents the ablation results on Reli and ReHol datasets. BOTE-NoReduction does not use any 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 Module 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**

5.3 Error Analysis

We conducted an analysis of false positives (identified by the model but not existing in ground truth) and false negatives (not identified by the model but existing in ground truth) on ReLi, ReHol, RestES, Rest14, Rest15 and Rest16 datasets. Figure 12 shows the proportions of false positives and negatives as a function of the overlapped case and triplet terms. The proportion of false negatives is significantly higher than the negatives, regardless of language, overlapped case, or the implied triplet term. Although the model proves to be superior to baselines, there is an opportunity for improvement for the extraction precision of the triplets. A detailed analysis of the sentences shows that many errors correspond to aspects or opinions formed by multiple words, mainly those composed of a noun and an adjective (Table 7, index 1 and 2). The model manages to locate the aspect/opinion term but fails in the precision to select the range. As another aspect to consider, sentiment dependency for the sarcastic and metaphoric sentences is usually challenging due to the difference in its literal meaning and actual meaning (Table 7, index 3). However, this is a still-pending challenge to be solved in most of the NLP Deep Learning models (KAMATH et al., 2021).

Table 7: Sentence cases.

Index	Sentence	Ground truth	BOTE Output
1	Well seasoned latin food eat	[(latin food, Well seasoned, POS)]	[(food, Well seasoned, POS)]
2	Lounge is a good attraction	[(Lounge, good attraction, POS)]	[(lounge, good, POS)]
3	Good luck getting a table	[(getting a table, Good luck, NEG)]	[(table, good, POS)]
4	Great eats , good times	[(eats, great, POS)]	[(eats, great, POS),(times, good, POS)]
5	Late nite omelletes are good here	[(omelletes, good, POS)]	[(Late nite omelletes, good, POS)]

Figure 13 shows where false positives and negatives are concentrated. We can observe that in the datasets in English, the prediction of the sentiment of the triplet has the highest error rate. The model classifies sentiment dependence into four categories (negative, positive, neutral, non-dependence) while aspect/opinion term is classified into three (B, I, O). Given that the greater the number of classes, the greater the entropy, it is expected that more errors will occur in this item. In Portuguese and Spanish datasets, sentiment error rate has a different behavior. Table 8 shows the entropy as a function of the number of triplets grouped by sentiment. The sentiment categories are more unbalanced in Portuguese and Spanish than in English. The greater the imbalance, the lower the entropy, which increases the probability of success and reduces the sentiment error rate in these datasets.

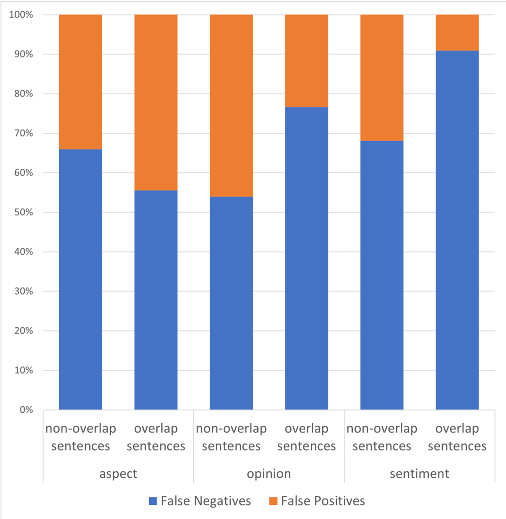
We can see that ReLi and RestES have similar entropies (0.212 and 0.181) and similar sentiment error percentages (even being datasets of different languages). RestES has a lower entropy and its percentage of false positives and negatives in sentiment prediction is also lower. This situation differs significantly from the ReHol dataset, whose entropy is 0.34. Despite these correlations, it is expected that other variables influence these results; entropy alone is not a sufficient condition to predict these distributions.

