






# Mining Comparative Opinions in Portuguese: A Lexicon-based Approach

Daniel Kansaon   [ Universidade Federal de Minas Gerais | [daniel.kansaon@dcc.ufmg.br](mailto:daniel.kansaon@dcc.ufmg.br) ]

Michele A. Brandão  [ Instituto Federal de Minas Gerais | [michele.brandao@ifmg.edu.br](mailto:michele.brandao@ifmg.edu.br) ]

Julio C. S. Reis  [ Universidade Federal de Viçosa | [jreis@ufv.br](mailto:jreis@ufv.br) ]

Fabrcio Benevenuto  [ Universidade Federal de Minas Gerais (UFMG) | [fabrcio@dcc.ufmg.br](mailto:fabrcio@dcc.ufmg.br) ]

 Federal University of Minas Gerais, Institute of Exact Sciences, Department of Computer Science. 31270010 - Belo Horizonte, MG - Brazil

Received: 01 August 2022 • Accepted: 26 March 2024 • Published: 26 September 2024

**Abstract** The constant expansion of e-commerce, recently boosted due to the coronavirus pandemic, has led to a massive increase in online shopping, made by increasingly demanding customers, who seek comments and reviews on the Web to assist in decision-making regarding the purchase of products. In these reviews, part of the opinions found are comparisons, which contrast aspects expressing a preference for an object over others. However, this information is neglected by traditional sentiment analysis techniques and it is not applicable for comparisons, since they do not directly express positive or negative sentiment. In this context, despite efforts in the English language, almost no studies have been done to develop appropriate solutions that allow the analysis of comparisons in the Portuguese language. This work presented one of the first studies on comparative opinion in Portuguese. Four main contributions are (1) A hierarchical approach for detecting comparative opinions, which consists of an initial binary step, which subdivides the regular opinions from the comparatives, to further categorize the comparatives into the five opinion groups: (1) Non-Comparative; (2) Non-Equal Gradable; (3) Equative, (4) Superlative; and (5) Non-Gradable. The results are promising, reaching 87% of Macro-F1 and 0.94 of AUC (Compute Area Under the Curve) for the binary step, and 61% of Macro-F1 in multiple classes; (2) An lexicon algorithm to detect the entity expressed as preferred in comparative sentences, reaching 94% of Macro-F1 for Superlative; (3) Two new datasets with approximately 5,000 comparative and non-comparative sentences in Portuguese; and (4) a lexicon with words and expressions frequently used to make comparisons in the Portuguese language.

**Keywords:** Opinion Mining, Sentiment Analysis, Comparative Opinions, Preference Detection

## 1 Introduction

The number of global digital buyers has been increasing, and the trend is that these numbers will continue to increase due to the boost given by the Covid-19 pandemic, which has forced people to adapt to this new online ecosystem [Berthene, 2022]. The main advantage of E-commerce is to reach a large number of people in different places despite distance and time [Nasti *et al.*, 2020]. All this online interaction in purchases, sales, and reviews generates a large amount of information, which is used by increasingly demanding customers to make decisions through reviews of opinions and reviews made on forums, blogs, and even on Online Social Networks (OSN) [Pitman, 2022].

In this context, opinions and reviews about a product can be divided into two types: (i) *regular opinions*, which are direct or indirect, criticizing or highlighting positive points of different aspects; and (2) *comparative opinions*, which contrast aspects of a given product with the same aspects of its competitors [Liu, 2012]. While regular opinions express a sentiment about a brand or a product, comparatives express a common form of evaluation indicating a contrast or similarity between different products. This ability to make comparisons expressing order and preference is a basic component

of human cognition, which is reflected in natural language through comparative sentences, a direct and efficient way of contrasting objects showing preferences [Sapir, 1944].

A large number of efforts in the literature focus on the application of sentiment analysis techniques to classify opinions into positive, negative and neutral, using lexical approaches [Taboada *et al.*, 2011; Ribeiro *et al.*, 2016; Araujo *et al.*, 2016; Melo *et al.*, 2019] and different supervised machine learning techniques [Bespalov *et al.*, 2011; Wang *et al.*, 2016; de O. Carosia *et al.*, 2019; Mehta *et al.*, 2020; Zhang *et al.*, 2018; Basiri *et al.*, 2021; Pathak *et al.*, 2021]. However, for mining comparative opinions, these traditional sentiment analysis techniques are not enough. For example, in the comparative sentence: “Smartphone X is *better* than Y”, the polarity is insufficient for a deeper analysis that aims to extract additional information, such as which products are compared and even which object is preferred. This information can be extracted from comparison and is extremely valuable not only for users, but also for companies looking to understand users’ views of their brand, product, and also their competitors.

Thus, in this field, one of the fundamental efforts to analyze and extract useful information from the comparisons is the creation of a mechanism to detect which sentences from a

set of revisions can be classified as comparative, distinguishing them from non-comparative sentences. This task is essential because the correct classification of sentences allows directing efforts to apply techniques suitable for each type of opinion, which is important, for instance, for a product recommendation, effective marketing plan generation, and reputation management.

Given the importance and applicability of the task of identifying expressions containing comparisons between distinct products considering different aspects, several emerging studies aim to propose techniques to solve the problem. Overall, these techniques are language-dependent, and although efforts are directed toward specific languages such as Arabic [El-Halees, 2012; Eldefrawi *et al.*, 2019], Chinese [Huang *et al.*, 2008], Vietnamese [Bach *et al.*, 2015] e Korean [Yang and Ko, 2009, 2011], to the best of our knowledge, is no effort to build an approach to detect comparative expressions in Portuguese, which is among the 10 most spoken languages in the world [Souza *et al.*, 2017]. Further, Brazil is the largest Portuguese-speaking country in the world and represents a vast e-commerce market that requires specific solutions for this context [Pompeo, 2022].

Therefore, this work aims to fill this gap, presenting a framework for studying comparative sentences in Portuguese, that ranges from detecting regular and comparative opinions to extracting preferences expressed in comparative sentences. Our primary objective can be segmented into two sub-goals. Initially, when presented with a collection of reviews, our emphasis is on recognizing comparative opinions. Subsequently, within the comparative opinions, we detect which entity expresses itself as the preferred one.

Our work is one of the first studies in Portuguese and proposes a novel automatic strategy based on machine learning algorithms for the detection of comparative sentences, categorizing them into five classes: (1) Non-Comparative; (2) Non-Equal Gradable; (3) Equative; (4) Superlative; and (5) Non-Gradative. The obtained results are promising, reaching Macro-F1 of up to 87%. Despite the challenges in the Portuguese language and the similarities in the comparative opinions, it is possible to find most of the comparisons, which allows a detailed analysis of preferences in the next steps.

Since we differentiate regular opinions from comparative ones, we turn our focus to extracting valuable information from them. In this direction, in this work, we propose an algorithm for analyzing preferences expressed in a comparative opinion by extracting the preferred entity. Our strategy achieved Macro-F1 of up to 84% and 95% for Non-Equal Gradable and Superlatives, respectively.

Finally, the main contributions of this work are the following: (1) A hierarchical approach to detecting comparative opinions, categorizing opinions into five comparative types, reaching up to 87% in terms of Macro-F1; (2) An lexicon algorithm to analyze preferences in comparisons, detecting the preferred entity in a comparison, reaching up to 84% Macro-F1; (3) Two new datasets with approximately 5,000 comparative and non-comparative sentences in Portuguese; and (4) a lexicon with words and expressions frequently used for making comparisons in the Portuguese language.

This paper is organized as follows: First, we start by presenting related efforts and prior knowledge about opinions

(Section 2 and 3). Next, we present the methodology adopted to collect and process the data (Section 4 and 5). Then, the strategy to automatically detect comparative opinions is discussed (Section 6), and an algorithm to detect the preferred entity in a sentence is proposed (Section 7) and evaluated (Section 8.1). Last, we present the conclusion and future work (Section 11).

## 2 Related Work

This section provides a description of related efforts along two main dimensions explored in this work, i.e., sentence classification (Section 2.1) and preference detection (Section 2.2).

### 2.1 Sentence Classification

There is a growing number of efforts that explore sentiment analysis in order to understand some phenomenon [Liu, 2012], or yet, that aim to propose more robust strategies to capture the sentiment associated with an input text, focusing on aspects [D'Addio *et al.*, 2017; de Melo *et al.*, 2018; Trisna and Jie, 2022], comparisons [Liu, 2012; Eldefrawi *et al.*, 2019], or even sentiment classification [Alaei *et al.*, 2019; Singh *et al.*, 2021] or summarization techniques [Asevedo Nóbrega and Salgueiro Pardo, 2018; Xu *et al.*, 2022]. Regarding strategies, there are three levels of granularity to classify feelings and/or emotions: (1) document level; (2) sentences; and (3) based on aspects and entities [Ganapathibhotla and Liu, 2008]. In this work, our focus is on the second direction, i.e., identifying comparative sentences and detecting their preferences. Specifically, we are interested in the comparative sentences of the opinions which express the similarities and differences among several entities [de Barros, 2019; Eldefrawi *et al.*, 2019; Serrano-Guerrero *et al.*, 2015; Younis *et al.*, 2020; Wei *et al.*, 2022]

The number of studies dealing with the automatic mining of comparative sentences is small, in Portuguese, there are almost no related studies. Portuguese is the second most common language on Twitter and is among the ten most spoken languages in the world [Souza *et al.*, 2017]. Souza *et al.* [2017] performed a systematic review of text mining in Portuguese and observed that most studies focus on text classification, which reinforces the gap in the studies of comparative sentences. In this context, lexical approaches have been used by several solutions in different languages [Jindal and Liu, 2006a,b; Yang and Ko, 2009; El-Halees, 2012]. Despite the limitations of lexical approaches, the study made by Jindal and Liu [2006a] shows that most comparative sentences use a group of comparative words to express comparisons, which means that lexical approaches may be able to capture most existing comparisons [Jindal and Liu, 2006a,b].

