UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Aspect extraction in sentiment analysis for Portuguese

Pedro Paulo Balage Filho

Tese de Doutorado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP SÃO CARLOS

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: ______________________

Pedro Paulo Balage Filho

Aspect extraction in sentiment analysis for Portuguese

Doctoral dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP, in partial fulfillment of the requirements for the degree of the Doctorate Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Thiago Alexandre Salgueiro Pardo

USP – São Carlos
October 2017

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

Balage Filho, Pedro Paulo
BB171a Aspect extraction in sentiment analysis for Portuguese / Pedro Paulo Balage Filho; orientador Thiago Alexandre Salgueiro Pardo. -- São Carlos, 2017.
74 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2017.

1. Natural Language Processing. I. Pardo, Thiago Alexandre Salgueiro, orient. II. Título.

Pedro Paulo Balage Filho

Extração de aspectos em análise de sentimentos para língua portuguesa

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Thiago Alexandre Salgueiro Pardo

USP – São Carlos
Outubro de 2017

ACKNOWLEDGEMENTS

Thanks to the Instituto de Ciências Matemáticas e de Computação (ICMC-USP), the professors and the general staff, for all the support given and for maintaining this graduate program of excellence.

Thanks to the Professor Dr. Thiago Alexandre Salgueiro Pardo, my advisor, for his support, comprehension and patience. I am very glad to be part of the NLP community which it would not be possible without his help.

Thanks to all my colleagues in NILC for the good moments and friendship.

Thanks to my family for their support and specially to my wife, Juliana, the person who accepted to join me in this great journey of life.

Thanks to the companies Dicionário Criativo, Boo Labs and Priberam for showing me how much natural langauge processing could impact real products and for everything I learned from them.

Finally, thanks to the Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP - Grant Number 2012/16131-4) for the financial aid and support without this research would not be possible.

ABSTRACT

BALAGE FILHO, P. P. **Aspect extraction in sentiment analysis for Portuguese**. 2017. 74 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2017.

Aspect-based sentiment analysis is the field of study which extracts and interpret the sentiment, usually classified as positive or negative, towards some target or aspect in an opinionated text. This doctoral dissertation details an empirical study of techniques and methods for aspect extraction in aspect-based sentiment analysis with the focus on Portuguese. Three different approaches were explored: frequency-based, relation-based and machine learning. In each one, this work shows a comparative study between a Portuguese and an English corpora and the differences found in applying the approaches. In addition, richer linguistic knowledge is also explored by using syntatic dependencies and semantic roles, leading to better results. This work lead to the establishment of new benchmarks for the aspect extraction in Portuguese.

Keywords: Aspect-based sentiment analysis, Sentiment analysis, Opinion mining.

RESUMO

BALAGE FILHO, P. P. **Extração de aspectos em análise de sentimentos para língua portuguesa**. 2017. 74 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2017.

A análise do sentimento orientada a aspectos é o campo de estudo que extrai e interpreta o sentimento, geralmente classificado como positivo ou negativo, em direção a algum alvo ou aspecto em um texto de opinião. Esta tese de doutorado detalha um estudo empírico de técnicas e métodos para extração de aspectos em análises de sentimentos baseadas em aspectos com foco na língua Portuguesa. Foram exploradas três diferentes abordagens: métodos baseados na frequências, métodos baseados na relação e métodos de aprendizagem de máquina. Em cada abordagem, este trabalho mostra um estudo comparativo entre um córpus para o Português e outro para o Inglês e as diferenças encontradas na aplicação destas abordagens. Além disso, o conhecimento linguístico mais rico também é explorado pelo uso de dependências sintáticas e papéis semânticos, levando a melhores resultados. Este trabalho resultou no estabelecimento de novos padrões de avaliação para a extração de aspectos em Português.

Palavras-chave: Análise de sentimentos orientada a aspectos, Análise de Sentimentos, Mineração de Opiniões.

LIST OF FIGURES

LIST OF CHARTS

LIST OF TABLES

CONTENTS

CHAPTER 1

INTRODUCTION

The emerging of a new and complex social environment is undeniable. The **user generated content (UGC)** is now part of our daily lives, present when we read news, check the social networks or search for information in a forum or interest group website. As production and consumption of social data increases at an exponential rate, demonstrates the importance for new methods and techniques that might automatically mine and convert the raw unstructured data into structured information. A study by Gwava (2016) shows that every day, in average, we produce:

- 205 billion e-mails;
- 500 million tweets;
- 4 million hours of content uploaded to Youtube;
- 3.6 billion Instagram Likewavas;
- 4.3 billion Facebook messages;
- 5.75 billion Facebook likes;
- 40 million Tweets shares;
- 6 billion Google searches.

Despite the major example of user generated content be in the social networks, it is in the electronic commerce where it causes direct impact on buying decision. According to Burke (2002), the **social opinion impacts the decisions we make**, considering a product better than other by the simple judgment of a real user opinion. Thus, websites for product reviews have become an important resource to find opinions and influence users (BAILEY, 2005).

According to the white paper published by Social Annex (ANNEX, 2016), the UGC has been a valuable asset for companies, targeting user marketplaces and generating the word-of-mouth marketing of their products. According to their findings, 60% of consumers seek out product reviews and other forms of UGC before making a purchase. Other numbers that are also compiled in the same research are:

- over 50% of online consumers read reviews before buying;
- 54% of millennial shoppers read online reviews before shopping in stores;
- 88% of consumers trust online reviews as much as personal recommendations from friends;
- 47% of UK residents have reviewed a product online before;
- 72% of shoppers say that positive reviews make them trust a business more;
- only 10% of consumers do not take any notice of online reviews.

According to Kaplan and Haenlein (2010), one of the main reasons for the popularity of UGC is the emerging of social media platforms, making possible the interaction in a highly accessible and in a scalable environment. This contrasts to conventional media platforms, which are expensive to produce, require professional authors and are one-way only communication. Social media platforms are cheap, allow many authors and favor the interaction with the content produced.

In this context, many companies focus on offering an automatic way to assess the sentiment present in the social data. In specific, the analysis of the sentiment transmited by the user towards a specific topic or subject. This need is the object of analysis by this PhD research. In this direction, this chapter shows an outline for the following chapters and how this dissertation is structured.

Natural language processing (NLP) approach the study of techniques to extract structured data from this unstructured form of data. In NLP, the area that studies opinions and the associated sentiment is denominated **Sentiment Analysis**. According to Pang and Lee (2008), sentiment analysis aims to process the opinion, sentiment and subjectivity.

Liu (2010) alleges that all textual information present in the world may be categorized in only two types: facts and opinions. According to him, facts are objective expressions regarding entities, events and properties, while opinions are subjective expressions that describe sentiment, assessments, or emotions. Most of the information retrieval systems operates over facts, hence dealing with opinions remains a challenge, given the expressiveness of the language.

The study of sentiment analysis is categorized in three levels: text level, sentence level, and aspect or entity level (LIU, 2010). In the text level, one needs to determine whether the text has a positive or negative polarity about the overall topic. In the sentence level, the objective is to classify the sentence as positive, negative or neutral. In the **aspect or entity level**, the objective is to extract the sentiment associated to each aspect or entity in the text.

The following example illustrates these three levels:

The iPhone is a good device. The battery is excellent. The quality is very good, but the price is not affordable.

At the text level, we observe a predominance of positive sentiment (*good device, excellent battery, good quality*) instead of negative (*price not affordable*), leading us to classify the text as positive. At the sentence level, we might classify the first two sentences as positive because of the positive words (*good, excellent*). The last sentence shows both positive and negative sentiments. Usually, in sentiment analysis, the coordinate conjunction *but* indicates the main sentiment, thus the sentence would be classified as negative (SARMENTO *et al.*, 2009).

It is in the aspect level that we get a more refined analysis of the sentiment present in the text. Instead of looking for textual constructions (text, sentence, phrase), we look directly into the opinion. This level is based on the idea that an opinion contains an associated sentiment (positive or negative) and a target aspect. In the aspect level, sentiment is determined considering these aspects, and its polarity is determined by the context that it was used. For the previous excerpt, we may have the following analysis.

Aspect	Evaluative Word	Sentiment
iPhone	good	Positive
battery (iPhone)	excellent	Positive
quality (iPhone)	very good	Positive
price (iPhone)	not affordable	Negative

As one may see, it is very difficult to do any textual analysis without considering the aspects about which the sentiment is expressed. An author might be favorable towards some aspects, but unfavorable to others. Thus, text-level and sentence-level analysis are often flawed in determining the many sentiments and aspects that might be evaluated in a text containing opinions.

The indicative of the importance of sentiment analysis and, in special, aspect-based sentiment analysis is the increasing number of papers and workshops dedicated to this topic. In specific, the most important event dedicated to the task is the **Aspect-Based Sentiment Analysis Shared Task** that was carried out in the last three editions of SemEval workshop (SemEval ABSA)[1]. In the 2014 edition, the task attracted 163 submissions from 32 teams (PONTIKI *et al.*, 2014). In the 2015 edition, the task attracted 93 submissions from 16 teams (PONTIKI *et*

[1] http://alt.qcri.org/semeval2016/task5/

al., 2015). In the 2016 edition, the task attracted more than 120 submissions from more than 15 teams (PONTIKI *et al.*, 2016).

In each new edition of the SemEval ABSA shared task, the organization makes some changes in the structure of the task. In the last edition, the novelty was the inclusion of datasets in different languages: Arabic, Chinese, Dutch, French, Russian, Spanish and Turkish. The effort and great investment in this topic demonstrates the importance and the current trends in the area of NLP, which is also supported by the large number of publications related to the topic in the main conferences and journals in the area.