Despite the limitations, the model acquired sufficient generalizability to suggest better labels (Table 7, index 4 and 5). When we join the English datasets to create a larger one, the model has better-predicted compound words. Therefore, one way to improve performance in the short term would be to increase the amount and variety of training data. Even in resource-rich languages like English, the number of triplets does not exceed 4000 data points (a small number compared to datasets for binary sentiment classification, 50k - 100k data points).

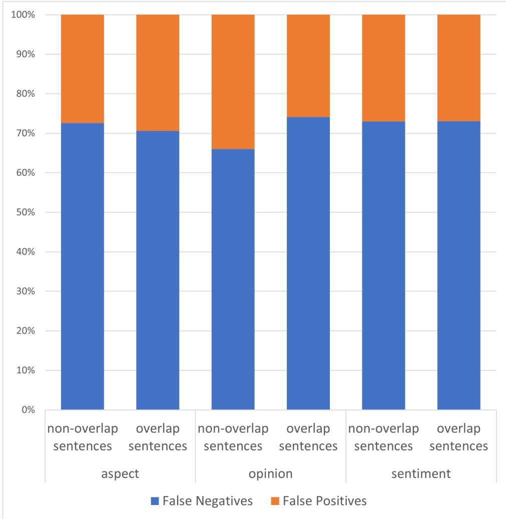
Table 8: Entropy as a function of the number of triplets .

Nomenclature	ReHol	ReLi	RestES	Rest14	Rest15	Rest16	Lap14
Language	PT	PT	ES	EN	EN	EN	EN
# Triplets	3199	1448	3085	3906	1747	2247	2349
# Positive triplets	1846	1170	2707	2867	1285	1674	1350
# Negative triplets	1242	278	334	753	401	483	774
# Neutral triplets	111	0	44	286	61	90	225
% Positive triplets	57.72	80.80	87.75	73.40	73.55	74.50	57.47
% Negative triplets	38.84	19.20	10.83	19.28	22.95	2150	32.95
% Neutral triplets	3.47	0.00	1.43	7.32	3.49	4.01	9.58
Entropy	0.348	0.212	0.181	0.32	0.296	0.295	0.395

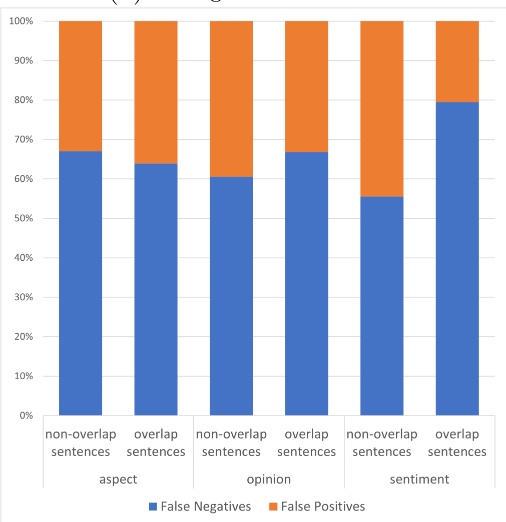
Figure 12: Proportion of false positives and false negatives.