Thus, in this work, we create a lexicon of comparative words to find comparative sentences in Portuguese. In this context, to the best of our knowledge, our effort is one of the first studies in Portuguese, and it complements previous studies that explore other languages [Jindal and Liu, 2006a,b]. In addition, this work complements the study with the use of a

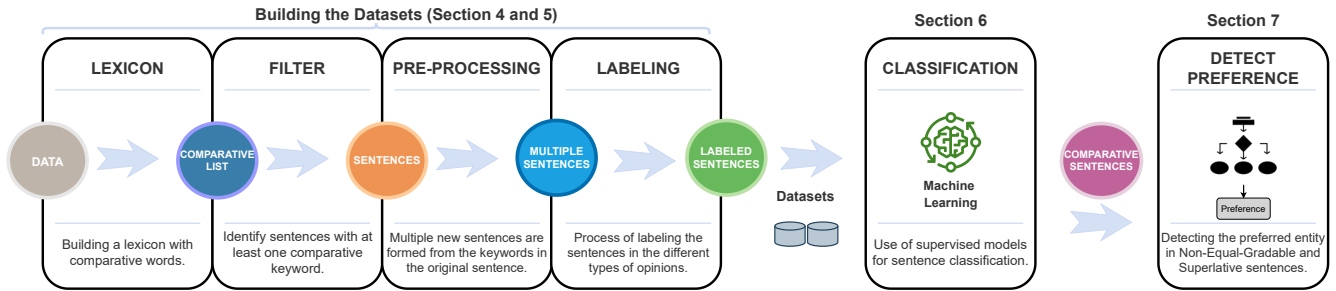


Figure 1. General steps of the methodology.

novel hierarchical machine learning strategy to classify sentences in their five different kinds of opinions.

## 2.2 Preference Detection

Some studies aim to develop techniques to analyze comparative sentences to extract their valuable information, such as detecting the preferred entity task [Ganapathibhotla and Liu, 2008; Ding *et al.*, 2008, 2009; Xu *et al.*, 2011; Eldefrawi *et al.*, 2019], opinion summarization [Kim and Zhai, 2009; Paul *et al.*, 2010; Ren and de Rijke, 2015; Ibeke *et al.*, 2017; Asevedo Nóbrega and Salgueiro Pardo, 2018; Nilashi *et al.*, 2022; Tsai *et al.*, 2020], and satisfactory analysis [Haque *et al.*, 2018; Dadhich and Thankachan, 2022].

The traditional task of determining whether a sentence has a positive or negative sentiment without indicating which entity is associated is uninformative [Liu, 2012]. Thus, the detection of the preferred entity is one of the main analyses that allow us to understand which entity is superior or inferior. Through textual characteristics and the comparative keyword used in the sentence, it is possible to determine which sentiment is associated with each entity and determine what is the preferred entity.

Preferred entity detection studies are usually carried out in specific and language-restricted solutions, such as the construction of lexicons and the treatment of language peculiarities [Ganapathibhotla and Liu, 2008; Ding *et al.*, 2008, 2009; Eldefrawi *et al.*, 2019]. In this context, we propose a lexical algorithm that extracts textual information in the sentence to determine the preferred entity in Non-Equal Gradable and Superlative sentences. The proposed approach is complementary to other works, dealing with the characteristics and peculiarities of the Portuguese language.

This work adds four new efforts to our previous work [Kansoon *et al.*, 2020], they are: (1) A validation of the recall and precision of the proposed lexical approach; (2) An important labeling process for words in the lexicon list, which may indicate superiority or inferiority; (3) We proposed a new lexicon algorithm that extracts the preferred entity in comparison; and (4) Finally, we evaluate the algorithm, discussing the impact of each step and its limitations.

## 3 Background Knowledge

Here, we define some concepts covered in this work about opinions that are commonly divided into regular and comparative.

**Regular opinion.** Regular opinion expresses a sentiment for an aspect of a particular entity and can be subdivided into two types [Jindal and Liu, 2006a,b]:

1. **Direct:** These opinions express a direct sentiment towards an entity or aspect, for example, “The battery of this smartphone is excellent.”
2. **Indirect:** They express an opinion about entities indirectly, usually manifested by the consequence or effect of some entities. In the sentence “After a long time driving this new car, I started to have back pain”, the criticism made to the car and the negative feeling related to the seat is indirectly demonstrated, making these opinions more complex for analysis.

**Comparative opinions.** The comparative opinions can be subdivided into four different groups [Liu, 2012].

1. **Non-Equal Gradable:** Relations of the type greater or less, expressing ordering and preference of an objects. “The car X is *better* than car Y”.
2. **Equative:** Two objects with relations of the type equal with respect to some features. “The smartphone camera X is *equals* Y”.
3. **Superlative:** An entity is greater or smaller than all others, rank one object over all others. “This is the *best* laptop in the world”.
4. **Non-Gradable:** Compares two or more entities, but expresses neither order nor preference for any one. “Laptop X design has some different features than laptop Y”.

This work explores all these comparative types, presenting a strategy to detect comparative sentences and categorize them into five different groups, in order to extract preference information in the next step.

**Lexicon approaches.** This work uses a lexicon with calculated weights, which can be sentiment (e.g., positive, negative, or neutral), word orientation (e.g., increase, decrease), or detailed word information [Buyya *et al.*, 2016]. The lexicon can be generated manually or automatically and is easily applicable to detect sentiment in a text, for example. For comparative opinions, a lexicon-based approach or a hybrid approach (i.e. uses a lexicon aggregated with another strategy such as machine learning) is very commonly used to find comparative opinions [Jindal and Liu, 2006b,a; Eldefrawi *et al.*, 2019] or even understand preferences in a sentence [Ganapathibhotla and Liu, 2008; Ding *et al.*, 2008, 2009].

## 4 Methodology

This section presents the main steps of the methodology proposed in this work, as shown in Figure 1. The three main steps of this methodology are:

**Building the datasets.** While comparative opinions are frequently employed to indicate preferences and similarities between elements [Liu, 2012], most of the content found in online reviews and discussion forums consists of regular opinions. Therefore, a strategy is necessary to find comparative sentences to build the datasets. Otherwise, if the comments are crawled indiscriminately many non-comparative opinions will be found, which could hinder the mining process. Section 5 describes this strategy. First, we create a lexicon with frequently used keywords for comparison purposes. Then, we validate and utilize this lexicon to filter opinions containing these keywords. Finally, we extract relevant sentences from all opinions to handle multiple comparisons.

**Classifying the comparative sentences.** In this step, we describe a supervised approach for hierarchical classification of comparative sentences, which has been divided into two steps: (1) binary classification, which separates opinions into comparative and regular; and (2) multiclass, which takes previously classified comparatives and breaks them down into four specific types of comparisons.

**Detecting the preferences.** After dividing the opinions into their types, we propose a lexicon strategy that uses textual information and word dependencies to extract the preferred entity in the comparisons, one of the most valuable information in comparison.

Our strategies were applied and evaluated in the two labeled datasets built in this work, in which metrics commonly used in machine learning and information retrieval tasks were used to evaluate the performance of proposed approaches (e.g., Precision, Accuracy, Recall, F1-Score, and Macro-F1) [Baeza-Yates et al., 1999].

## 5 Building Datasets

This section describes the construction of a lexicon in Portuguese (Section 5.1) used to build two comparative datasets presented in Sections 5.2 to 5.4), respectively.

### 5.1 Lexicon Approach

When analyzing comparative sentences, we detected a group of words commonly used to express most of the comparative opinions. This set of words is capable of covering most comparisons made in Portuguese.

From comparative keywords found in English [Jindal and Liu, 2006a] and some analysis to find out how comparative sentences are constructed in Portuguese, a list of words commonly used to make comparisons was built. The list was expanded by incorporating additional synonyms, encompassing verbs such as “win,” “overcome,” and “recommend,” adverbs like “more” and “less,” adjectives including “best,” “worst,” and “similar,” as well as common expressions in Portuguese such as “first one” and “not far behind.” In total, the list now comprises 59 words.

Although the comparative list has the main words used to express comparison, there may be words not considered because they are used only in specific contexts. Here, we analyze two important contexts existing in the online environment, which are: (1) review/evaluation websites; and (2) Online Social Networks (OSN). Thus, the initial lexicon is expanded through the inclusion of new comparative words found in these contexts.

For the context of online reviews, we chose Buscapé<sup>1</sup>, a prominent Brazilian platform for product and price searches in online stores. To enrich our lexicon with additional words and comparative expressions, a manual reading of some reviews was conducted. In total, we identified 107 new comparative keywords.

In addition, Twitter is the platform chosen for the study of comparisons on Online Social Networks (OSN) as it is one of the main opinion networks, composed largely of texts that express opinions about brands, and products and also comments on various subjects, and ranks among the six most used Online Social Networks in Brazil. Our analysis of tweets led to the discovery of 10 new comparative keywords not present in Buscapé. Consequently, we formed a lexicon with 176 comparative words<sup>2</sup>.

### 5.2 Lexical Filter

We built two datasets, one for each context. Regarding Buscapé, we use a large corpus with 85,910 Portuguese reviews collected in September 2013 Hartmann et al. [2014]. The dataset contains reviews of 230 different products, such as electronics, cars, bikes, guitars, etc. From this dataset, 48,311 reviews were found that present at least one keyword or comparative expression.

For Twitter, we conducted a comprehensive crawl of all Portuguese-language tweets posted by users on a specific day (2018-01-10), totaling 759.111 raw tweets encompassing both comparative and non-comparative opinions. After that, we used the lexicon list to filter all tweets that have at least one comparative keyword, the remaining 130.459 tweets. Both datasets were created and used in the work Kansaon et al. [2020].

**Lexicon Validation.** Applying a lexical filter is an interesting solution to find comparative opinions in a bunch of opinions. However, it is important to note the lexicon’s ability to find the comparisons. The decision to use a keyword strategy to find comparative sentences depends on evaluating whether this built set of words is sufficient to capture the expressive majority of the comparisons.

Thus, validating the capacity of the constructed lexicon on a sample of the data, we found a recall rate of 91.5% for Buscapé and 90% for Twitter, along with a precision rate of 37.5% for Buscapé and 24.3% for Twitter.

This means that this approach can find the most comparative sentences. Although not all sentences with comparative words are comparative, the lexicon can capture more than 90% of comparisons, that is, it has high recall and low precision, which can be improved with supervised techniques as further presented in Section 6.

<sup>1</sup>Buscapé Website: <https://www.buscape.com.br/>

<sup>2</sup>Lexicon: <http://doi.org/10.5281/zenodo.4124410>

### 5.3 Pre-Processing

The study of comparisons at the sentence level requires the extraction of all sentences for each crawled review. These sentences contain important characteristics that are relevant to the analysis and processing of existing comparisons.