For the Portuguese language, the focus of this work, the corpus **ReLi** (FREITAS *et al.*, 2012) was studied for the research in question. ReLi is a corpus of book reviews, composed by 2,056 reviews from 13 different books with about 200 comments each. The annotations include manual sentence and aspect level sentiment analysis as an automatic tokenization and part-of-speech tagging. Figure 1 shows the aspect level annotation for the sentence *"Eu não me apaixonei pela história"* ("I did not fall in love with the plot") extracted from the corpus.

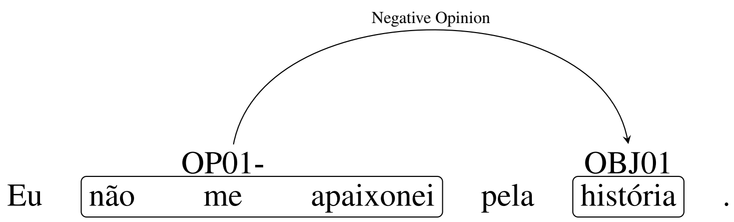

Figure 1 – Example of an excerpt from Reli corpus

In this sentence, the word *"história"* ("plot") labeled as OBJ01 is the object or aspect of the evaluation. The number 01 distinct from other aspects evalued in the same text. The excerpt *"Eu não me apaixonei"* ("I did not fall in love") labeled as OP01- is a negative opinion (minus signal) connected to the aspect *"história"* by the same number.

This research also studies the English corpora SemEval ABSA 2015 (PONTIKI *et al.*, 2015) and SemEval ABSA 2016 (PONTIKI *et al.*, 2016). These corpora allowed to benchmark the evaluated methods against state-of-art results for English as well as to study phenomena specifics to the Portuguese language or to the corpus under analysis.

The next sections show the objetives of this work followed by the hypotheses and contributions.

1.1. Objetives

The objective for this dissertation is to show different methods for the aspect extraction and show how these methods improve or not with more linguistic knowledge. In more specific, the **objectives** of this work comprehend:

1. To explore approaches based on frequency, relation and machine learning in aspect-based sentiment analysis and establish new benchmarks for the Portuguese;
2. to compare state-of-art approaches for English with Portuguese corpora;
3. to investigate the use of syntax and semantics in Portuguese ABSA methods;
4. to develop new tools and lexicons for sentiment analysis.

1.2. Hypotheses

The **hypotheses** investigated in this work are:

1. Deeper linguistic knowledge, such as syntax and semantics, improve aspect-based sentiment analysis.
2. Aspect-based sentiment analysis approaches do not differ between English and Portuguese.
3. Corpora from different domains show different challenges.

1.3. Contributions

The contributions brought by this work are:

1. The exploration of methods based on the frequency in new variations overcame the classical methods of this approach;
2. The exploration of classic methods based on relation and the inclusion of an automated step to learn patterns and thus leverage the approach;
3. A compilation for linguistic patterns in sentiment analysis for aspect extraction in Portuguese;
4. The exploration of methods based on machine learning and its enrichment with linguistic information of a syntactic and semantic nature, producing better results than the original methods;
5. Linguistic compilation for the semantic nature of the most frequent aspects in Portuguese language;
6. The training of a new syntactic parser based on Universal Tree Bank corpus, thus providing a new tool for the research area;
7. The pre-process and availability of the ReLi corpus with syntactic and semantic information.

1.4. Organization

Chapter 2 introduces to the reader the field of study of Natural Langauge Processsing, Sentiment Analysis, and the associated terminology.

Chapter 3 defines the problem of aspect-based opinion mining and presents the main works related to this topic. This chapter presents a comprehensive review of the state-of-the-art methods following the categorization proposed by Liu (2010): frequency-based methods, relation-based methods, topic modeling approaches and machine learning based approaches.

Chapter 4 presents the existing datasets and a detailed analysis for the datasets ReLi and SemEval used in this work. Complementary, the chapter also presentes the NLP tools that were used in this work.

Chapter 5 shows the analysis of methods based on frequency and relation for the task of aspect extraction.

Chapter 6 focuses on the analysis for methods based on machine learning. The entire chapter is dedicated to these methods, which demonstrated to perform as the state-of-art in many evaluations. In specific, the chapter shows conditional random fields algorithms and the inclusion of lexicon, syntactic and semantic features to improve the aspect extraction.

Chapter 7 presents a final discussion and concludes pointing direction for future researches.

CHAPTER

2

INTRODUCTION TO SENTIMENT ANALYSIS

Sentiment analysis, or **opinion mining**, is the area in Natural Language Processing that deals with the computational treatment of opinion, sentiment and subjectivity in text (PANG; LEE, 2008). In spite of its importance, this research area is still very new, having its initial works in the last decade (WIEBE, 2000; DAS; CHEN, 2001; MORINAGA *et al.*, 2002; PANG; LEE; VAITHYANATHAN, 2002; TONG, 2001; TURNEY, 2002).

The next section defines some terminology adopted by this work, which are the key concepts associated with sentiment analysis. As different authors may address different meanings to each term, this work seeks to present the terms define in the perspective of its author.

2.1. Sentiment, subjectivity, emotion, appraisal, opinion

Sentiment: Sentiment can be seen as a generic term to designate every text expressing positive, negative, or neutral characteristics. The term "sentiment" is broadly used and it may refer to subjectivity, emotion, appraisal and opinion.

As a better way to determine the definition of sentiment is to define the terms subjectivity, emotion, appraisal and opinion.

Subjectivity: is the presence in the text of sentiment, points of view or personal beliefs. Thus, a subjective sentence is that which contains any belief. In opposition, the sentence can be objective. An objective sentence is one that presents factual information about the world, and therefore has no subjectivity.

The following examples better define what is subjectivity.

(1) I have a cell phone.
(2) My cell phone is beautiful.

In the above examples, we have for (1) an objective sentence, since it is a factual information. In example (2), we have a subjective sentence, because it reflects a belief that the cell phone is beautiful, thus attributing a positive feeling to the cell phone entity.

It is important to realize that the concept of subjectivity is not directly related to the concept of sentiment. We may have the sentence *"I think I'm better"*, which is subjective, but has no sentiment. In the other hand, the sentence *"The battery did not last 2 minutes"* which is objective, presents implicit negative sentiment.

According to Riloff, Patwardhan and Wiebe (2006) and Wiebe, Wilson and Cardie (2005), subjective expressions may even take different forms, for example, opinions, claims, wishes, beliefs, suspicions and speculations.

In the literature, there is still some confusion between the definitions of subjective text and opinion text. For a text to be considered opinionated there must be the expression of sentiment related to an entity or a characteristic. Opinion texts will be better explained in Section 2.2. The task of determining whether a sentence is objective or subjective is called the subjectivity classification (WIEBE; WILSON; CARDIE, 2005).

We shall now describe the concept of emotion.

Emotion: Emotions are our subjective feelings and thoughts (LIU, 2012).

The study of emotions has been a broad field in psychology, philosophy, and sociology. One of the best known emotion-related studies is reported in Ekman (1992) and Parrott (2001). These works present six possible primary emotions: joy, surprise, anger, sadness, fear. In the study, illustrated by Figure 2, the researchers studied the expression of the human face in relation to different types of situations and reached to these six primary emotions. The study further subdivides these emotions into secondary categories, into secondary and tertiary emotions and at different intensities.

Emotions are directly related to sentiment. It is understood that the strength of a sentiment or opinion is tied to a certain emotion.

Another important term to be defined is evaluation. Evaluations can be categorized into two types (LIU, 2012; CHAUDHURI, 2006):

Rational assessments: These are assessments that are expressed through reason about beliefs and attitudes. For example, *"The battery of this phone is good "*, *"This phone has a great cost/benefit"* and *"I'm happy with this phone"*.

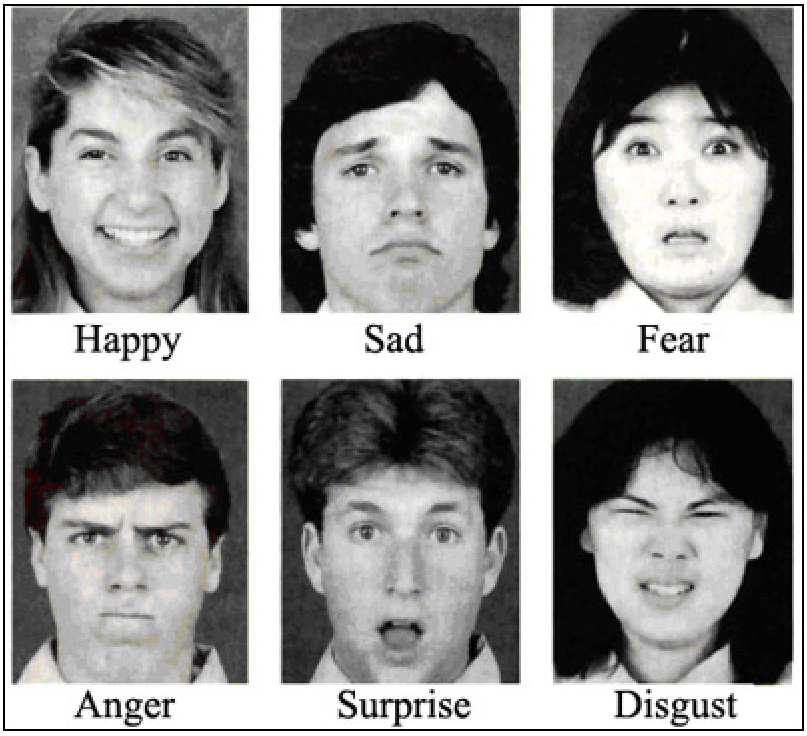

Figure 2 – Six primary emotions described by Ekman (1992).

Emotional assessments: assessments that are expressed from emotional and non-tangible responses. Examples are: *"I love this phone"*, *"I'm nervous about this service"*, *"The best shoe I've ever used"*.