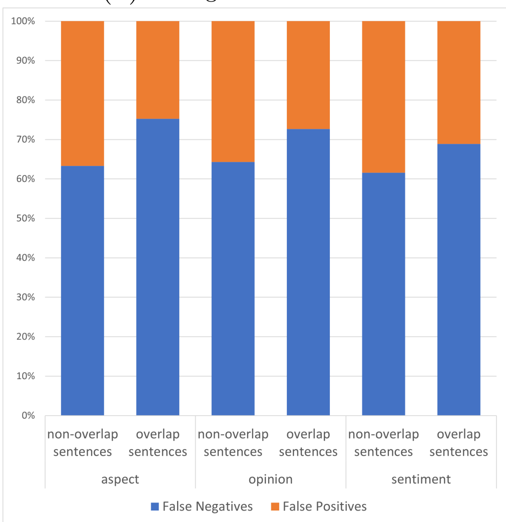
(a) Portuguese ReLi Dataset



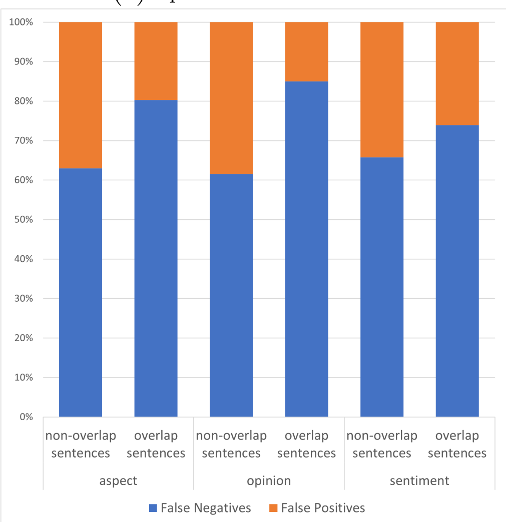
(b) Portuguese ReHol Dataset



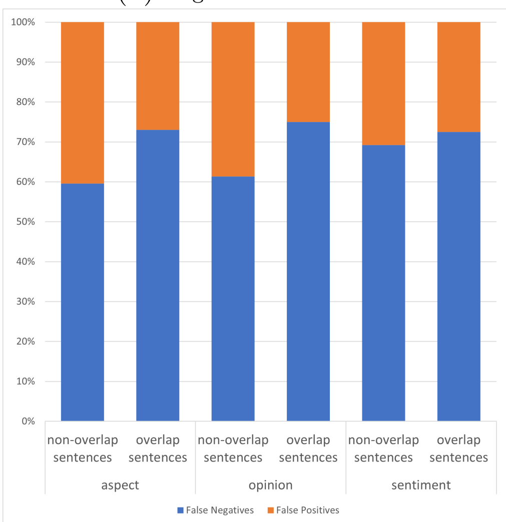
(c) Spanish RestES Dataset



(d) English Rest14 Dataset



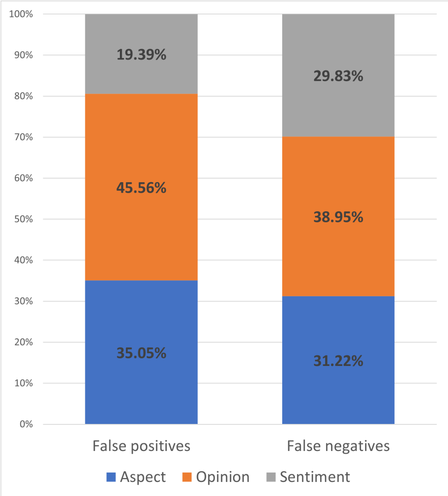
(e) English Rest15 Dataset



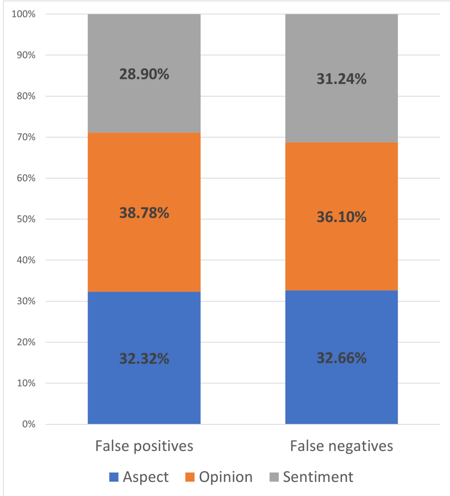
(f) English Rest16 Dataset

Source: Authors' own creation.