1. **The compared entity is not directly specified in the sentence.** Sometimes, the sentence does not contain the compared entity because: (1) Before making an online review, users select which product they will evaluate. Hence, the users do not put the evaluated product to avoid redundancy or just use a relative pronoun. For example, “is better than Smartphone Y”, this sentence does not specify which product is better than Smartphone Y; (2) In the Portuguese language, some sentences can have a hidden subject, i.e., a subject is not present in the written sentence. For example, “The best series”.

**Proposed Solution:** Even if there is no comparative entity, the missing entity can be inferred from the context.

2. **Multiple Comparisons.** There are sentences with multiple comparisons, for example “The TV is incredible, worth buying it, the price is inferior to the Samsung, but it is superior all others.”, which has two distinct comparisons: (1) the price is inferior to the Samsung, and (2) is superior all others.

**Proposed Solution:** Rather than considering the complete sentence as a structure, it should be analyzed as a structure with several parts, which may or may not be comparative.

By focusing on the keyword used for comparison, it becomes possible to identify the focal point in which each comparison is conducted readily. Therefore, rather than attempting to extract information from a sentence containing multiple comparisons, we can instead split it down into simpler subsentences that contain only one comparison.

Therefore, for each comparative keyword within the sentence, a range of three words before and after the keyword is extracted to ensure that each sentence contains only a single comparison. Some studies have indicated that the next three words often encapsulate the most relevant information in the comparison Jindal and Liu [2006b,a]. This process generates new sentences. In the example in Figure 2, a new sentence is created for each comparative word. By dividing a complex sentence with multiple comparisons into simpler sentences with only one comparison, the task becomes more manageable. The choice of three as the interval size is significant since it captures the most relevant aspects in proximity to the comparison. Even if this range occasionally includes a word from a subsequent sentence, it does not impact the analysis as the range is small enough to avoid distorting or altering the original meaning of the comparison.

### 5.4 Labeling Process

After processing all reviews and obtaining the multiple comparisons, two datasets were built<sup>3</sup>. Labeling all these sen-

<sup>3</sup>Datasets: <http://doi.org/10.5281/zenodo.4124410>

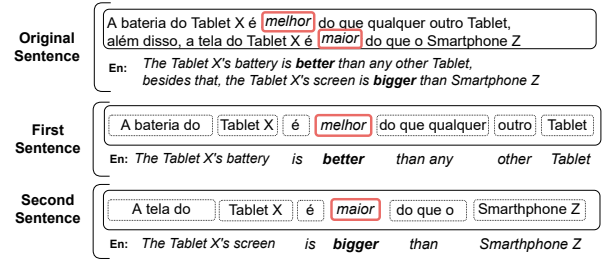


Figure 2. Strategy used to extract comparisons in a sentence (En: translated to English).

Table 1. Labeled sentences in each dataset.

Sentences	Buscapé	Twitter	Total
Comparatives	1,282	918	2,200
Non-Comparatives	1,472	1,135	2,607
<b>Total</b>	<b>2,754</b>	<b>2,053</b>	<b>4,807</b>

tences still takes effort due to the amount of data available. Thus, a sample was created keeping the original distribution. These sentences were manually labeled by a group of three volunteers who indicated whether the sentences were comparative. In addition, the entities, aspects, and preferences were identified. The Fleiss Kappa Cohen [1960] coefficient was calculated and the labelers’ agreement was 88.09% ( $\pm 0.007$ ) for Buscape and 87.73% ( $\pm 0.009$ ) for Twitter.

In total, 2,754 sentences were labeled in Buscapé, in which 1,282 comparative and 1,472 non-comparative sentences were found. For Twitter, 2,053 sentences were labeled, with 918 comparative and 1,135 non-comparative. Table 1 shows the number of labeled comparisons in each dataset.

## 6 Classification Strategy

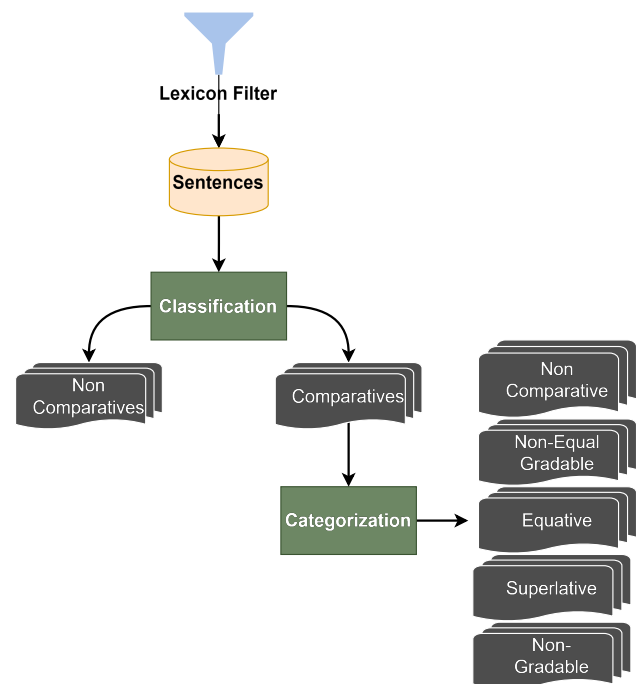


Figure 3. Hierarchical approach to sentence classification.

This section presents the supervised approach for the automatic classification of comparative sentences, which has been divided into two steps, as shown in Figure 3. First, after

the lexical filter is used to find likely comparisons, an approach to binary classification is presented, separating comparative from non-comparative (Section 6.1). Then, a classification strategy is applied to categorize these comparative sentences into five groups, which represent each type of opinion (Section 6.2).

## 6.1 Detecting Comparative Opinions

One of the fundamental steps in opinion mining is to separate comparatives from non-comparatives. The division of opinions is the most practical and important task, as it allows the application of classification techniques and the extraction of more detailed information about the comparisons.

The performance of four state-of-the-art classifiers was analyzed for the classification of binary sentences: Multinomial Naive Bayes (NB) [McCallum *et al.*, 1998], Support Vector Machine with the Radial Basis Function (SVM) [Joachims, 1998], Logistic Regression (LR) [Kleinbaum *et al.*, 2002], and Random Forest (RF) [Breiman, 2001]. Our emphasis is on using statistical models because of their ability to deal effectively with high-dimensional data, as well as their suitability for text classification problems, which often exhibit linear separability. Deep learning techniques were not explored, as they need a large set of labeled data for training [Koppe *et al.*, 2021].

The four algorithms were applied with a combination of three textual features<sup>4</sup>: (1) Tf-idf of words, (2) Tf-idf of bigrams of words, (3) Tf-idf of trigrams of words. Table 2 shows the results obtained for the experiments performed, which were replicated 35 times to allow the calculation and reporting of the confidence interval (95%), through 5-fold cross-validation.

For both datasets, the probabilistic algorithm NB showed the best results in terms of Macro-F1 (i.e., 87.3% and 86.1% for Buscapé and Twitter, respectively). The SVM (i.e., 87.1% for Buscapé and 84.7% for Twitter) presented statistically similar values, considering Macro-F1. However, the probabilistic model is superior, since it has a superior evocation for the comparative class. Jindal and Liu [2006b] applied a CRF strategy to detect gradable comparison sentences, achieving 81% Macro-F1.

When assessing the methodology, the metric of recall holds great importance as it reveals the frequency at which the model accurately detects examples from each class. In the hierarchical approach, the initial classifier identifies comparative sentences and subsequently categorizes them into different types. Thus, optimizing the recall metric for comparative sentences becomes crucial.

For the Buscapé dataset, although the NB has a similar Macro-F1 value to SVM, the first has a recall rate of 90.3% for the comparative class, well above the other models, which have values close to 80%. The same occurs in the Twitter database, where NB has the best recall rate, with 87.4%.

Table 3 shows the confusion matrix obtained with the NB, which presented the best results to distinguish the two classes. Note that this model can correctly detect 90.3% of the comparative sentences existing on Buscapé and 87.4% on Twitter,

<sup>4</sup>Others word representations were tested, such as embeddings, but Tf-idf was chosen because it shows the best results.

which highlights the model's good ability to cover comparisons. Finally, the best results were obtained with the NB, with an AUC of 0.94 for both datasets. Observing the ROC curve in Figure 4, one can see the possibility of choosing a classification threshold to correctly detect almost 90% of all comparisons with only 10% classification error (rate of classification false positive rating  $\approx 0.1$ ).

## 6.2 Categorization into Multiples Classes

After the binary detection of the sentences, the classification of the results was started to categorize the sentences previously classified as comparative into five groups: (1) Non-Comparative; (2) Non-Equal Gradable; (3) Equative; (4) Superlative; and (5) Non-Gradable. The categorization plays an important role in detailing the comparisons found, facilitating the visualization of information, and enabling more systematic analyses.

The sentences classified as comparative in the previous step with the NB were added to a new dataset, which has about 88% of the comparative sentences initially labeled. Table 1 shows the number of sentences for each comparative type. In addition, some non-comparative sentences identified as false positives in the previous step were brought, 170 for Twitter and 234 for Buscapé.

For categorization, we used the four machine learning algorithms previously explored through 35 replications performed using 5-fold cross-validation. The number of repetitions is determined to ensure we have the minimum required values to achieve a 95% confidence level in our results. The dataset is divided into 5 folds, with each fold used once as a validation set, while the remaining folds are for training. This number of folds is appropriate for the task and is commonly used in many machine-learning models. This iterative process covers all 5 folds, ensuring each one is validated once. After all 5 folds have been validated, we create a new set of five folds and repeat the validating process. We repeat this process seven times, resulting in a total of 35 validations. Averaging results across all iterations gives a reliable performance estimate, mitigating overfitting and aiding robust model evaluation. In particular, LR and SVM had similar results in terms of Macro-F1 across both datasets. However, LR demonstrated superior performance on the Buscapé dataset, achieving a Macro-F1 of 61.9% ( $\pm 0.01$ ). On the other hand, for the Twitter dataset, SVM demonstrated better performance, achieving a Macro-F1 score of 61.6% ( $\pm 0.012$ ). Consequently, these algorithms emerged as the most effective for distinguishing the non-comparative class from the other comparative groups.