Finally, we shall define the term opinion.

Opinion: An opinion basically consists of two components, a target object and an associated sentiment.

This dissertation focuses on the study of texts containing opinions. This way, the following section shows the concept of opinion in more details.

2.2. Opinion

Formally, a sentiment or opinion is defined by Liu (2010) as a quintuple (o_j, f_{jk}, oo_{ijkl}, h_i, t_l) where o_j is an object, f_{jk} is a feature of the object o_j, oo_{ijkl} is the semantic orientation or polarity of the opinion on feature f_{jk} of object o_j, h_i is the opinion holder and t_l is the time when the opinion is expressed by h_i.

In this definition, the source of the opinion is the object or entity being analyzed. By entity we mean a product, service, topic, person, organization or event. It can be described by a set of attributes or sub-parts, which we define as features or aspects of the entity or object.

Although many works can be found with the terms features or entities, this work will follow the suggestion from Liu (2010) and use the term aspect to characterize this concept. According to Liu (2010), the term feature may be misinterpreted with the term feature used in machine learning. Other names found to designate aspects are facets, attributes, or topics.

The aspect of the entity or object is the part about which the opinion is expressed. For example, in the sentence *"iPhone battery life is good"*, we have the aspect *"battery life"* and the entity *"iPhone"* being evaluated.

The orientation or polarity of an opinion is the sentiment associated with this opinion, which may be positive, negative or neutral. This sentiment can be expressed through an evaluative word, such as the term *"great"*, in the sentence *"this product is great"*.

The opinion holder is the one who is expressing the opinion. In the sentence, *"President Obama believes the economy is good"*, the opinion-holder is *"President Obama"*. When the opinion holder is not explicit, it could be associated with the writer. This happens in product reviews where it is usual to assign the opinion holder directly to the user who wrote the review.

The last factor, the time when the opinion is expressed is the time at which the assessment was made. This attribute is especially important when we want to follow the evolution of opinions about aspects of a certain entity over time.

Liu (2010) also defines some concepts for the extraction of aspect-oriented opinions. For him, an opinion can be classified into two types: direct opinions and comparative opinions.

Direct opinions are those in which we have an evaluation or feeling about an aspect present in an object referred into the text. For example, *"I like the Android system"*.

Comparative opinion expresses a relation between two or more objects. This type of opinion is evident in the text by the use of comparative or superlative forms of adverbs or adjectives. For example: *"Android platform is better than iOS"*.

A direct opinion may still take the explicit or implicit form.

Explicit direct opinion is expressed explicitly in the sentence. For example, *"Quality is poor"*.

Implicit direct opinion occurs when an inference of the context and world knowledge is required to understand the expressed opinion. For example, *"The device broke in two days"*.

Aspects can appear in two forms in the text: explicit aspects and implicit aspects.

Explicit aspects that are explicit present in text. For example, the passage *"Its battery is excellent"* presents the aspect *"battery"* explicitly.

Implicit aspects that are only perceptible through inference. For example, in the passage *"The phone is too large"*, the dimension or size aspect is implicitly present.

2.3. Opinionated texts

Opinionated texts are texts that have user reviews and opinions about an entity, product, or object. They are intended to offer a critical view on the assessed entity as well as the author's argumentation about its opinion.

Opinionated texts can occur in several ways. The most common repositories of opinions on the Internet are through forums and in product reviews websites. It is important here also to distinguish the term "evaluation". An evaluative text is a text in which the author reports his/her experience with a certain product. Evaluations are also examples of opinionated texts.

As an example, the Chart 1 shows an evaluation of the iPhone 4 handset extracted from the website <www.buscape.com.br>.

Chart 1 – Excerpt for an evaluation of the iPhone 4 handset

> The iPhone 4 is certainly the best phone I've ever had, the design is very beautiful, it's made with a durable material that inspires confidence, I can play and listen to songs without fear that when I'm receiving some connection the phone don't hang.
> It has a very good camera, perfect for those who do not like to be carrying several devices, there are not many free applications but for a few dollars you can buy high quality applications in the AppStore.
> It's the best phone I've ever had without a doubt, I recommend it.

We note that the text provides several opinions about the iPhone 4 handset. These opinions are in its majority explicit direct opinions, such as: *"best cell phone I ever had"*, *"beautiful"*. In these evaluations, the author is motivated to provide his/her experience with the product to another future consumer. In this way, the author of the text gives a clearer and simpler possible vision of everything he/her believes to be positive and of everything he/her believes to be negative. In contrast, we display below other opinionated text extracted from the Folha de São Paulo newspaper website. It is an excerpt from the Trends/Debates column extracted from the Opinion section. The Chart 2 shows the original text in Portuguese followed by its translation into English.

Unlike the text in Chart 1, this article show only the opinion of the author and not an evaluation of a product or entity. Hence, it is an opinionated text but not an evaluation.

This work focuses on evaluations since the goal is to identify and measure the sentiment towards the aspects. The next section shows a brief literature review for the main topics in sentiment analysis.

Chart 2 – Example of an opinionated text that it is not an evaluation

> Você se sente vivendo num Estado policial?
> JÁ VIVI num Estado policial, a ditadura militar, que durou dos meus 14 aos 35 anos. Palavras como essas não podem ser banalizadas. Lembro que, em 1976, recém-retornado da França, onde fora estudar (sem ser exilado, só bolsista), recém-contratado professor da USP, me deparava toda semana com barreiras policiais na ponte da Cidade Universitária. Em algum momento de 1978, elas acabaram. Mas demorei para perceber que não havia mais essas blitze, que sempre estavam atrás dos supostos subversivos. E me perguntei por que não tinha notado o fim delas. A resposta que me ocorreu foi muito simples: é tão contra a nossa natureza, é tão fora de propósito viver sob a tutela de um Estado policial, que é mais fácil notar quando ele surge do que quando desaparece. A não ser, claro, que caia com estrondo, como o muro de Berlim. Mas a lenta, gradual e quase interminável redução do caráter policial de nossa ditadura passava até despercebida.
> Alguém pode comparar aquele tempo ao atual? Sim, há hoje recursos de controle que na época nem se imaginavam. Desenvolveram-se mecanismos de escuta que permitem captar conversas de quase todas as pessoas. O simples uso do cartão de crédito ou do celular permite retraçar os trajetos pela cidade, pelo país, pelo mundo. Mas, tudo isso somado, não é a mesma coisa que viver no confronto direto com o policial que pode prender você a qualquer momento, sem lhe dar razões ou satisfações.
> —
> Do you feel like living in a police state?
> I have ALREADY lived in a police state, the military dictatorship, which lasted from 14 to 35 years. Words like these can not be trivialized. I remember that in 1976, when I had just returned from France, where I had gone to study (without being exiled, I was only a scholar), I was a professor at the University of São Paulo.
> Sometime in 1978, they were over. But it took me a while to realize that there were no more blitze, who were always behind the supposed subversives. And I wondered why I had not noticed the end of them. The answer that occurred to me was very simple: it is so against our nature, it is so out of place to live under the tutelage of a police state, that it is easier to notice when it arises than when it disappears. Unless, of course, it crashes, like the Berlin Wall. But the slow, gradual and almost endless reduction of the police character of our dictatorship went unnoticed.
> Can anyone compare that time to the current one? Yes, there are today control resources that at the time could not even imagine. We have developed mechanisms of listening that allow us to capture conversations of almost all the people. The simple use of the credit card or the cell phone allows to retrace the routes by the city, by the country, by the world. But all in all, it is not the same as living in direct confrontation with the police officer who can arrest you at any moment without giving you reasons or satisfactions.

2.4. Sentiment Analysis

The study of sentiment analysis can be performed in three different levels: text level, sentence level, and aspect or entity level (LIU, 2010).

Document level: aims to classify documents based on the overall sentiment expressed in the document. At this level, it is assumed that the document is about only one topic or entity and the sentiment is usually classified as positive or negative (PANG; LEE; VAITHYANATHAN, 2002; TURNEY, 2002).

Sentence level: aims to classify each sentence based on the expressed sentiment. To do so, some authors establish a pre-classification to determine whether the sentence is objective (without sentiment) or subjective (with sentiment) (WIEBE; BRUCE; O'HARA, 1999). For subjective sentences, besides the positive and negative options, some authors also use a third class named neutral for sentences with no sentiment or undetermined sentiment (WILSON; WIEBE; HWA, 2004).

Aspect level: aims to determine the polarity of the sentiment expressed towards a particular aspect, entity or characteristic. This level is motivated by the fact that a sentiment is never expressed without taking into account a target object. Thus, understanding the importance of the target also helps to understand the opinion contained in the sentences and in the text. In this way, the aspect-based sentiment analysis is seen as a more refined analysis, where instead of looking at the textual constructions (text, sentence or locution), one looks directly at the opinion.

The Subsection 2.4.1 shows the initial works in the field of sentiment analysis. The objective of this chapter is to show the evolution of the field leading to the problems tackled by this dissertation.

2.4.1. Initial works

Although sentiment analysis encompasses many activities, the area began with few works with focus on lexicon generation and in sentence/text sentiment classification.

Lexicon generation

To compile a dictionary for opinion words, i.e., words with an associated sentiment, there are three possible approaches: manual, dictionary-based, and corpus-based. The manual approach consists of collecting and building the dictionary manually, which is a very time-consuming task. The dictionary-based approach uses a standard or domain speciific language dictionary in order to determine the polarity of the words. The corpus based approach uses the corpus to extract contexts which the words can be expressed positively or negatively.