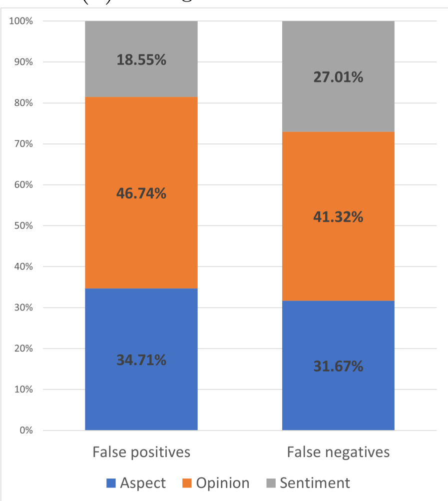
Figure 13: Distribution of false positives and negatives according to the triplet components.



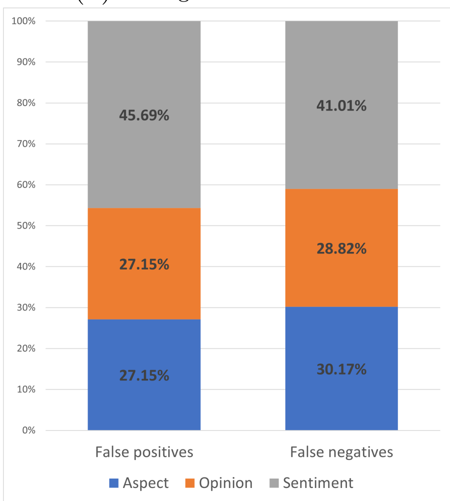
(a) Portuguese ReLi Dataset



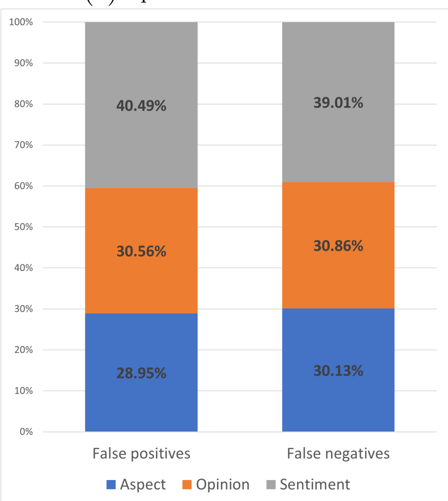
(b) Portuguese ReHol Dataset



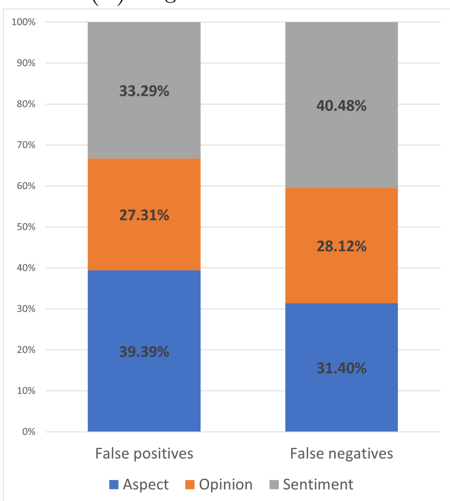
(c) Spanish RestES Dataset



(d) English Rest14 Dataset



(e) English Rest15 Dataset



(f) English Rest16 Dataset

Source: Authors' own creation.

6 CONCLUSIONS AND FUTURE WORK

In this work, we proposed a Machine Learning framework to deal with Aspect Sentiment Triplet Extraction in Portuguese and Spanish. Our model used a BERT model to obtain context-aware word vectors, a Graph Attention Module 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 and 70.32% in Spanish. 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. Error analysis showed that the distribution of false positives and negatives is influenced by entropy.

The experimental results verify the effectiveness of our framework compared to the baselines. To the the best of our knowledge, we have developed the most fine-grained model to deal with aspect sentiment tasks in Portuguese and Spanish. The proposed model even outperforms its English counterparts in some cases. The monolingual BOTE models showed to be more competent than the multilingual ones. 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.