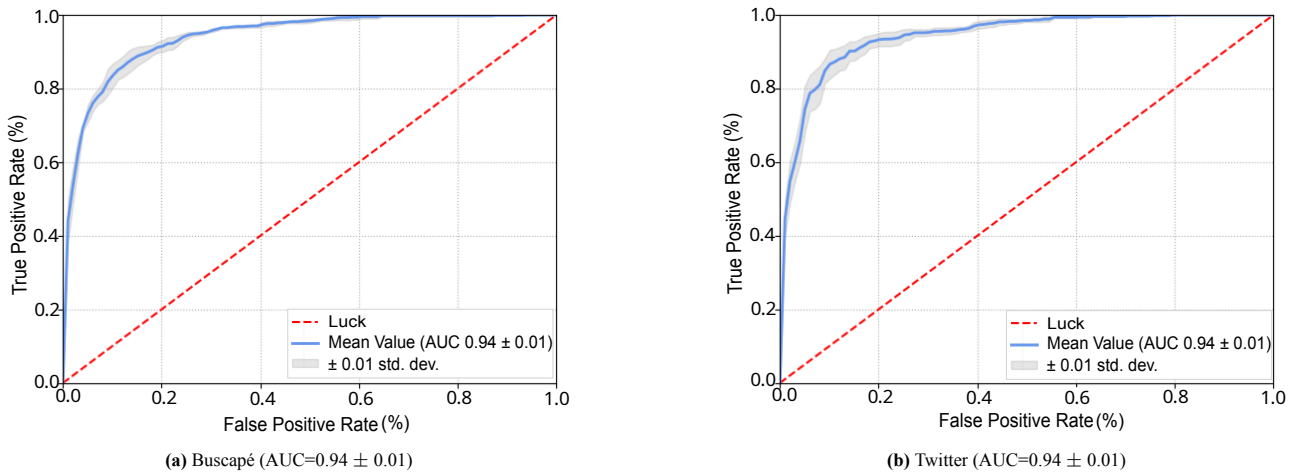
While accuracy is an important metric, the frequency of the right classification of each class (i.e. recall) is critical to categorizing comparisons into different classes. Table 5 presents the result, where the values indicate the frequency of classification of each class. Although there are different types of comparisons, the main challenge remains the distinction between comparative and non-comparative sentences. The false positives detected in the binary classification show significant similarities with the comparative sentences, leading to a relatively lower classification result for this class due to the lack of discernible patterns compared to the other

**Table 2.** Precision, Recall and F1-Score with 95% confidence in binary classification for Buscapé and Twitter.

Buscapé								
	Non-Comparative			Comparative			Total	
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Accuracy	Macro-F1
<b>RF</b>	0.741 ± 0.006	0.801 ± 0.008	0.77 ± 0.006	0.749 ± 0.008	0.679 ± 0.001	0.712 ± 0.007	0.744 ± 0.006	0.741 ± 0.006
<b>LR</b>	0.861 ± 0.006	0.863 ± 0.007	0.862 ± 0.005	0.843 ± 0.007	0.839 ± 0.008	0.841 ± 0.006	0.852 ± 0.005	0.851 ± 0.006
<b>SVM</b>	0.869 ± 0.005	<b>0.895 ± 0.006</b>	<b>0.882 ± 0.004</b>	<b>0.875 ± 0.007</b>	0.845 ± 0.006	<b>0.86 ± 0.005</b>	<b>0.872 ± 0.005</b>	<b>0.871 ± 0.005</b>
<b>NB</b>	<b>0.909 ± 0.005</b>	0.847 ± 0.006	<b>0.877 ± 0.005</b>	0.838 ± 0.006	<b>0.903 ± 0.006</b>	<b>0.869 ± 0.005</b>	<b>0.873 ± 0.004</b>	<b>0.873 ± 0.004</b>

Twitter								
	Non-Comparative			Comparative			Total	
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Accuracy	Macro-F1
<b>RF</b>	0.741 ± 0.007	0.84 ± 0.011	0.787 ± 0.008	0.764 ± 0.013	0.637 ± 0.011	0.695 ± 0.01	0.749 ± 0.008	0.741 ± 0.009
<b>LR</b>	0.831 ± 0.006	0.874 ± 0.007	0.851 ± 0.005	0.833 ± 0.008	0.779 ± 0.01	0.805 ± 0.007	0.831 ± 0.006	0.828 ± 0.006
<b>SVM</b>	0.834 ± 0.007	<b>0.912 ± 0.005</b>	<b>0.871 ± 0.005</b>	<b>0.878 ± 0.007</b>	0.775 ± 0.011	0.823 ± 0.007	<b>0.851 ± 0.006</b>	0.847 ± 0.006
<b>NB</b>	<b>0.894 ± 0.007</b>	0.851 ± 0.008	<b>0.872 ± 0.005</b>	0.827 ± 0.007	<b>0.874 ± 0.009</b>	<b>0.85 ± 0.006</b>	<b>0.862 ± 0.005</b>	<b>0.861 ± 0.005</b>



**Figure 4.** ROC Curve for Multinomial Naive Bayes (NB).

**Table 3.** Classification frequency for each class with Multinomial Naive Bayes (NB).

Buscapé			
		Predicted Label	
		Non-Comparative	Comparative
True Label	Non-Comparative	<b>84.7%</b>	15.3%
	Comparative	9.7%	<b>90.3%</b>

Twitter			
		Predicted Label	
		Non-Comparative	Comparative
True Label	Non-Comparative	<b>85.1%</b>	14.9%
	Comparative	12.6%	<b>87.4%</b>

classes. However, the low recall rate for the non-comparative sentence group is not problematic since approximately 85% of the sentences have already been separated through binary classification.

Moreover, these non-comparative sentences constitute a minority within the new dataset. Additionally, these non-comparative sentences share patterns with non-gradable sentences, which represent the least crucial comparisons as they do not convey preferences. This observation is emphasized in Table 5, where the recall rates for non-comparative and non-gradable classes are the lowest. This suggests that the impact of false positives is minimal, as misclassifying them as non-gradable does not affect the main objective of detecting preference.

On the other hand, comparative sentences showed good results. Superlatives and Equatives make use of specific expres-

**Table 4.** Sentences classified as comparative by Multinomial Naive Bayes (NB).

Sentences	Buscapé	Twitter	Total
Non-Equal Gradable	502	279	781
Equative	255	172	427
Superlative	290	270	560
Non-Gradable	115	81	196
<b>Total Comparatives</b>	<b>1,162</b>	<b>802</b>	<b>1,964</b>
<b>Total Non-Comparatives</b>	<b>234</b>	<b>170</b>	<b>404</b>

sions that allow them to be more easily distinguished from others, as can be seen in Table 5. The opposite is true for Non-Gradable, which are more complex and generally lack a clear pattern. These sentences can often be confused with Non-Equal Gradable or even non-comparative sentences.

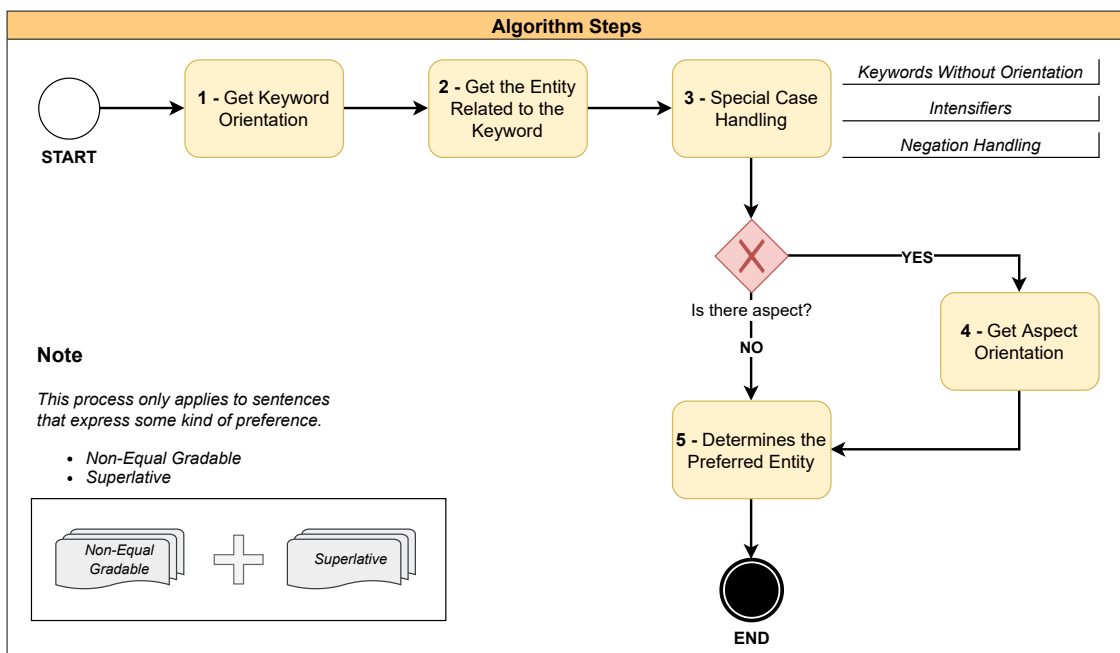
Despite the differences in the contexts, the results show that there is no significant difference in the task of classifying sentences in the datasets.

## 7 Preference Detection

After finding the comparative sentences, determining the sentiment expressed by each entity is one of the most relevant pieces of information, as it allows for identifying the preferred entity in a sentence. In this work, preference is considered as the act of choosing one item over others or the action

**Table 5.** Categorization frequency for Buscapé (LR - Macro-F1 = 61.9%±0.008) and Twitter (SVM - Macro-F1 = 61.6%±0.09).