One of the simplest techniques for the dictionary-based approach is reported by Hu and Liu (2004) and Kim and Hovy (2004). This approach is based on bootstrapping using a small set of seed opinion words and an online dictionary, e.g., WordNet (FELLBAUM, 1998). This strategy first collects a small set of opinion words and then uses WordNet to grow this set for their synonyms and antonyms. The new words are included in the dictionary and the process is restarted. The iterative process stops when no more new words are found. Usually, after this process, the words are verified and corrected by hand, finalizing the dictionary construction.

Some researchers have appointed that additional information like glosses in WordNet and additional techniques (e.g., machine learning) generates better lists (ANDREEVSKAIA; BERGLER, 2006; ESULI; SEBASTIANI, 2006a; ESULI; SEBASTIANI, 2006b). Despite the efficiency of these methods to construct polarity dictionaries, they are unable to generate opinion words with domain specific orientations. For this problem, a corpus-based approach represents a better solution.

The corpus-based approach relies on syntactic or co-occurrence patterns and a list of seeds to find other opinion words in a large corpus. One of the first works is described in Hatzivassiloglou and McKeown (1997), where the authors use a regression model to classify conjoined adjectives into positive or negative categories. They use a set of linguistic constrains or conventions on connectives to identify additional adjective opinion words and their orientations. For example, in the sentence *“this phone is beautiful and small"*, if *beautiful* is known to be positive, so *small* can be categorized as positive as well. Other connectives like *or*, *but*, *either-or* and *neither-nor* are used in the work.

Qiu *et al.* (2009) propose a propagation approach that exploits the relations between sentiment words and topics or product features that the sentiment words modify. The extraction rules are designed based on relations described in dependency trees.

Ding, Liu and Yu (2008) explore the idea of intra-sentential and inter-sentential sentiment consistency. They showed that the same word might have different orientations in different contexts. Their method determines opinion words and their orientations together with the object features that they modify.

Sentiment Classification

Hatzivassiloglou and McKeown (1997) present a work on the prediction of polarity, or semantic orientation, of adjectives. The authors used a large corpus-based approach from which syntactic patterns of co-occurrence among adjectives were extracted. A regression model to classify co-occurring adjectives as positive or negative based on these patterns. For example, in their method, if the unknown adjective “elegant” was found frequently in sentences like “nice and elegant”, with “nice” being known as positive, this would lead the word *“elegant”* to be also labeled as positive.

Pang, Lee and Vaithyanathan (2002) performed supervised classification in a movie reviews dataset[1] collect from the *Rotten Tomatoes Website*[2] which consists of 2000 movie reviews, where 1000 are positive labeled and 1000 are negative labeled. Pang, Lee and Vaithyanathan (2002) show that using unigrams, as a bag-of-words feature model, and the relative position for the words, the performance is higher than when they used either naïve Bayes, maximum entropy or SVM algorithms. The maximum accuracy obtained for their classifier was 81% in the movie

[1] Available at <http://www.cs.cornell.edu/people/pabo/movie-review-data/>
[2] <http://www.rottentomatoes.com/>

reviews dataset.

Wilson, Wiebe and Hoffmann (2009) present a further study with a more elaborated set of features for supervised machine learning. They present an exploratory study on features for phrase level sentiment analysis with the Multi-Perspective Question Answering (MPQA) corpus (WIEBE; WILSON; CARDIE, 2005). The listing below describes some of the features used in their work as well as in other supervised learning works:

- *Terms and their frequency*: words present in the text, as individual words or n-grams and their frequency counts. In some cases, word positions may also be considered. In this feature it is important to apply a selection filter like TF-IDF weighting scheme, that can distinguish the most valuable attributes for each class.
- *Part of speech tags*: part of speech tags are important indicators of subjectivity and opinions.
- *Negation*: negation and its scope are very important features. For example, sentences like *I don't recommend* have an opposite polarity to *I recommend*.
- *Opinion words and phrases*: opinion words can be inserted as characteristics to express each sentiment or class.
- *Syntactic dependency*: words dependency based features generated from parsing or dependency trees.

Taboada *et al.* (2011) use a lexicon method to determine the polarity, or semantic orientation, for the individual words in the text. This is based on the same linguistic concept used by the reader when it assesses a text. In this method, a classifier can simply averages the semantic orientations (i.e., the sentiment valence for each word) found in the text, or it can use semantic constitutionality to infer more elaborated constructions. Linguistic features like negation scope (e.g., not good), *irrealis* (e.g., could be good) and intensifiers (e.g., very good) could be easily addressed in this method by the incorporation of steps in the process. Some authors also prefer lexicon methods, in opposite to supervised machine learning methods, since the dictionary used is independent from domain (TABOADA *et al.*, 2011).

At the other side, we have the unsupervised learning algorithms. They differ from supervised methods in the way that they do not require a corpus with labeled examples, thus often making use of bootstrapping methods, i.e., a small seed of examples is given and the algorithm is able to retrieve by similarity other instances of training. Despite the robustness of this method, it is susceptible to semantic drifts so that, in the bootstrapping process, an example of positive instance is learned as negative or vice versa.

Turney (2002) presents an unsupervised classification of reviews for extraction fixed syntactic phrases that are likely to express opinions. The algorithm has four steps: extract phrases containing adjectives or adverbs; estimate the orientation of the extracted phrases using the pointwise mutual information (PMI) measure (CHURCH; HANKS, 1990); compute the average orientation of all phrases and classifies the review into positive or negative. In the end, these list is used as input for the system to classify new instances from the corpus.

Subsection 2.4.2 presents some initial works of sentiment analysis for Portuguese.

2.4.2. Sentiment Analysis for Portuguese

The initial works for Portuguese are related with the elaboration of corpus (SARMENTO *et al.*, 2009; CARVALHO *et al.*, 2009; SCOPIM *et al.*, 2012), sentiment lexicons (SILVA *et al.*, 2010; PASQUALOTTI, 2008; SOUZA *et al.*, 2011; FILHO; PARDO; ALUÍSIO, 2013) and methods for sentiment classification at the document level (AFONSO *et al.*, 2011; AMANCIO *et al.*, 2010; SOUZA; VIEIRA, 2012).

The Section 3.5 will approach in details the most important works for Portuguese in the context of aspect-based sentiment analysis.

2.5. Final Remarks

Despite the numerous papers for sentiment analysis, the great focus still on English language. Faced with this fact, few works focus on the transfer of knowledge between languages and few works are specific to the Portuguese language.

Chapter 3 will now focus in aspect-based sentiment analysis which is to field better covered by this dissertation.

CHAPTER

3

ASPECT-BASED SENTIMENT ANALYSIS

Distinctly from the text and sentence level approaches, the aspect-based sentiment analysis shows great challenges. In general, works at this level have two main challenges: **extraction of information** and **textual classification**. Through information extraction we want to identify the aspects and evaluations from the opinions present in a text. Through textual classification, we want to classify the evaluation towards the aspect based on the sentiment it expresses.

According to Liu (2010), the aspect-based sentiment analysis is composed of three main tasks:

Aspect extraction: This is the task responsible for extracting aspects and their modifiers. For example, in the sentence *"I like the quality of this television"*, the opinion target is *"quality"* while the word *"like"* is the modifier, transferring positive sentiment to the target. In some cases, it is possible to not have a specific target in the text, as in the example *"This television is great!"*. In this case, there is no particular aspect for the assessment but the entity as a whole. In these cases, we denote the aspect as GENERAL.

Group entity, aspects and modifiers: This task consists of grouping entities, aspects, and modifiers together. For example, we may have *"Apple phone"* and *"iPhone "*, both of them refering to the same entity. For aspects we may also have *"cost"* and *"price"*, and for modifiers a similar case would be *"pretty"* and *"cute"*. In all these cases it is important that the grouping method identifies that it is the same.

Sentiment classification: After extraction, it is necessary to classify aspects according to their modifiers. In this sentiment classification, it is usual a first step do determine whether it is polar (contains sentiment) or neutral (contains no sentiment or the sentiment is undetermined). In the case of containing sentiment, it will be further be classified either as positive or negative.

Some authors also add to these tasks a last step for the visualization and gathering information.

To better understand the proposed steps for aspect-based sentiment analysis, the Figure 3 shows the usual architecture for an aspect-based sentiment analysis system.

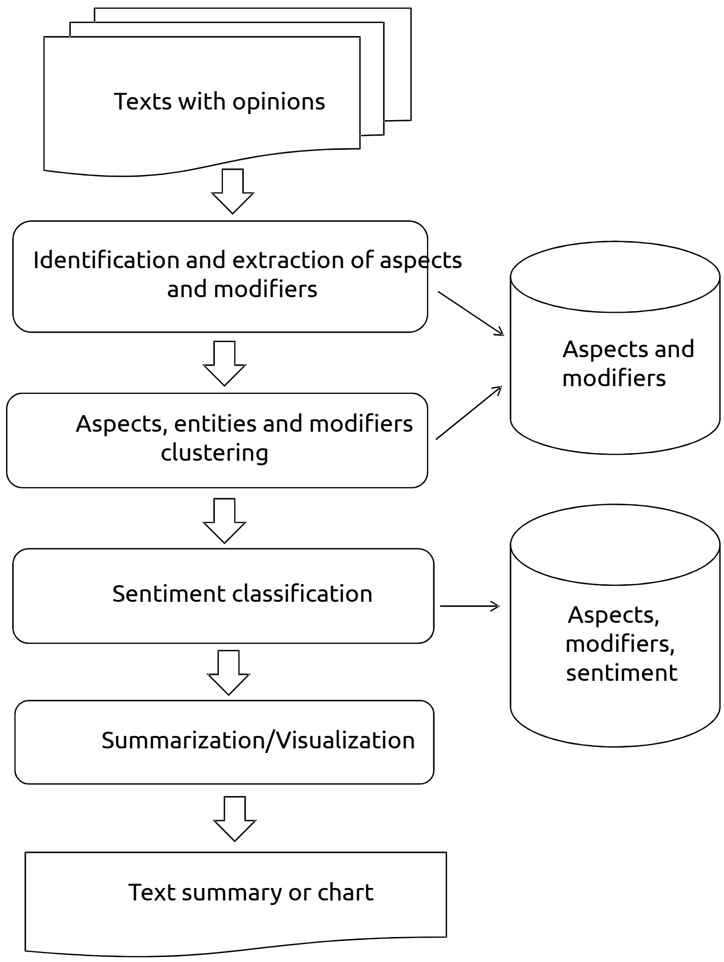

Figure 3 – Usual steps for the aspect-based sentiment analysis

3.1. Aspect Extraction

This is the first step of the aspect-based sentiment analysis and it could be categorized as an information extraction task. Formally, according to Liu (2010), this task aims to extract quintuples from the text in the form of $(o_j, f_{jk}, oo_{ijkl}, h_i, t_l)$ where o_j is an object, f_{jk} is a feature of the object o_j, oo_{ijkl} is the semantic orientation or polarity of the opinion on feature f_{jk} of object o_j, h_i is the opinion holder and t_l is the time when the opinion is expressed by h_i.