REFERENCES

AIRES, J. P. et al. A deep learning approach to classify aspect-level sentiment using small datasets. In: *2018 International Joint Conference on Neural Networks, IJCNN 2018, Rio de Janeiro, Brazil, July 8-13, 2018*. IEEE, 2018. p. 1–8. Disponível em: ⟨https://doi.org/10.1109/IJCNN.2018.8489760⟩.

AKHTAR, M. S. et al. Language-agnostic model for aspect-based sentiment analysis. In: *Proceedings of the 13th International Conference on Computational Semantics-Long Papers*. [S.l.: s.n.], 2019. p. 154–164.

ANGEL, S. O.; NEGRÓN, A. P. P.; ESPINOZA-VALDEZ, A. Systematic literature review of sentiment analysis in the spanish language. *Data Technologies and Applications*, Emerald Publishing Limited, 2021.

ARAQUE, O. et al. Aspect based sentiment analysis of spanish tweets. In: *Proceedings of TASS 2015: Workshop on Sentiment Analysis at SEPLN co-located with 31st SEPLN Conference (SEPLN 2015), Alicante, Spain, September 15, 2015*. CEUR-WS.org, 2015. (CEUR Workshop Proceedings, v. 1397), p. 29–34. Disponível em: ⟨http://ceur-ws.org/Vol-1397/gsi.pdf⟩.

BALAGE, P. P. *Aspect extraction in sentiment analysis for portuguese language*. Tese (Doutorado) — Universidade de Sao Paulo, 2017.

Balaji, P.; Nagaraju, O.; Haritha, D. Levels of sentiment analysis and its challenges: A literature review. In: *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*. [S.l.: s.n.], 2017. p. 436–439.

BEKOULIS, G. et al. Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Systems with Applications*, Elsevier, v. 114, p. 34–45, 2018.

CANETE, J. et al. Spanish pre-trained bert model and evaluation data. In: *PML4DC at ICLR 2020*. [S.l.: s.n.], 2020.

CARDOSO, B.; PEREIRA, D. Evaluating an aspect extraction method for opinion mining in the portuguese language. In: SBC. *Anais do VIII Symposium on Knowledge Discovery, Mining and Learning*. [S.l.], 2020. p. 137–144.

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*. New York, NY, USA: Association for Computing Machinery, 2019. (EBIMCS '19). ISBN 9781450366496. Disponível em: ⟨https://doi.org/10.1145/3377817.3377834⟩.

DAGAR, A. et al. *Graph Representation Learning*. [S.l.]: GitHub, 2020. ⟨https://github.com/dsgiitr/graph_nets⟩.

DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, v. 7, n. 1, p. 1–30, 2006. Disponível em: ⟨http://jmlr.org/papers/v7/demsar06a.html⟩.

DEVLIN, J. et al. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. Disponível em: ⟨http://arxiv.org/abs/1810.04805⟩.

DEVLIN, J. et al. 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 4171–4186.

DIETTERICH, T. G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, MIT Press, v. 10, n. 7, p. 1895–1923, 1998.

DOZAT, T.; MANNING, C. D. Deep biaffine attention for neural dependency parsing. *CoRR*, abs/1611.01734, 2016. Disponível em: ⟨http://arxiv.org/abs/1611.01734⟩.

DUBIAU, L.; ALE, J. M. Análisis de sentimientos sobre un corpus en español: Experimentación con un caso de estudio. In: *XIV Argentine Symposium on Artificial Intelligence (ASAI)-JAIIO 42 (2013)*. [S.l.: s.n.], 2013.

FAN, Z. et al. 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: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 2509–2518. Disponível em: ⟨https://www.aclweb.org/anthology/N19-1259⟩.

FARIAS, D. S. et al. Opinion-meter: A framework for aspect-based sentiment analysis. In: *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web, Webmedia 2016, Teresina, Piauí State, Brazil, November 8-11, 2016*. ACM, 2016. p. 351–354. Disponível em: ⟨https://doi.org/10.1145/2976796.2988214⟩.