		Buscapé				
		Predicted Label				
		Non-Comparative	Non-Equal Gradable	Equative	Superlative	Non-Gradable
True Label	Non-Comparative	34.3%	29.7%	13.1%	14.2%	8.7%
	Non-Equal Gradable	9.8%	75.1%	5.4%	5.7%	4.0%
	Equative	11.3%	9.9%	74.3%	2.3%	2.2%
	Superlative	215.9%	11.6%	2.3%	79.8%	0.5%
	Non-Gradable	17.3%	22.7%	10.6%	3.7%	45.7%
		Twitter				
		Predicted Label				
		Non-Comparative	Non-Equal Gradable	Equative	Superlative	Non-Gradable
True Label	Non-Comparative	37.9%	21%	15.9%	20.7%	4.5%
	Non-Equal Gradable	5.6%	85.6%	3.7%	4.9%	0.3%
	Equative	11.0%	7.8%	78.9%	2.2%	0%
	Superlative	13.8%	16.5%	2.5%	66.9%	0.3%
	Non-Gradable	33.0%	0.7%	30.2%	0.5%	35.6%



**Figure 5.** Algorithm steps to determine the preferred entity.

of indicating the superiority of one item over the others. In the sentence “Smartphone X is *better* than Y”, we can see the existence of two items, Smartphone X and Y, the former being pointed out as preferred.

A comparative opinion is composed of a set of elements that have been identified and analyzed together to collect relevant information. A comparison can be represented by a tuple with six elements as  $(E_1, E_2, A, PE, h, t)$  [Liu, 2012], in which  $E_1$  and  $E_2$  represent the entities being compared in the sentence, with  $E_1$  being mentioned before  $E_2$ . The  $A$  element, on the other hand, represents the aspect associated with the entities being compared. Finally,  $PE$  represents the preferred entity, which is indicated by an opinion leader  $h$  at a given time  $t$ .

This section presents the five steps of the proposed algorithm to find the preferred entity in a comparative sentence, as shown in Figure 5. Initially, the algorithm detects the orientation of the keyword used to make the comparison (Section 7.1) and then finds the entity associated with that word

(Section 7.2). Next, the algorithm handles the special cases (Section 7.3), such as keywords without orientation, intensifiers, and negation. Also, handle cases that require aspect-based context analysis to determine preference (Section 7.4). Finally, if the final orientation is positive, we can assume that the entity related to the keyword is preferred (Section 7.5). The proposed algorithm specifically targets Non-Equal Gradable and Superlative sentences, as they constitute the exclusive types of comparisons that articulate a preference, encompassing approximately 70% of the sentences within the datasets.

### 7.1 STEP 1: Getting Comparative Keyword Orientation

Comparative opinions have a comparative keyword that establishes a relationship between the entities being compared. One of the effective ways of interpreting a comparative opinion is the comparative analysis of the keyword, as it contains



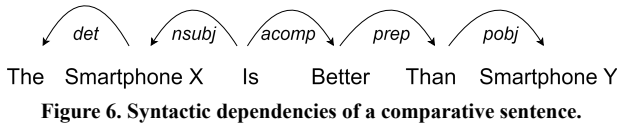


Figure 6. Syntactic dependencies of a comparative sentence.

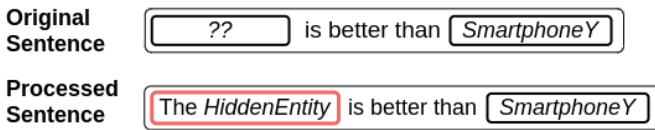


Figure 7. Processing hidden entity comparisons.

properties that indicate the relationship between the entities of the sentence.

These comparative keywords have an orientation, which refers to the situation in which that keyword places its related entity, which can be superiority or inferiority. Thus, the orientation value of each comparative keyword of the lexicon was mapped by adding a positive value (+1) if this word indicates superiority and a negative (-1) if it means inferiority.

### 7.2 STEP 2: Entity Detection

In addition, it is necessary to identify which entity is related to the comparative keyword, which allows for establishing the relationship of degrees between the entities and determining which is superior. In comparison, an entity can be a product, service, brand, organization, person, event, situation, or a topic [Liu, 2012]. In this context, the analysis of syntactic dependencies is one of the ways to find the relationship between words in a sentence, discovering which terms are directly related to the keyword, as shown in Figure 6.

Through syntactic dependencies, the proposed strategy uses the related entity and goes through the list of comparative keyword dependencies, prioritizing dependencies of type *nsubj* and *obj*. By tracing the links of keywords, we can determine their associated entities, as illustrated in Figure 6.

**Hidden Entity.** To ensure that syntactic dependency analysis works correctly, the compared objects must be explained in the sentence. Thus, for cases where the entity is hidden, a word indicating this hidden entity is automatically included at the beginning of the sentence, as shown in Figure 7. In this way, the sentence has a syntactic structure that allows the correct identification of the entity associated with the comparative keyword.

### 7.3 STEP 3: Special Case Handling

The main idea of our strategy is to use word orientation to define preferences. If the comparative keyword in the sentence indicates superiority, the associated entity is automatically indicated as preferred. Otherwise, the other entity in the comparison is preferred. This strategy works well for the majority of sentences, but there are some cases where we need to consider some other information to define the preference. Here, we start processing to handle special cases that fall outside the normal flow of algorithm execution.

**Comparative Keywords Without Orientation.** Although most comparative words contain orientation information, in words like *compared*, *difference*, *relation* and *as good as* it

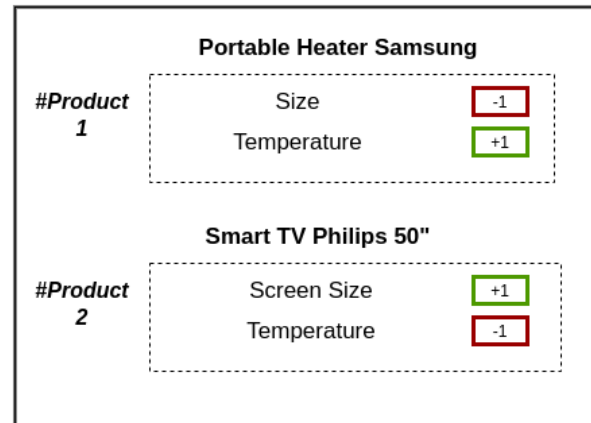


Figure 8. List with the mapping of the objects' aspects.

is not possible to determine the orientation without analyzing the context and sentiment involved in the close words. To handle these cases is proposed a strategy that consists of finding the closest term to the comparative keyword, following the order of priority (adjectives, adverbs, and verbs). Thus, the sentiment is obtained by combining three Portuguese lexicons, SentiStrength [Thelwall and Buckley, 2013], LIWC [Balage Filho et al., 2013], and Onto.PT [Gonçalo Oliveira and Gomes, 2014]. In this case, the final sentiment value represents the keyword orientation.

**Increase/Decrease Words.** The presence of an adverb next to a keyword can change the meaning and the way in which the comparative sentence should be interpreted. In the sentence “Product X is **almost** better than Product Y”, the decreased word *almost* makes the entity Product Y preferred. Thus, we proposed a strategy to interpret these expressions (increase/decrease word + comparative keyword), consisting of four rules:

1. Increase word + Keyword (+1) = **Superiority**
2. Increase word + Keyword (-1) = **Inferiority**
3. Decrease word + Keyword (+1) = **Inferiority**
4. Decrease word + Keyword (-1) = **Superiority**

**Negation.** Orientation inversion is one of the techniques adopted to deal with the existence of negation. Therefore, the strategy proposed here reverses the original orientation of the comparative keyword if there is any negation. In the sentence: “X is **not** *superior* than Product Y”, the keyword *superior* expresses superiority, however, this orientation is reversed due to the existence of negation, making Product Y the preferred one.

### 7.4 STEP 4: Comparisons with Aspects

One of the main challenges in detecting preference is identifying the preferred entity in a sentence with aspects, as it is necessary to analyze the context to determine preference. In the sentence “Smartphone battery X is *longer* than Y”, (battery, bigger) is a positive feature, but the same keyword in other context “Application A has a *longer* runtime than application B”, (runtime, longer) refers to negative feature. Therefore, the same keyword can have a completely different meaning depending on the aspect.

In some cases, it is possible to determine preference without analyzing the context, sentences with words like *better*,

worse, and *prefer* express a preference regardless of the aspect mentioned. The same occurs in aspects common to products, such as price, cost-effectiveness, and weight, which always have the same context (e.g., low price is always good).

To address other aspects, it is necessary to map them within the context, extracting the orientation of each aspect and constructing a dictionary-like structure, illustrated in Figure 8. The aspects of each product receive an orientation, which can be positive (+1) when a greater intensity of the aspect is desired, otherwise a negative value (-1). By employing this strategy, we can ascertain preferences through the multiplication of aspects and the orientation derived from keywords. Thus, an aspect and a keyword with a negative orientation indicate superiority: “The temperature (-1) of Smart TV X is *lower* (-1) than Smart TV Y”

---

#### Algorithm 1 Preference detection in Non-Equal-Gradable sentences.

---

```

1: procedure getPreference(text, keyword, aspect, ent1, ent2)
2:   wordOrientation = getWordOrientation(keyword)
3:   relatedEntity = getRelatedEntity(text, keyword, ent1, ent2)
4:   if keyword contains decrement expression then
5:     wordOrientation = wordOrientation * (-1)
6:   end if
7:   if keyword contains negative expression then
8:     wordOrientation = wordOrientation * (-1)
9:   end if
10:  if aspect is not null then
11:    aspectOrientation = getFeatureOrientation(aspect)
12:    wordOrientation = wordOrientation * aspectOrientation
13:  end if
14:  if wordOrientation > 0 then
15:    preferredEntity = relatedEntity
16:  else
17:    preferredEntity = if relatedEntity == ent1 then ent2 else ent1
18:  end if
19:  return preferredEntity
20: end procedure

```

---

## 7.5 STEP 5: Determines the Preferred Entity

Summarizing the previous steps, we can determine preference with Algorithm 1 for Non-Equal-Gradable and Algorithm 2 to Superlative opinions. After all, if the final orientation is positive, the preferred entity is related to the comparative keyword. The number of entities is the main difference between the Non-Equal-Gradable and Superlative comparisons. While the first opinion has two or more entities, the Superlatives compare an object to a group of others, expressing superiority or inferiority between them.

To detect the preferred entity in Superlatives it is not necessary to get the entity related to the comparative keyword, we can consider the only mentioned entity as a related entity. This is the main difference between Algorithm 1 and Algorithm 2. In this way, if the mentioned entity in the superlative opinion is not preferred, then the product group is superior to the explicitly mentioned entity.