There are different techniques for aspect extraction. Liu (2012) shows a categorization based on its approaches.

1. Frequency-based
2. Relation-based
3. Machine learning
4. Topic modeling

3.1.1. Frequency based aspect extraction

This approach for extraction of explicit aspects is based on the fact that users are likely to use the same vocabulary when referring to aspects and qualities (LIU, 2012). For example, if a text comments on a digital camera, nouns and noun phrases such as *"image quality"*, *"sharpness"*, *"price"* should often be used across several evaluations in this same domain.

Hu and Liu (2004) shows a standard approach for this method. After the part-of-speech tagging, the authors extract the most frequent nouns and noun phrases as candidate for aspects. A cut-off threshold is empirically proposed by the authors.

Popescu and Etzioni (2005) improved the method presented by Hu and Liu (2004) removing candidates that could not be related to the analyzed aspect. To make this decision, the authors measured the Pointwise Mutual Information (PMI) between the aspect and the candidates. If the measure returns a low co-occurrence value, the candidate is then discarded.

Blair-goldensohn *et al.* (2008) refined the candidates to the ones which obey certain syntactic patterns. The authors also included filters and normalization that led to the improvement of the method.

Long, Jie and Xiaoyan (2010) extract nouns based on information frequency and distance. The work first uses the frequency to determine the main aspects, then uses the distance information to find other words related to the aspects. This method can infer, for example, that the aspect price is related to the currency symbol and also to the currency word.

3.1.2. Relation-based aspect extraction

This approach is known to explore the relationship between opinion and its targets. This method often acts as a complement to the frequency-based extraction method. In Hu and Liu (2004), for example, infrequent aspects were extracted by the relationship with words of sentiments. Thus, in the example, *"The screen is great"*, the adjective *"great "* establishes a qualifying relationship with the word *"screen"*. In Zhuang, Jing and Zhu (2006), a dependency parser was used to indicate the relationship between aspects and their targets. The works presented by Kobayashi, Inui and Matsumoto (2007), Somasundaran *et al.* (2009) and Kessler and Nicolov (2009) are similar to Zhuang, Jing and Zhu (2006) and also used a dependency parser approach.

3.1.3. Machine learning for aspect extraction

This approach explores the use of already established algorithms for extracting information. The most dominant method is sequential learning.

There are two known methods for sequential learning: hidden Markov models, which Jin, Ho and Srihari (2009) introduces the concept for aspect extraction by appling a hidden and lexicalized Markov chain to learn patterns to extract expressions of aspects and opinions; and Conditional Random Field (CRF), which Jakob and Gurevych (2010) demonstrated to be more effective than Markov models for the extraction of aspects and opinions.

The most proeminent works in this approach will be better covered in Section 3.6.

3.1.4. Topical modeling for aspect extraction

Extraction using topical modeling is based on the principle that each document has several topics and that each topic has its own probability of distribution over its words. Extraction using topical modeling is basically a form of unsupervised machine learning. There are two basic topic models: Probabilistic Latent Semantic Analysis (pLSA) (HOFMANN, 1999) and *Latent Direchlet Allocation* (LDA) (BLEI; NG; JORDAN, 2003).

Mei *et al.* (2007) propose a joint model based on pLSA for the extraction of aspects and sentiments.

Lin and He (2009) propose a joint model where sentiment and predicates are extracted as a single entity.

Li, Huang and Zhu (2010) proposed two joint models for extraction, one for sentiment and the other for aspects that depend on this sentiment.

In Titov and McDonald (2008), the authors argue that topical modeling such as LDA may not be good for detecting aspects. The reason the authors found is that product reviews have a lot of homogeneity and high co-occurrence between words. Thus, the topical modeling techniques can not statistically differentiate the aspects present in the document. Liu (2012) states that techniques based on topical modeling, while conceptually and mathematically interesting, are noteworthy to use in purely statistical fashion. Liu (2012) says that a more natural language approach is needed.

3.2. Clustering aspects

After identifying and extracting the aspects from the text, it is necessary to group them and relate them. Thus, a second step is to find synonyms and links between concepts so that the analysis of the text can be complete.

Popescu and Etzioni (2005) use a web-based method to discover hyperonyms and hyponym. They look in online search engines for expressions associating the word with phrases like "part of" or "has". Guo *et al.* (2009) use multilevel latent semantic analysis to group aspects into groups. Zhai *et al.* (2011) use some information for grouping expressions into the appropriate categories for each aspect. Examples are: lexical similarity by WordNet (FELLBAUM, 1998); similarity of the distribution of words in the corpus; syntactic constraints.

Mukherjee and Liu (2012) use a semisupervisioned model for grouping similar aspects using contextual information of co-occurrence of these terms.

3.3. Sentiment classification

After the identification and grouping of similar aspects, it is necessary to classify the associated sentiment. For classification of sentiment there are two main approaches: classifiers based on machine learning and classifiers based on the use of lexicon. The lexical-based method uses a dictionary of terms and their respective polarities, also known as semantic orientations. This method calculates the polarity of a document, phrase, or resource based on the number of positive or negative terms in the text. The approach based on machine learning can be supervised or unsupervised. The supervised method uses a training corpus with labeled examples to learn the specific lexicon of each sentiment class in order to construct a classification model. The unsupervised method uses a corpus of unlabeled examples to cluster by similarity examples of each sentiment classes.

Jiang *et al.* (2011) discuss in their work the importance of associating the learned lexicon with the aspects it modifies, thus facilitating the disambiguation of meanings in the classification by machine learning. Ding, Liu and Yu (2008) present a lexical-based method and explore the evidence and linguistic conventions that modify the original semantic orientation of this lexicon. Such linguistic conventions lead to an analysis of compositional semantics, in which several linguistic factors can influence the *a priori* polarity of words.

3.4. Visualization and summarization

Finally, the final step in the sentiment analysis at the aspect level is the summarization of sentiment and opinions. This step is seen as a multi-document summarization problem. In this step, it would be possible to apply classical methods based on content selection for the composition of a textual summary. However, content selection only allows an informative summary of the text. For a quantitative assessment (e.g., "60 % of users liked product price"), other approaches are required. In this way, the aspect extraction and sentiment classification play a major role, helping the system to structure the information in order to compile it textually or graphically to the user. Examples of interfaces for summarizing aspects are presented by Liu, Hu and Cheng (2005), Lerman, Blair-Goldensohn and McDonald (2009), Hsieh *et al.* (2012) and Filho, Brun and Rondeau (2012).

Section 3.5 will show some works in aspect-based sentiment analysis for Portuguese language.

3.5. Aspect-based sentiment analysis for Portuguese

Carvalho *et al.* (2011) describes the construction of a corpus in the political domain with the annotation of opinions and their targets. The OPTIMISM (SILVA *et al.*, 2009) system was developed from research with this corpus. Although the system does not provide a complete analysis at the aspect level, it performs the recognition of named entities and uses a political domain ontology to connect them. A parallel work was applied in the 2011 elections in Portugal (AFONSO *et al.*, 2011). In this work, the online Twitometer tool measures the expression of positive, negative and neutral sentiment in texts of the online Twitter tool regarding candidates for the election.

Chaves *et al.* (2012) present the PIRPO tool for rating sentiment in online evaluations for the hotel sector. The authors use an ontology in the hotel domain and a lexicon of sentiment (SOUZA *et al.*, 2011) for extracting and classifying aspects contained in the text. Its approach is simple and uses only the direct comparison of terms present in the text with the terms present in the ontology. If a term is found, the words around it that match those present in the dictionary of sentiment are used to compute the semantic orientation, or polarity, of this aspect.

The work described by Ribeiro *et al.* (2012) presents a polarity classification system for aspects in vehicle evaluation texts. In this paper, the authors compare lexical-based classifiers with classifiers based on machine learning. In the evaluation of the authors, classifiers based on machine learning have a better performance and require less effort in relation to lexical-based methods. The extraction of aspects was not covered by the work.

Fernandes (2010) presents the HowGood tool that performs the analysis of sentiment at the aspect level in the Portuguese language. The system determines the aspects present by the most frequent nouns. In this system, the user must manually filter the aspects that interest him as well as assign the polarity to each found predicate. Silva (2010) made an adaptation of this work for the production of the BestChoice system. In this system, the polarity of words is determined by the lexical feature SentiWordNet (ESULI; SEBASTIANI, 2006b) available for English language. Authors use the Google Translate tool to translate each term into English. The main limitation of these works is in the user's need to filter the aspects returned by the system.