FREITAS, C. et al. Vampiro que brilha... rá! desafios na anotaçao de opiniao em um corpus de resenhas de livros. *Encontro de Linguística de Corpus*, v. 11, p. 22, 2012.

FREITAS, L. A. de; VIEIRA, R. Exploring resources for sentiment analysis in portuguese language. In: *2015 Brazilian Conference on Intelligent Systems, BRACIS 2015, Natal, Brazil, November 4-7, 2015*. IEEE Computer Society, 2015. p. 152–156. Disponível em: ⟨https://doi.org/10.1109/BRACIS.2015.52⟩.

GARCíA, S.; HERRERA, F. An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. *Journal of Machine Learning Research*, v. 9, n. 89, p. 2677–2694, 2008. Disponível em: ⟨http://jmlr.org/papers/v9/garcia08a.html⟩.

GOLDBERG, Y. Assessing bert's syntactic abilities. *CoRR*, abs/1901.05287, 2019. Disponível em: ⟨http://arxiv.org/abs/1901.05287⟩.

HAN, H. et al. Multi-attention network for aspect sentiment analysis. In: *Proceedings of the 2019 8th International Conference on Software and Computer Applications*. New York, NY, USA: Association for Computing Machinery, 2019. (ICSCA '19), p. 22–26. ISBN 9781450365734. Disponível em: ⟨https://doi.org/10.1145/3316615.3316673⟩.

HE, R. et al. 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: Association for Computational Linguistics, 2019. p. 504–515. Disponível em: ⟨https://www.aclweb.org/anthology/P19-1048⟩.

HENRIQUEZ, C.; SANCHEZ-TORRES, G. Aspect extraction for opinion mining with a semantic model. *Engineering Letters*, v. 29, n. 1, 2021.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019. p. 5469–5477. Disponível em: ⟨https://www.aclweb.org/anthology/D19-1549⟩.

KAMATH, A. et al. Sarcasm detection approaches survey. In: *Advances in Computer, Communication and Computational Sciences*. [S.l.]: Springer, 2021. p. 593–609.

LI, Y. et al. Self-attentive biaffine dependency parsing. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI*. [S.l.]: ijcai.org, 2019. p. 5067–5073.

LI, Z. et al. Dependency or span, end-to-end uniform semantic role labeling. *Proceedings of the AAAI Conference on Artificial Intelligence*, v. 33, n. 01, p. 6730–6737, Jul. 2019.

LIU, B. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, 2012. (Synthesis Lectures on Human Language Technologies). Disponível em: ⟨https://doi.org/10.2200/S00416ED1V01Y201204HLT016⟩.

LIU, L. et al. On the variance of the adaptive learning rate and beyond. In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. Disponível em: ⟨https://openreview.net/forum?id=rkgz2aEKDr⟩.

LU, Z.; DU, P.; NIE, J.-Y. Vgcn-bert: Augmenting bert with graph embedding for text classification. In: JOSE, J. M. et al. (Ed.). *Advances in Information Retrieval*. Cham: Springer International Publishing, 2020. p. 369–382.

MACHADO, M. T.; PARDO, T. A. S.; RUIZ, E. E. S. Analysis of unsupervised aspect term identification methods for portuguese reviews. *Anais do XIV Encontro Nacional de Inteligncia Artificial e Computacional (ENIAC), SBC*, p. 239–249, 2017.

MELENDEZ, J.; De Bona, G. A deep learning approach for aspect sentiment triplet extraction in portuguese. In: *10th Brazilian Conference on Intelligent Systems, BRACIS 2021 (to appear)*. [S.l.: s.n.], 2021.

MIRANDA, C. H.; BUELVAS, E. Aspectsa: Unsupervised system for aspect based sentiment analysis in spanish. *Prospectiva*, v. 17, n. 1, p. 87–95, 2019.