**Table 6.** Results of the preference detection algorithm for Non-Equal Gradable in Buscapé and Twitter datasets.

Buscapé			
Preference	Precision	Recall	F1-Score
First entity is preferred	0.799	0.846	0.822
Second entity is preferred	0.873	0.832	0.852
<b>Accuracy</b>	<b>0.838</b>		
<b>Macro-F1</b>	<b>0.837</b>		
Twitter			
Preference	Precision	Recall	F1-Score
First entity is preferred	0.802	0.873	0.836
Second entity is preferred	0.877	0.808	0.841
<b>Accuracy</b>	<b>0.839</b>		
<b>Macro-F1</b>	<b>0.839</b>		

---

#### Algorithm 2 Preference detection in Superlative sentences.

---

```

1: procedure isSuperlativeEntityPreferred(text, keyword, aspect, ent1, ent2)
2:   wordOrientation = getWordOrientation(keyword)
3:   if keyword contains decrement expression then
4:     wordOrientation = wordOrientation * (-1)
5:   end if
6:   if keyword contains negative expression then
7:     wordOrientation = wordOrientation * (-1)
8:   end if
9:   if aspect is not null then
10:    aspectOrientation = getFeatureOrientation(aspect)
11:    wordOrientation = wordOrientation * aspectOrientation
12:   end if
13:   if wordOrientation > 0 then
14:     return True
15:   else
16:     return False
17:   end if
18: end procedure

```

---

## 8 Preference Detection Results

This section presents the results of both algorithms (i.e., 1 and 2), proposed for the detection of the preferred entity in Non-Equal-Gradable and Superlatives. Next, we evaluate the importance of each step of the algorithm for the final result.

### 8.1 Evaluation of the Algorithms 1 and 2

For evaluation, metrics used in the evaluation of machine learning models [Baeza-Yates *et al.*, 1999] were used. In addition, the sentences labeled in the Twitter and Buscapé datasets were used in the evaluation. In all, there are 910 sentences marked as Non-Equal Gradable and 602 sentences as Superlative. These sentences, in addition to having the type of opinion, have information about the entities compared and which one is preferred.

Initially, the proposed algorithm was tested on each of the datasets, reaching a Macro-F1 of 83.7% in the Buscapé and 83.9% in the Twitter base, as shown in Table 6. The results are summarized into two classes. The first class is when the first entity mentioned is indicated as preferred, and the second class is for cases where the second entity is pointed out as preferred in the comparison.

Observing the results of preference detection in both datasets, there is no significant difference between the sentences analyzed in each context. Social networks are known for more informal language, and when it comes to comparisons, they also have a variety of objects being compared that go beyond the context of products and services seen on most

**Table 7.** Results of the preference detection algorithm for Superlative in Buscapé and Twitter datasets.

Buscapé			
Preference	Precision	Recall	F1-Score
Mentioned entity is preferred	0.993	0.979	0.986
Second entity is <b>not</b> preferred	0.788	0.913	0.840
<b>Accuracy</b>	<b>0.974</b>		
<b>Macro-F1</b>	<b>0.913</b>		
Twitter			
Preference	Precision	Recall	F1-Score
Mentioned entity is preferred	0.979	0.974	0.976
Second entity is <b>not</b> preferred	0.895	0.911	0.903
<b>Accuracy</b>	<b>0.962</b>		
<b>Macro-F1</b>	<b>0.940</b>		

websites, all these differences go unnoticed by looking only at the results obtained.

One of the reasons is that, as we are working at the sentence level, the contextual factors that could influence a difference in the results are mitigated due to the treatments, as is the case of the hidden entity, which without treatment can cause inconsistencies in the detection of preferences.

Analyzing the Superlative sentences, the Algorithm 2 was evaluated in both datasets in a similar way to the Non-Equal-Gradable. Since these sentences do not explicitly compare two entities as Non-Equal-Gradable, how do the classes used to evaluate the approach refer to when: (1) the entity mentioned in the sentence is preferred, and (2) the entity mentioned is inferior to the group of opposing objects, or the entity mentioned is not superior in the comparison.

Evaluating the result of the Algorithm 2 for superlative comparisons, we can observe a superior precision when compared to the result obtained for Non-Equal-Gradable. For Buscapé, a Macro-F1 value of 91.3% was obtained, while for Twitter, a value of 94%, as shown in Table 7.

The approach presented by Ganapathibhotla and Liu [2008] addresses Superlative and Non-Equal-Gradable simultaneously, employing a modeling approach similar to ours. has the same modeling as our case. The outcomes indicate an improved Macro-F1 (99.6% compared to 89.6%) for preference detection when the first entity is favored. Nevertheless, our algorithm demonstrates superior performance with a higher Macro-F1 (88.4% compared to 82.5%) in scenarios where the preference lies with the second entity.

The variation in results observed between Non-Equal Gradable and Superlatives stems from the distinctive characteristics of each opinion and the differing number of steps required to discern preference in each scenario. To comprehend these outcomes, it is essential to examine the specific reasons that lead the algorithm to inaccurately identify preference, as detailed in Section 9.

## 8.2 Evaluation of the Preference Detection Steps

To fully comprehend the impact of each algorithmic step, it is essential to have a thorough understanding of each step in the aggregate process of preference detection. This understanding plays a crucial role in identifying the most critical steps in the algorithm.

Figures 9 and 10 show the accuracy and Macro-F1 metrics

**Table 8.** Reasons of Misclassification of Non-Equal Gradable sentences.

Reasons of Misclassification (Non-Equal Gradable)	Buscapé	Twitter
Hidden Preference	2.09%	0%
Dependency Parsing Error	11.65%	11.94%
Keyword Orientation Poorly Estimated	1.22%	0.59%
Preference Changed by the Context	0.87%	2.09%
Negation	0.34%	1.5%
<b>Total Error (%)</b>	<b>16.17%</b>	<b>16.12%</b>

**Table 9.** Reasons of Misclassification of Superlative sentences.

Reasons of Misclassification (Superlative)	Buscapé	Twitter
Keyword Orientation Poorly Estimated	1.28%	2.41%
Negation	0.64%	0%
Preference Changed by the Context	0.64%	1.38%
<b>Total Error (%)</b>	<b>2.56%</b>	<b>3.79%</b>

for each of the five steps of the algorithm. The first bar of the graph represents the simple preference detection procedure (SIMPLE DETECTION), which finds the entity associated with the keyword and determines the preference without any treatment. Next, we have the inclusion of treatments for the three special cases, they are: (1) when the keyword has no orientation (KEYWORD); (2) adverbs of intensity (ADV); and (3) negation (NEG). Finally, the last bar represents the complete algorithm, with all the previous steps plus the inclusion of aspect handling (ASPECTS).

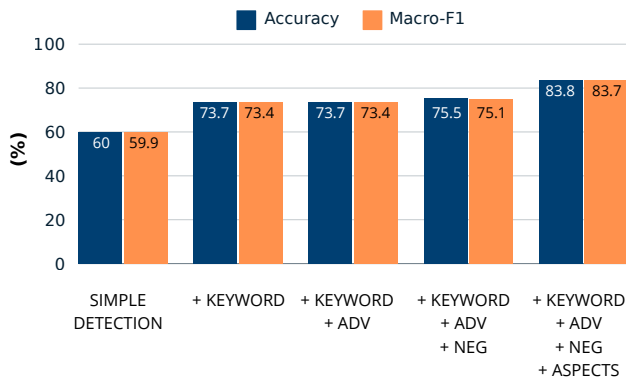
Analyzing each of the steps represented in the graphs, it is clear that the treatment of keywords without guidance (second slash) and the treatment of aspects (last slash) are the most important, as they are the main ones responsible for the considerable increase in the assertiveness of the algorithm. These two steps together represent a 14% to 20% increase in the Macro-F1 of the preference detection algorithm.

On the other hand, the processing steps for adverbs of intensity and negation have a negligible impact on the performance of the algorithm. One of the reasons is that, despite being important treatments, the datasets do not have many sentences that need these treatments. Therefore, these procedures end up not generating a significant difference in the result of the algorithm.

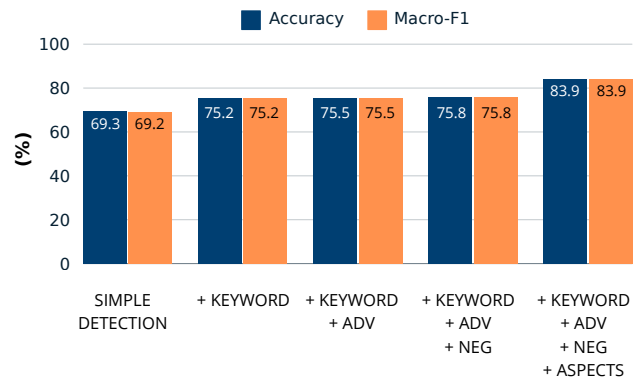
In general, it can be concluded that the detection of preference without any treatment is not so effective, however, with the inclusion of the steps presented, the algorithm becomes able to cover the vast majority of comparisons satisfactorily, both Non-Equal-Gradable, to Superlatives.

## 9 Reasons of Misclassification

Within this Section, we will examine the factors contributing to the errors in the preference detection algorithm. To gain insights into the disparity in results, we analyze the outcomes of Non-Equal-Gradable and Superlative cases. Table 8 highlights the reasons behind the incorrect determination of preference in Non-Equal-Gradable cases. Out of the total error rate of almost 16%, slightly over 11% can be attributed to dependency parsing. When the dependency parsing fails to accurately capture the dependencies within a sentence, it becomes challenging to identify the entity associated with the keyword, and, consequently, determine the preferred entity.