Siqueira and Barros (2010) present a process of extraction of aspects in the analysis of sentiment for texts in Portuguese language in the e-commerce domain. The WhatMatter system, described by them, performs four steps: identifies frequent nouns, identifies relevant nouns, maps aspect indicators, and removes unrelated nouns. The first two steps identify nouns by frequency and occurrence with adjectives. The third step focuses on the acquisition of implicit aspects. For these, the authors used a manually compiled list of 20 indicators of particular aspects of that domain. According to the authors, there is no automatic way available to generate this list. In the fourth, and last step, the authors measure the PMI-IR (TURNEY, 2002) to filter out infrequent nouns. The threshold value for the PMI-IR measure was determined empirically. In the evaluation, the authors used a corpus of 200 opinions and obtained an accuracy of 77.24%, coverage of 90.94% and F-measure of 83.54%, which are considered very good values. The deficiency of this method is the need to indicate specific lists and thresholds for the work domain.

3.6. State-of-the-art

According to Pontiki *et al.* (2014), the activity of aspect-based sentiment analysis is analyzed for different tasks on different datasets from different perspectives. These variety of forms makes difficult to define what the activity is and what is the state-of-art. Thus, these authors then propose a joint assessment to be held in Semantic Evaluation workshop (SemEval) for the activity of aspect-based sentiment analysis. According to the authors, the activity comprises four subtasks:

1. **Aspect Term Extraction**: given a set of sentences, the task aims to identify all terms / objects present in the sentence that are aspects for the entity that is being qualified. For example, in the following review "The service is excellent!", the term "service" should be extracted, because it expresses an aspect of the restaurant that is being evaluated.

2. **Aspect Term Polarity**: given the set of sentences and identified aspects for the subtask above, this task aims to determine the polarity assigned to each aspect, which may be positive, negative, neutral or conflicting. As an example, in the same sentence "The service is excellent!", The term "service" previously identified as aspect should have its polarity assigned to positive due to the use of the qualifier term "excellent".

3. **Aspect Category Detection**: given a pre-defined set of categories and a set of sentences without any annotation, this task is to identify which categories are involved in the sentence. For example, the sentence "I liked the lasagna and rice" could be classified in the category "food".

4. **Aspect Category Polarity**: given the set of identified sentences and categories in subtask above, this task aims to determine the polarity assigned to each category, which may be

positive, negative, neutral or conflicting. Again, in the example "I liked the lasagna and rice", the category "food" could be classified as positive.

Subtasks 1 and 2 enter a higher level of detail in the text, since they identify the terms of the sentence that qualify as aspects of the analyzed entity. Subtasks 3 and 4 are evaluations at a higher level, in which it is not important to explicitly define the aspect, but to identify what each sentence addresses.

The state of the art in the area is mainly found in the works of SemEval's 2015 (PONTIKI *et al.*, 2015) and 2016 (PONTIKI *et al.*, 2016). The task entitled Opinion Target Expression (OTE), where systems were asked to identify the aspect mentioned within the sentence, attracted 14 teams in 2015, with the best system reaching the f-score of 70.05%; and 29 teams in 2016 with the best system reaching the **f-score of 72.34%**

The EliXa (VICENTE; SARALEGI; AGERRI, 2015) system was best scored in the OTE task in the 2015 edition, with 70.05% for f-score. This system used machine learning based on the *Averaged Perceptron* (COLLINS, 2002) algorithm with the following machine learning characteristics: n-grams; part-of-speech label; n-grams of prefixes and suffixes; Brown clusters (BROWN *et al.*, 1992) and word embeddings (MIKOLOV; DEAN, 2013).

The NLANGP (TOH; SU, 2015) system, the second highest score for the OTE task in the 2015 edition, reached the f-score of 67.11%. This system was based on the machine learning algorithm Conditional Random Fields (LAFFERTY; MCCALLUM; PEREIRA, 2001) with the following characteristics: the word itself; the head of the syntactic constituents (obtained from a dependency parser); lists of names (extracted base on frequency from a corpus); and Brown clusters.

In the 2016 edition of SemEval, the NLANGP (TOH; SU, 2016) system was the best system in the OTE task with the f-measure score of 72.34%. The system, an enhancement of the same system that won the second place in the previous edition of the competition, brings the addition of a new learning feature based on the extracted probability of a recurrent neural network. The system AUEB (XENOS *et al.*, 2016), second in the OTE task in the 2016 edition, reached the score of 70.44% through a system based on the algorithm *Conditional Random Fields* with the following set of characteristics: part-of-speech tags; lexicon (KARAMPATSIS; PAVLOPOULOS; MALAKASIOTIS, 2014), list of aspects and word embeddings.

3.7. Final Remarks

Despite the many existing work on sentiment analysis at the aspect level, many challenges and limitations still need to be addressed. The main issues that arise in this area are related to discovering aspects without the need to use resources compiled manually for the domain.

Few papers also explore the use of linguistic knowledge in the extraction and classification of aspects. In general, most research is focused on new algorithms of machine learning rather than modeling the linguistic phenomena that occur when an opinion is expressed.

The focus of this dissertation is to investigate methods for the first three approaches reported on Section 3.1, frequency-, relation- and machine learning-based methods for Portuguese language, proposing variations and enrichment for these methods (using syntactic and semantic knowledge) and comparing them with the actual state-of-art. The decision to not explore methods based on topical modeling was made because these methods present great difficulty in the insertion of linguistic knowledge and, in general, do not present results close to the state-of-art (MOGHADDAM, 2013).

CHAPTER

4

DATASETS AND TOOLS

4.1. Datasets

For the Portuguese language, the focus of our work, we use the corpus **ReLi** (FREITAS *et al.*, 2012). [1]. The reviews from ReLi have been extracted from the website www.skoob.com, a social network of books and readers. On this site, readers freely write opinions about the books they have read.

The corpus consists of 2,056 reviews. The annotations include manual sentence and aspect level sentiment analysis of 13 different books with about 200 comments each. The corpus contains 300,000 words and 15,000 sentences that have been annotated with part-of-speech labels and chucking. The annotations include manual sentence and aspect level sentiment analysis. The corpus has 4,210 positive opinions and 1,024 negative opinions. At the sentence level, the corpus has 2,883 positive sentences and 596 negative ones. Figure 4 shows the annotation for the sentence *"Eu não me apaixonei pela história"* ("I did not fall in love with the plot") extracted from the corpus.

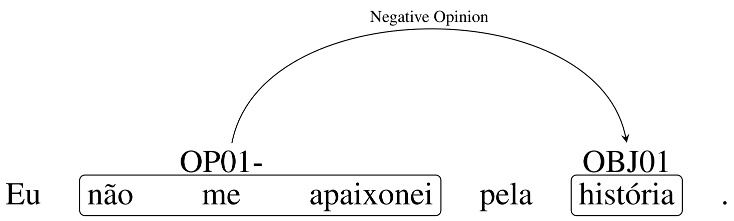

Figure 4 – Example of an excerpt from Reli corpus

In this sentence, the word *"história"* ("plot") labeled as OBJ01 is the object or aspect of the evaluation. The number 01 distinguishes this aspect from other aspects evalued in the same text which receive a incremental numbering. The excerpt *"Eu não me apaixonei"* ("I did not fall

[1] Available at <http://www.linguateca.pt/Repositorio/ReLi/>

in love”) labeled as OP01- is a negative opinion (minus signal) connected to the aspect *“história”* by the same number.

According to Freitas *et al.* (2012), the main difficulty in the process of annotating the corpus was to distinguish subjective information from factual information. One of the most recurrent phenomena of this type of evaluation is the description of the characters. In many of these cases, the researchers interpreted this as subjective and peripheral information to the opinion of the book itself. In the annotation, the researchers report that they adopt a conservative approach.

Taking only a subset of 6,000 sentences (850 reviews), the researchers reanalyzed their content and reported some statistics. According to them, 26% of these sentences contain opinion, 76% of them being positive. 32.5% of sentences expressing opinion contain contrasting opinions, that is, positive and negative aspects of the book in the same sentence. In 18% of sentences the opinion can not be identified by some word but by the expression of the whole sentence. Freitas *et al.* (2012) present a table with the distribution of expressions of opinion according to the number of words, which is displayed in Table 1.

Table 1 – Distribution of expressions of opinion according to the number of words (FREITAS *et al.*, 2012).

N-gram size	Frequency
1-3	69%
4-6	15%
7+	15%

Freitas *et al.* (2012) also report other phenomena found in the text. According to the authors, there are several words or expressions whose polarity varies according to the context. Thus, the word *“different”* can be both positive and negative. Expressions with diminutive can also indicate both a positive and a negative judgment. In some sentences, the polarity of the sentiment demands a knowledge of the context, for example, in the phrase *“E no final você vai ser odiada, pode ter certeza!”* (“And in the end you will be hatred, you can be sure!”), which in context is associated with a positive judgment on the book. Finally, the researchers also verified the presence of neologisms *“Bella é muito tonga”* (“Bella is very tonga/*dummy*”), typical expressions of the internet (*“rá!”*, *“Nah neh Noh ”*) and emoticons, as well as curse words and misspelled phrases.

4.1.1. Other available corpora

There are also other corpora produced for aspect-level opinions for Portuguese. For example, Ribeiro *et al.* (2012) reports the construction of a corpus with identified aspects in the vehicle domain; Siqueira and Barros (2010) use a corpus with 200 opinions with annotated aspects in the online shops domain; Carvalho *et al.* (2011) present a series of political debates with annotated entities.

Although studies with these corpora may be possible, only ReLi is freely available for access and research at the moment. In addition, the corpus of Ribeiro *et al.* (2012) and Siqueira and Barros (2010) are presented only in the context of their experiments and not as a resource built to be made available. In Carvalho *et al.* (2011), the constructed corpus focuses on a very specific purpose and therefore does not provide all the desired possibilities for this research.