MIWA, M.; SASAKI, Y. Modeling joint entity and relation extraction with table representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 2014. p. 1858–1869. Disponível em: ⟨https://www.aclweb.org/anthology/D14-1200⟩.

PENG, H. et al. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, v. 34, n. 05, p. 8600–8607, Apr. 2020. Disponível em: ⟨https://ojs.aaai.org/index.php/AAAI/article/view/6383⟩.

PEREIRA, D. A. A survey of sentiment analysis in the portuguese language. *Artif. Intell. Rev.*, v. 54, n. 2, p. 1087–1115, 2021. Disponível em: ⟨https://doi.org/10.1007/s10462-020-09870-1⟩.

PONTIKI, M. et al. SemEval-2015 task 12: Aspect based sentiment analysis. In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Denver, Colorado: Association for Computational Linguistics, 2015. p. 486–495.

PONTIKI, M. et al. SemEval-2016 task 5: Aspect based sentiment analysis. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. San Diego, California: Association for Computational Linguistics, 2016. p. 19–30.

PONTIKI, M. et al. SemEval-2014 task 4: Aspect based sentiment analysis. In: *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Dublin, Ireland: Association for Computational Linguistics, 2014. p. 27–35.

SAIAS, J.; MOURÃO, M.; OLIVEIRA, E. Detailing sentiment analysis to consider entity aspects: an approach for portuguese short texts. *Transactions on Machine Learning and Artificial Intelligence*, Society for Science and Education, United Kingdom, v. 6, p. 26–35, 2018.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. de A. Bertimbau: Pretrained BERT models for brazilian portuguese. In: *Intelligent Systems - 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20-23, 2020, Proceedings, Part I*. Springer, 2020. (Lecture Notes in Computer Science, v. 12319), p. 403–417. Disponível em: ⟨https://doi.org/10.1007/978-3-030-61377-8_28⟩.

VARGAS, F. A.; PARDO, T. A. S. Hierarchical clustering of aspects for opinion mining: A corpus study. *Linguística de Corpus: Perspectivas; Finatto, MJB, Rebechi, RR, Sarmento, S., Bocorny, AEP, Eds*, p. 69–91, 2018.

VELICKOVIC, P. et al. Graph attention networks. In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. Disponível em: ⟨https://openreview.net/forum?id=rJXMpikCZ⟩.

WANG, K. et al. Relational graph attention network for aspect-based sentiment analysis. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020. p. 3229–3238. Disponível em: ⟨https://www.aclweb.org/anthology/2020.acl-main.295⟩.

WANG, W. et al. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In: SINGH, S. P.; MARKOVITCH, S. (Ed.). *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*. AAAI Press, 2017. p. 3316–3322. Disponível em: ⟨http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14441⟩.

WU, Z. et al. Grid tagging scheme for aspect-oriented fine-grained opinion extraction. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, 2020. p. 2576–2585. Disponível em: ⟨https://www.aclweb.org/anthology/2020.findings-emnlp.234⟩.

XU, L. et al. Aspect sentiment classification with aspect-specific opinion spans. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2020. p. 3561–3567. Disponível em: ⟨https://www.aclweb.org/anthology/2020.emnlp-main.288⟩.

ZHANG, C. et al. A multi-task learning framework for opinion triplet extraction. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, 2020. p. 819–828. Disponível em: ⟨https://www.aclweb.org/anthology/2020.findings-emnlp.72⟩.

ZHAO, H. et al. SpanMlt: A span-based multi-task learning framework for pair-wise aspect and opinion terms extraction. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020. p. 3239–3248. Disponível em: ⟨https://www.aclweb.org/anthology/2020.acl-main.296⟩.

ZHOU, J. et al. Deep learning for aspect-level sentiment classification: Survey, vision, and challenges. *IEEE Access*, v. 7, p. 78454–78483, 2019.