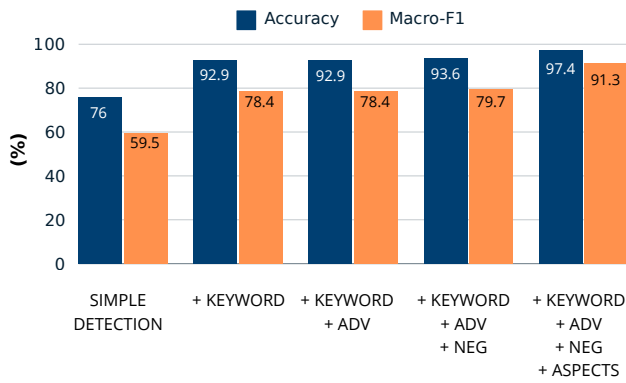


(a) Buscapé Dataset

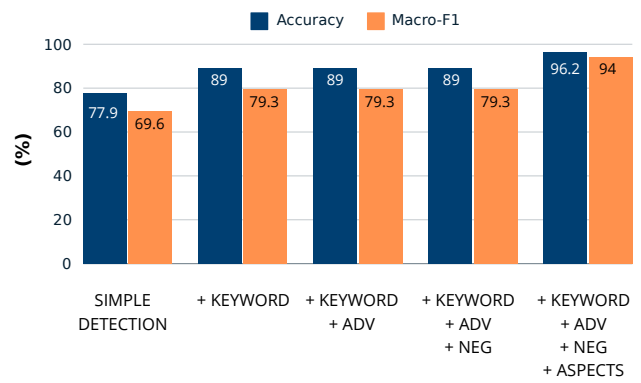


(b) Twitter Dataset

Figure 9. Accuracy of each algorithm step in Non-Equal Gradable opinions.



(a) Buscapé Dataset



(b) Twitter Dataset

Figure 10. Accuracy of each algorithm step in superlative opinions.

Despite dependency parsing being a very useful tool and, in general, having a success rate, there are still many challenges to be faced when it comes to the Portuguese language. Slang, typos, missing commas, informal expressions, and other grammatical errors make the dependency parsing task difficult, making it difficult to analyze dependencies. Even with these challenges, dependency analysis still achieves satisfactory results for the Portuguese, showing a precision of approximately 90% for Non-Equal-Gradable sentences.

Analyzing Table 9 with the reasons for the classification errors of the Superlative opinions, we can observe the absence of the dependency parsing step results in a much smaller aggregate error ( $\approx 3\%$ ), since the main task, detection of the entity associated with the keyword is not required.

The absence of the keyword-related entity detection step is one of the main reasons for better performance for the Superlative opinions. As these sentences have only one entity explicitly mentioned, it can be assumed that the comparative keyword refers to this mentioned entity. This allows determining preference without the need for dependency parsing, minimizing aggregate error and simplifying the preference detection process from superlative opinions. In addition, other reasons for misclassification have been found, they are:

when the treatment of the hidden entity is not sufficient, the existence of expressions in specific contexts that modify the way of determining preference, errors to identify a negation and, finally, situations in which the keyword has no orientation and, even analyzing nearby words, it is not possible to correctly estimate the orientation. Nonetheless, such situations are infrequent, and in most cases, the proposed approach can effectively handle the majority of comparisons by accurately capturing the sentence preference.

## 10 Limitation

The study has a limitation associated with its reliance on a lexical approach in the methodology. While this approach successfully captures a significant portion of comparative sentences, it is dependent on contextual factors that may involve specific patterns or expressions. Nonetheless, this limitation can be effectively addressed by expanding the lexicon to incorporate new words. This expansion will enhance the methodology’s capacity to handle diverse contexts, leading to more comprehensive coverage and improving the overall effectiveness of the approach. Additionally, preference de-

tection involves a series of interconnected steps, wherein the failure of one step can have a cascading impact on subsequent steps, affecting the overall accuracy of preference determination. However, it is important to note that the algorithm's steps do not exhibit a substantial error that would significantly impact the overall result by magnifying the aggregate error. Additionally, the algorithm's strength lies in the compartmentalization of these steps, making it easier to modify individual components without requiring changes to the entire process. This enables targeted enhancements and optimizations within the preference detection process, improving its overall performance.

## 11 Conclusion

This work presented a new opinion-mining study in online reviews written in Portuguese. In this context, one of the fundamental tasks in opinion mining is to separate regular from comparative opinions. Thus, firstly, based on a set of comparative keywords we build new labeled datasets containing comparative sentences in Portuguese from distinct scenarios: (1) review sites; and (2) Online Social Networks (OSN).

We then proposed a hierarchical supervised machine learning approach to distinguish comparative sentences from the others, reaching 87% in terms of Macro-F1. Next, each comparative sentence identified is categorized into five types of opinions: (1) Non-Comparative; (2) Non-Equal-Gradable; (3) Equative; (4) Superlative; and (5) Non-Gradable. Overall, the results show that the approach is promising in detecting comparative sentences and allows for a more systematic analysis of opinions and preferences.

Last, an algorithm for detecting the preferred entity in comparative sentences was proposed. It can find the preference in a sentence through the comparative keyword and textual elements. The obtained results also are promising, with 84% Macro-F1 for Non-Equal-Gradable and 94% for Superlatives, respectively. In sum, our main contributions are four-fold: (1) two new datasets; (2) a hierarchical machine learning approach to detecting multiple comparisons; (3) an algorithm to detect preferred entities in comparative sentences; and (4) the building of a lexicon with words used in Portuguese comparisons.

In future work, we plan to explore new datasets, recognizing that a larger volume of data provides the opportunity to explore alternative techniques, including deep learning, and incorporate novel classification features. Additionally, we intend to build a Web tool that encompasses the developed approach. We hope it will allow researchers and companies to apply opinion-mining strategies to detect comparisons and preferences in different scenarios.

## Declarations

## Acknowledgements

This work was partially supported by CAPES, FAPEMIG, CNPq, FAPESP, and the Big Data fellowship program.

## Authors' Contributions

All authors contributed equally to the conception of this study. Daniel Kansoon planned and implemented the approach and performed the analysis. Michele A. Brandão, Julio C. S. Reis, and Fabricio Benevenuto contributed to designing the methodology and conducting the experiments. Daniel Kansoon is the primary contributor and writer of this article. All authors reviewed and approved the final manuscript.

## Competing interests

The authors declare that they have no competing interests.

## Availability of data and materials

The produced datasets and lexicon are available at: <https://zenodo.org/records/4124410>

## References

- Alaei, A. R., Becken, S., and Stantic, B. (2019). Sentiment analysis in tourism: capitalizing on big data. *Journal of Travel Research*, 58(2):175–191. DOI: 10.1177/0047287517747753.
- Araujo, M., Diniz, J. P., Bastos, L., Soares, E., Ferreira, M., Ribeiro, F., and Benevenuto, F. (2016). ifeel 2.0: A multi-lingual benchmarking system for sentence-level sentiment analysis. In *Tenth International AAAI Conference on Web and Social Media*. DOI: 10.1609/icwsm.v10i1.14705.
- Asevedo Nóbrega, F. A. and Salgueiro Pardo, T. A. (2018). Update summarization: building from scratch for portuguese and comparing to english. *Journal of the Brazilian Computer Society*, 24(1):1–12. DOI: 10.1186/s13173-018-0075-1.
- Bach, N. X., Van, P. D., Tai, N. D., and Phuong, T. M. (2015). Mining vietnamese comparative sentences for sentiment analysis. In *Proc. of the KSE*, pages 162–167. DOI: 10.1109/KSE.2015.36.
- Baeza-Yates, R., Ribeiro-Neto, B., et al. (1999). *Modern information retrieval*, volume 463. ACM press New York. Book.
- Balage Filho, P., Pardo, T. A. S., and Aluísio, S. (2013). An evaluation of the brazilian portuguese liwc dictionary for sentiment analysis. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology (STIL)*, pages 215–219. Available at: <https://aclanthology.org/W13-4829.pdf>.
- Basiri, M. E., Nemati, S., Abdar, M., Cambria, E., and Acharya, U. R. (2021). Abcdm: An attention-based bidirectional cnn-rnn deep model for sentiment analysis. *Future Generation Computer Systems*, 115:279–294. DOI: 10.1016/j.future.2020.08.005.
- Berthene, A. (2022). Coronavirus pandemic adds \$219 billion to us ecommerce sales in 2020-2021. Available at: <https://www.digitalcommerce360.com/article/coronavirus-impact-online-retail/>. Accessed on April 16, 2022.
- Bespalov, D., Bai, B., Qi, Y., and Shokoufandeh, A. (2011). Sentiment classification based on supervised latent n-gram