4.1.2. Language Processing Tools

The syntax parser PALAVRAS (BICK, 2000) allows automatic parts-of-speech tagging and syntactical analysis of texts in Portuguese. The analyzer outputs a tokenized text, the part-of-speech annotations, a dependency parsing of each sentence and a semantic type for some words.

The system is constructed from a lexicon of 50,000 lemmas and thousands of grammatical rules using the grammatical formalism of Constraint Grammar (CG). The parser claims to achieve an accuracy level of 99% in terms of morphology, and 97-98% in terms of syntax (BICK, 2000).

There are also other syntactic parsers available for Portuguese, however for the purpose of this research the parser PALAVRAS was the most indicated. This parser also is well known and generally accepted by the scientific community.

The Chart 3 shows an excerpt from the ReLi corpus in XML format annotated with syntactic and semantic information by PALAVRAS. The information obtained from the parser PALAVRAS are: lemma, morphological flexion, part-of-speech tag, head of the dependency relation, dependency function, lexical semantic attributes and the semantic role.

Semantics is a very broad area of linguistics. In this monograph, we will limit ourselves to talk about semantic role labeling, since they are a practical application and with good results in NLP tasks.

There are two main works that report semantic role annotation tools for Brazilian Portuguese. The work of Alva-Manchego and Rosa (2012) presents a semantic role writer trained by supervised machine learning. In Fonseca and Rosa (2012), the authors present a annotator using a connectionist approach.

Chart 3 – Example for the ReLi corpus in XML annotated with PALAVRAS parser

```xml
<sentence id="0:0:1" place="title" polarity="neutral">
<text>Pode existir um livro bom sem uma história boa.</text>
<tokens>
<word id="1" form="Pode" base="poder" postag="v-fin" morf="PR 3S IND VFIN" extra="fmc * aux" head="0" deprel="STA" srl="" obj="O" opinion="O" from="0" to="4"/>
<word id="2" form="existir" base="existir" postag="v-inf" morf="--" extra="mv" head="1" deprel="Oaux" srl="PRED" obj="O" opinion="O" from="5" to="12"/>
<word id="3" form="um" base="um" postag="pron-indef" morf="M S" extra="--" head="4" deprel="DN" srl="" obj="O" opinion="O" from="13" to="15"/>
<word id="4" form="livro" base="livro" postag="n" morf="M S" sem="sem-r" extra="--" head="1" deprel="S" srl="TH" obj="O" opinion="O" from="16" to="21"/>
<word id="5" form="bom" base="bom" postag="adj" morf="M S" extra="np-close" head="4" deprel="DN" srl="" obj="O" opinion="O" from="22" to="25"/>
<word id="6" form="sem" base="sem" postag="prp" morf="--" extra="--" head="2" deprel="fA" srl="" obj="O" opinion="O" from="26" to="29"/>
<word id="7" form="uma" base="um" postag="pron-indef" morf="F S" extra="--" head="8" deprel="DN" srl="" obj="O" opinion="O" from="30" to="33"/>
<word id="8" form="história" base="história" postag="n" morf="F S" sem="per domain sem-r" extra="--" head="6" deprel="DP" srl="COM-ADV" obj="O" opinion="O" from="34" to="42"/>
<word id="9" form="boa" base="bom" postag="adj" morf="F S" extra="jh np-close" head="8" deprel="DN" srl="" obj="O" opinion="O" from="43" to="46"/>
<word id="10" form="." base="--" postag="pu" morf="--" extra="--" head="0" deprel="PU" srl="" from="46" to="47"/>
</tokens>
</sentence>
```

CHAPTER 5

FREQUENCY- AND RELATION-BASED ASPECT EXTRACTION

5.1. Frequency-based methods

The simplest method for the problem of extraction of aspects is based on the most frequent aspects. This method consists of checking the aspects that appeared in a training corpus and then in the subsequent identification of these in the test corpus. A cut-off threshold can be applied in order to only identify the aspects within a frequency range.

The method follows a simple algorithm, which is based on the premise that the main aspects are the most commented by the users/writers of opinions. This method serves as a first baseline method for checking more sophisticated models. Also, it allows the exploration for linguistic knowledge in the domain. Evidence of its relevance (despite the simplicity) is that this algorithm was also used as **baseline** in the joint evaluation SemEval for identifying aspects in texts (PONTIKI *et al.*, 2014).

In Table 2, we present the results for four frequency-based methods evaluated in the ReLi corpus.

The first method persists all aspects present in the training corpus and then performs the identification of these in the test set. The second method is similar to the first, but with the removal of *stopwords* from the persisted aspects.

In the third method, we verified several cutoff thresholds related to the aspect frequency in the training set. Here, we use the relative frequency as a function of the number of occurrences of the word in the set of all aspects mentioned in the training set. The best result was for cutting at 7% of frequency, which in our training set represents aspects mentioned at least 40 times.

The fourth method is a variation of the previous one. We check the relative frequency of the aspect term in relation with the whole vocabulary of the training set. Thus, we seek to only recognize aspects with a high degree of probability to be related to aspects and not to common words of the text. The best result was found when we cut 65% of the relative frequency.

Table 2 – ReLi Aspect Extraction Scores for Frequency Methods

Run	ReLi corpus – Aspect Extraction		
	Precision	Recall	F-measure
All aspects	7,14%	**82,26%**	13,13%
Stopwords cut	14,32%	79,11%	24,25%
Frequency cut	30.27%	55.51%	39,17%
Relative frequency cut	**36,44%**	78,25%	**49,73%**

Table 3 and Table 4 show the results for the same four methods reported in Table 2 replicated for the corpus SemEval ABSA 2015 and SemEval ABSA 2016, both for the English language, respectively.

We observed from the results that methods 1, 2 and 3 achieved identical results in both corpora. This is due to the fact that no stopword is noted as aspect and also that the best cut by frequency is selecting all aspects (no cut). The best cut for method 4, relative frequency, was 30% for the SemEval ABSA 2015 and 45% for the Semeval ABSA 2016 corpus [1].

Table 3 – Extraction results of aspects in the SemEval ABSA 2015 corpus using frequency based methods

Method	Precision	Recall	F-score
1. All aspects	50,81%	**58,12%**	54,22%
2. Stopwords cut	50,81%	**58,12%**	54,22%
3. Frequency cut	50,81%	**58,12%**	54,22%
4. Relative frequency cut	**61,34%**	55,28%	**58,15%**

Table 4 – Extraction results of aspects in the SemEval ABSA 2016 corpus using frequency based methods

Method	Precision	Recall	F-score
1. All aspects	50,88%	**62,31%**	56,02%
2. Stopwords cut	50,88%	**62,31%**	56,02%
3. Frequency cut	50,88%	**62,31%**	56,02%
4. Relative frequency cut	**60,35%**	58,77%	**59,55%**

In the evaluation of the results, we observed that the best f-score results for the SemEval ABSA corpus were slightly higher (59.55%) than the results in the ReLi corpus (49.73%). We observed that the ReLi corpus is more complex in relation to the SemEval corpus, since the

[1] A complete breakdown of the results and parameters of the algorithm can be found in <https://github.com/pedrobalage/ABSA_Experiments/tree/master/3_FrequencyBased>

annotation of opinions is more refined and differentiates several factors that are not considered in the English corpus. For this reason, the challenges to the methods are greater and the results for Portuguese tend to be lower. As evidence of the complexity differences between the corpus, we show in Table 5 and Table 6 the distribution of the 10 most frequent aspects in the corpus ReLi and SemEval ABSA 2015, respectively.

Table 5 – Distribution of the gold annotated aspects in the ReLi corpus

Extracted aspects	Word frequency in Corpus	Relative Freq. word in corpus	Freq. as aspect	Relative Freq. as aspect
livro (book)	2779	1.07%	916	33.0%
história (story)	864	0.33%	208	24.1%
leitura (reading)	409	0.16%	112	27.4%
personagens (story characters)	321	0.12%	85	26.5%
crepúsculo (Twilight)	260	0.10%	62	23.8%
narrativa (narrative)	141	0.05%	61	43.3%
final (final)	193	0.07%	57	29.5%
romance (romance)	274	0.11%	55	20.1%
obra (book)	251	0.10%	48	19.1%
ele (him)	1053	0.40%	43	4.1%

Table 6 – Distribution of the gold aspects in the SemEval ABSA 2015 corpus

Extracted aspects	Word frequency in Corpus	Relative Freq. word in corpus	Freq. as aspect	Relative Freq. as aspect
food	190	1.03%	158	83.2%
service	127	0.69%	117	92.1%
place	135	0.73%	82	60.7%
restaurant	82	0.44%	29	35.4%
staff	33	0.18%	27	81.8%
pizza	42	0.23%	26	61.9%
atmosphere	26	0.14%	21	80.8%
sushi	32	0.17%	20	62.5%
decor	19	0.10%	16	84.2%
ambience	13	0.07%	13	100.0%

We observed that, in general, the relative frequency of the word presented in the text as being labeled as aspect (last column) is significantly higher in the SemEval ABSA corpus with respect to the ReLi corpus. This indicates that the words labeled as aspect in the ReLi corpus are used in a more varied manner and therefore lead to greater difficulty in the extraction process. For comparison to the state of the art, only possible in the SemEval ABSA corpus, the best system in the 2015 competition obtained a f-measure of 70.00% (VICENTE; SARALEGI; AGERRI, 2015). This observation leads us to consider that, even though they are simple methods, the frequency-based methods have good results.

Frequency-based methods can offer a good starting point, but there is still much that can be improved. In particular, we believe that the ocurring context of the instance can make a difference. To do so, relationship-based methods try to identify patterns of occurrence of aspects, based on existing syntax dependencies (hence the name of this class of methods). Investigations conducted in this line are reported below.