- analysis. In *Proc. of the CIKM*, pages 375–382. DOI: 10.1145/2063576.2063635.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32. DOI: 10.1023/A:1010933404324.
- Buyya, R., Calheiros, R. N., and Dastjerdi, A. V. (2016). *Big data: principles and paradigms*. Morgan Kaufmann. Book.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46. DOI: 10.1177/001316446002000104.
- Dadhich, A. and Thankachan, B. (2022). Sentiment analysis of amazon product reviews using hybrid rule-based approach. In *Smart Systems: Innovations in Computing*, pages 173–193. Springer. DOI: 10.1007/978-981-16-2877-1\_17.
- de Barros, L. M. M. (2019). A correlação em construções comparativas da língua portuguesa. *Anais do IX SAPPIL-Estudos de Linguagem*. Available at: <http://www.anaisdosappil.uff.br/index.php/IXSAPPIL-Ling/article/view/932>.
- de Melo, T., da Silva, A., and de Moura, E. S. (2018). An aspect-driven method for enriching product catalogs with user opinions. *Journal of the Brazilian Computer Society*, 24(1):1–19. DOI: 10.1186/s13173-018-0080-4.
- de O. Carosia, A. E., Coelho, G. P., and Silva, A. E. d. (2019). The influence of tweets and news on the brazilian stock market through sentiment analysis. In *Proc. of the Web-Media*, pages 385–392. DOI: 10.1145/3323503.3349564.
- Ding, X., Liu, B., and Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining*, pages 231–240. DOI: 10.1145/1341531.1341561.
- Ding, X., Liu, B., and Zhang, L. (2009). Entity discovery and assignment for opinion mining applications. In *Proc. of the SIGKDD*, pages 1125–1134. DOI: 10.1145/1557019.1557141.
- D’Addio, R. M., Domingues, M. A., and Manzato, M. G. (2017). Exploiting feature extraction techniques on users’ reviews for movies recommendation. *Journal of the Brazilian Computer Society*, 23(1):1–16. DOI: 10.1186/s13173-017-0057-8.
- El-Halees, A. M. (2012). Opinion mining from arabic comparative sentences. In *Proc. of the ACIT*, pages 265–271. Available at: [https://www.researchgate.net/publication/276140611\\_OPINION\\_MINING\\_FROM\\_ARABIC\\_COMPARATIVE\\_SENTENCES](https://www.researchgate.net/publication/276140611_OPINION_MINING_FROM_ARABIC_COMPARATIVE_SENTENCES).
- Eldefrawi, M. M., Elzanfaly, D. S., Farhan, M. S., and Eldin, A. S. (2019). Sentiment analysis of arabic comparative opinions. *SN Applied Sciences*, 1(5):411. DOI: 10.1007/s42452-019-0402-y.
- Ganapathibhotla, M. and Liu, B. (2008). Mining opinions in comparative sentences. In *Proc. of the Coling*, pages 241–248. Available at: <https://aclanthology.org/C08-1031.pdf>.
- Gonçalo Oliveira, H. and Gomes, P. (2014). Eco and onto.pt: A flexible approach for creating a portuguese wordnet automatically. *Language Resources and Evaluation Journal*, 48(2):373–393. DOI: 10.1007/s10579-013-9249-9.
- Haque, T. U., Saber, N. N., and Shah, F. M. (2018). Sentiment analysis on large scale amazon product reviews. In *2018 IEEE international conference on innovative research and development (ICIRD)*, pages 1–6. IEEE. DOI: 10.1109/ICIRD.2018.8376299.
- Hartmann, N., Avanço, L., Balage Filho, P. P., Duran, M. S., Nunes, M. D. G. V., Pardo, T. A. S., Aluisio, S. M., et al. (2014). A large corpus of product reviews in portuguese: Tackling out-of-vocabulary words. In *Proc. of the LREC*, pages 3865–3871. Available at: <https://www.pedrobalage.com/assets/pdf/Hartmann2014LargeOpinionCorpus.pdf>.
- Huang, X., Wan, X., Yang, J., and Xiao, J. (2008). Learning to identify comparative sentences in chinese text. In *Proc. of the PRICAI*, pages 187–198. DOI: 10.1007/978-3-540-89197-0\_20.
- Ibeke, E., Lin, C., Wyner, A., and Barawi, M. H. (2017). Extracting and understanding contrastive opinion through topic relevant sentences. In *Proceedings of the International Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 395–400. Available at: <https://aclanthology.org/I17-2067/>.
- Jindal, N. and Liu, B. (2006a). Identifying comparative sentences in text documents. In *Proc. of the SIGIR*, pages 244–251. DOI: 10.1145/1148170.1148215.
- Jindal, N. and Liu, B. (2006b). Mining comparative sentences and relations. In *Proc. of the AAAI*, number 13311336, page 9. Available at: <https://cdn.aaai.org/AAAI/2006/AAAI06-209.pdf>.
- Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Proc. of the European Conference on Machine Learning (ECML)*, pages 137–142. DOI: 10.1007/BFb0026683.
- Kansaon, D., Brandão, M. A., Reis, J. C., Barbosa, M., Matos, B., and Benevenuto, F. (2020). Mining portuguese comparative sentences in online reviews. In *Proc. of the WebMedia*, pages 333–340. DOI: 10.1145/3428658.3431081.
- Kim, H. D. and Zhai, C. (2009). Generating comparative summaries of contradictory opinions in text. In *Proceedings of the 18th ACM International Conference on Information and knowledge management (CIKM)*, pages 385–394. DOI: 10.1145/1645953.1646004.
- Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., and Klein, M. (2002). *Logistic regression. A Self-Learning Text*. Springer, 3 edition. DOI: 10.1080/00401706.1995.10485899.
- Koppe, G., Meyer-Lindenberg, A., and Durstewitz, D. (2021). Deep learning for small and big data in psychiatry. *Neuropsychopharmacology*, 46(1):176–190. DOI: 10.1038/s41386-020-0767-z.
- Liu, B. (2012). *Sentiment analysis and opinion mining*, volume 5. Morgan & Claypool Publishers. Book.
- McCallum, A., Nigam, K., et al. (1998). A comparison of event models for naive bayes text classification. 752(1):41–48. Available at: <http://yangli-feasibility.com/home/classes/1fd2022fall/media/aaaiws98.pdf>.
- Mehta, R. P., Sanghvi, M. A., Shah, D. K., and Singh, A. (2020). Sentiment analysis of tweets using supervised

- learning algorithms. In *Proc of the ICTSCI*, pages 323–338. DOI: 10.1007/978-981-15-0029-9\_26.
- Melo, P. F., Dalip, D. H., Junior, M. M., Gonçalves, M. A., and Benevenuto, F. (2019). 10sent: A stable sentiment analysis method based on the combination of off-the-shelf approaches. *Journal of the Association for Information Science and Technology*, 70(3):242–255. Available at: [https://homepages.dcc.ufmg.br/~fabricio/download/Melo\\_et\\_al-2018-Journal\\_of\\_the\\_Association\\_for\\_Information\\_Science\\_and\\_Technology.pdf](https://homepages.dcc.ufmg.br/~fabricio/download/Melo_et_al-2018-Journal_of_the_Association_for_Information_Science_and_Technology.pdf).
- Nasti, S. J., Asger, M., and Butt, M. A. (2020). Automatic extraction of product information from multiple e-commerce web sites. In *Prof. of the ICRIC*, pages 739–747. DOI: 10.1007/978-3-030-29407-6\_53.
- Nilashi, M., Fallahpour, A., Wong, K. Y., and Ghabban, F. (2022). Customer satisfaction analysis and preference prediction in historic sites through electronic word of mouth. *Neural Computing and Applications*, pages 1–15. DOI: 10.1007/s00521-022-07186-5.
- Pathak, A. R., Pandey, M., and Rautaray, S. (2021). Topic-level sentiment analysis of social media data using deep learning. *Applied Soft Computing*, 108:107440. DOI: 10.1016/j.asoc.2021.107440.
- Paul, M. J., Zhai, C., and Girju, R. (2010). Summarizing contrastive viewpoints in opinionated text. In *Proceedings of the International Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 66–76. Available at: <https://aclanthology.org/D10-1007.pdf>.
- Pitman, J. (2022). Local consumer review survey 2022. Available at: <https://www.brightlocal.com/research/local-consumer-review-survey/>. Accessed on June 15, 2022.
- Pompeo, C. (2022). Brazil, argentina, and mexico ranked in the top 10 countries where e-commerce will grow fastest in 2022. Available at: <https://www.ebanx.com/en/>. Accessed on April 16, 2022.
- Ren, Z. and de Rijke, M. (2015). Summarizing contrastive themes via hierarchical non-parametric processes. In *Proceedings of the 38th International Conference on Special Interest Group on Information Retrieval (SIGIR)*, pages 93–102. DOI: 10.1145/2766462.2767713.
- Ribeiro, F. N., Araújo, M., Gonçalves, P., Gonçalves, M. A., and Benevenuto, F. (2016). Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):1–29. Available at: <https://link.springer.com/content/pdf/10.1140/epjds/s13688-016-0085-1.pdf>.
- Sapir, E. (1944). Grading, a study in semantics. *Philosophy of science*, 11(2):93–116. Available at: <https://www.cambridge.org/core/journals/philosophy-of-science/article/abs/grading-a-study-in-semantics/B8CC0CBE65DFAC87FBE142AD263478F9>.
- Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., and Herrera-Viedma, E. (2015). Sentiment analysis: A review and comparative analysis of web services. *Information Sciences*, 311:18–38. DOI: 10.1016/j.ins.2015.03.040.
- Singh, M., Jakhar, A. K., and Pandey, S. (2021). Sentiment analysis on the impact of coronavirus in social life using the bert model. *Social Network Analysis and Mining*, 11(1):1–11. DOI: 10.1007/s13278-021-00737-z.
- Souza, E., Costa, D., Castro, D. W., Vitorio, D., Teles, I., Almeida, R., Alves, T., Oliveira, A. L., and Gusmão, C. (2017). Characterising text mining: a systematic mapping review of the portuguese language. *IET Software*, 12(2):49–75. DOI: 10.1049/iet-sen.2016.0226.
- Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307. DOI: 10.1162/COLI\_a00049.
- Thelwall, M. and Buckley, K. (2013). Topic-based sentiment analysis for the social web: The role of mood and issue-related words. *Journal of the American Society for Information Science and Technology*, 64(8):1608–1617. DOI: 10.1002/asi.22872.
- Trisna, K. W. and Jie, H. J. (2022). Deep learning approach for aspect-based sentiment classification: A comparative review. *Applied Artificial Intelligence*, pages 1–37. DOI: 10.1080/08839514.2021.2014186.
- Tsai, C.-F., Chen, K., Hu, Y.-H., and Chen, W.-K. (2020). Improving text summarization of online hotel reviews with review helpfulness and sentiment. *Tourism Management*, 80:104122. DOI: 10.1016/j.tourman.2020.104122.
- Wang, Y., Huang, M., Zhu, X., and Zhao, L. (2016). Attention-based lstm for aspect-level sentiment classification. In *Proc. of the EMNLP*, pages 606–615. Available at: <https://aclanthology.org/D16-1058.pdf>.
- Wei, N., Zhao, S., Liu, J., and Wang, S. (2022). A novel textual data augmentation method for identifying comparative text from user-generated content. *Electronic commerce research and applications*, 53:101143. DOI: 10.1016/j.elerap.2022.101143.
- Xu, K., Liao, S. S., Li, J., and Song, Y. (2011). Mining comparative opinions from customer reviews for competitive intelligence. *Decision support systems*, 50(4):743–754. DOI: 10.1016/j.dss.2010.08.021.
- Xu, S., Zhang, X., Wu, Y., and Wei, F. (2022). Sequence level contrastive learning for text summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11556–11565. DOI: 10.1609/aaai.v36i10.21409.
- Yang, S. and Ko, Y. (2009). Extracting comparative sentences from korean text documents using comparative lexical patterns and machine learning techniques. In *Proc. of the AACL-IJCNLP*, pages 153–156. Available at: <https://aclanthology.org/P09-2039.pdf>.
- Yang, S. and Ko, Y. (2011). Extracting comparative entities and predicates from texts using comparative type classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 1636–1644. Available at: <https://aclanthology.org/P11-1164.pdf>.
- Younis, U., Asghar, M. Z., Khan, A., Khan, A., Iqbal, J., and Jilani, N. (2020). Applying machine learning techniques for performing comparative opinion mining. *Open Computer Science*, 10(1):461–477. DOI: 10.1515/comp-2020-0148.

Zhang, L., Wang, S., and Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253. DOI: 10.1002/widm.1253.