5.2. Relation-based methods

One of the classic methods that introduce the use of **templates in extracting aspects** of texts is presented in (NASUKAWA; YI, 2003). In this work, the authors identify the main **verbs** present in the corpus and categorize them according to the type of relation they exert with the subject and the predicate. For example, the verb “admire” transfers a positive sentiment to the evaluated aspect; The verb “accuse” transfers a negative sentiment; the verb “to provide” only transfers the subject’s sentiments to the predicate, whatever it may be. Nasukawa and Yi (2003) performed the classification of these patterns/templates entirely manually.

Unlike the original proposal, we performed the automatic learning of the templates, learning these from the training section of the ReLi corpus. In this way, the syntactic analysis and template extraction is done automatically from statistics of the corpus. This makes the method more scalable and easier to apply and, according to empirical observation, without compromising the quality of the extracted patterns. This was one of the contributions of this work.

In this work, the extraction of templates involving the verb, the aspect, the predicate and its assigned polarity was taken as a basis. We considered all the verbs contained in the training corpus that could form patterns according to the original work of Nasukawa and Yi (2003). We do not restrict here the number of templates to be learned.

Examples of learned templates are presented in Table 7, for two sentences of simple opinions. The proposal is that, whenever these standards are recognized in new data, **pattern matching** will identify the aspects of interest. Although we focus on the extraction of aspects, it is interesting to note that the templates allow us to go further, identifying the polarity and opinion associated with the aspect, if that is of interest.

Table 7 – Templates extracted from the ReLi corpus according to Nasukawa and Yi (2003)

Sentence	O livro é bom	Eu adorei o livro
Learned Patterns		
Verb	*ser*	*adorar*
Aspect	*livro*	*livro*
Opinion	*bom*	*adorar*
Polarity	*positiva*	*positiva*

Table 8 report the results for the Nasukawa and Yi (2003) method, only for the ReLi corpus, where we test two different ways to formalize the seized templates. In the first one, we use the lemmatized word, where we observe a very large specification of the rules. In the second, we used the part-of-speech label for pattern matching. In this attempt to generalize, we ended up getting worse results, and this variation did not prove advantageous.

Table 8 – Extraction results of aspects in the ReLi corpus using the method of Nasukawa and Yi (2003)

Method	Precision	Recall	F-score
Templates with lemma	**22,60%**	**17,85%**	**19,94%**
Templates with PoS	8,30%	17,85%	11,33%

What we observed in the learned templates is that they were very specific to the lexicon and led to a low range of coverage. In total, 2028 templates were extracted from the training corpus, but only 19 of these had a frequency greater than 2. For testing, we used all templates, because, given the low frequency of occurrence, it was not interesting to make any restrictions.

This verb-based method presents the advantages of using the context to extract the aspect and to make explicit the linguistic knowledge from the syntactic relations in the sentences. However, due to the fact that very specific rules are generated, this method does not obtain good results. It is important to emphasize that here, unlike the method of frequent aspects, the classification occurs not only in the aspects, but in the aspect and opinion pair.

Nasukawa and Yi (2003) method makes use of the annotation of the predicate aspect modifier, which is not provided to us in the SemEval ABSA 2015 and 2016 corpus. Therefore, it was not possible to replicate this same adaptation of the algorithm to the English language.

Using the Nasukawa and Yi (2003) method, we explore the use of the verb - aspect - predicate relation in the extraction of templates for later extraction of aspects. As we have seen, the attempt to insert more knowledge in this method ends up leading to a major generalization and thus to worse results. To circumvent this problem, we believe that we must insert not only a specific characteristic, but a set of characteristics that help in the elaboration of more complex templates. For this, the best approach to be used may be machine learning. With methods of this approach, we can feed the learning algorithms with sets of characteristics extracted from the text and let the algorithm itself solve all the complexity of deciding the set of characteristics most important for each case. In addition, the machine learning approach represents the current state-the-art in the task.

5.3. Final Remarks

Frequency-based methods may offer a good baseline, but also with some limitations. In order to explore more the relationship between the aspect and the opinion, the relation-based methods offer a good approach in constructing templates of co-occurrences. However, as verified, these templates are also very limited in capturing the contexts which ocurr aspects and opinions, mainly due the diversity of these constructions.

The next chapter will present the machine learning approach where we will explore in more detail the introduction of syntactic and semantic knowledge.

CHAPTER

6

MACHINE LEARNING-BASED ASPECT EXTRACTION

The methods based on machine learning allow to feed the learning algorithms with sets of features extracted from the text and let the algorithm itself solve all the complexity of deciding the most important set of characteristics for each case. Thus, we take out the complexity of learning how to identify patterns to extract aspects in the text and leave it to the learning algorithm, which incorporates these patterns through a **statistical modeling**.

These methods represent the current state of the art, as evidenced by the results achieved in the SemEval ABSA 2015 (PONTIKI *et al.*, 2015) and 2016 (PONTIKI *et al.*, 2016) shared tasks. In this way, this dissertation focus on this class of methods.

The learning algorithms that best fit the problem of aspect extraction are the algorithms belonging to the class of **sequential learning**. Examples of tasks within this class are: part-of-speech tagging (SILFVERBERG *et al.*, 2014), shallow parsing (SHA; PEREIRA, 2003), entity recognition (FINKEL; GRENAGER; MANNING, 2005), among others.

There are two algorithms better known in the sequential learning algorithm class: Hidden Markov Model (HMM) (BAUM; PETRIE, 1966), which is a generative Baysian learning algorithm; and Conditional Random Fields (CRF) (LAFFERTY; MCCALLUM; PEREIRA, 2001), which is a discriminative learning algorithm. This class of algorithms is based on the idea that the decision to classify a new instance is linked to the previous classifications and the presented context. Thus, an aspect is not only classified in isolation, but rather taking into account the context and sequence of tokens being analyzed.

In the methods proposed in this section, we use the CRF algorithm to predict the classification of each word as an aspect or not, which is also used in several works of aspect extraction and present in many systems from SemEval (PONTIKI *et al.*, 2016).

The positive point of using a machine learning algorithm lies in the fact that it is only necessary to provide the desired characteristics and the algorithm will learn what is statistically more important. Also, in the class of sequential labeling algorithms, such as CRF, the relationship between the sequence of words within a text is also computed. Thus, **contextual relations** are implicit in the use of this algorithm. A deficiency of these methods is in the use of many characteristics, which leads to a low performance due to the increase of space vector for decision. There are also limitations on modeling linguistic knowledge as it is necessary to send information word by word. The biggest challenge, therefore, resides in what is called **feature engineering**, where one must properly model the task to select the most relevant features.

Next, experiments are reported using the CRF algorithm through the framework **CRF-Suite** (OKAZAKI, 2007), used in conjunction with the machine learning library **Scikit-Learn** (PEDREGOSA *et al.*, 2011).

The experimental approach adopted was to test a set of features and check the impact of each in the system performance. Within the various possible features that might be extracted from the text for modeling in machine learning, we seek to extract those that have already been verified in the best systems reported in SemEval and those that are the focus of this research (namely, syntactic and semantic characteristics). In particular, due to constrains in the sequential labeling approach, our modeling of the features was performed token by token.

Next, we discuss the features evaluated in our experiments, explaining the way we believe that these would contribute to the understanding of the problem of aspect extraction:

word: we use the word or token extracted as a form of lexicalization of the learning model;

lemma: the use of the lemma (as opposed to the word) brings generalization to learning, which usually improves learning for morphologically rich languages (SEDDAH *et al.*, 2010);

part-of-speech label (PoS) : The PoS label enriches learning by generalizing the word by its function;

dependency relation: the syntactic relation obtained through a dependency parser adds context to the learning model. It provides the syntactic function for the word in the sentence;

head of the syntactic relation: The head of a syntactic relation is the word that governs the syntactic relation towards the word in analysis. With this, we were able to capture the modification and effect relationships between words in the sentence, thus adding context.

To exemplify the use of syntactic knowledge, we display in Figure 5 the output of parser PALAVRAS for the sentence *"Este livro é muito bom!"* ("This book is very good!") [1].

[1] The present syntax was obtained through the graphical visualization of the PALAVRAS parser available at <https://visl.sdu.dk/visl/en/parsing/automatic/dependency.php>

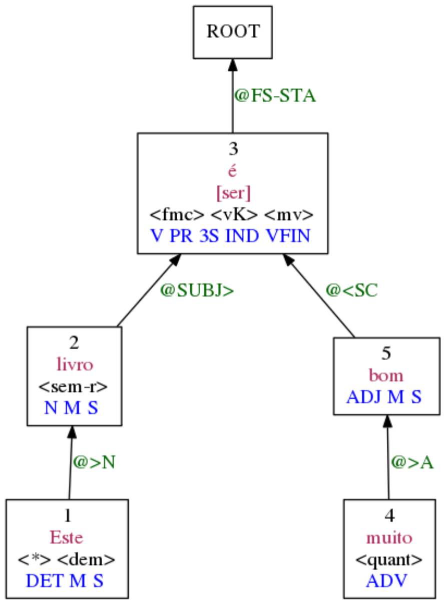

Figure 5 – Example of syntatic parser produced by PALAVRAS

We observe in Figure 5 the relationship between *"book"*, aspect in analysis, and *"good"*, the modifier. The verb *"be"* connects these two words through the syntax relations SUBJ and SC [2].

hierarchical clusters of words: the clusters used here are extracted through the methodology defined by Brown *et al.* (1992); In this technique, words are organized hierarchically into clusters according to their meaning obtained through the context of their use in a corpus;

word representation vectors: word representation vectors capture context and semantic relations between words expressed by their numerical values within a vector space of representation. The generation methodology of the representation vectors adopted here is based on the work of Mikolov and Dean (2013), known as **Word2Vec**;

semantic category from parser PALAVRAS: the syntactic parser PALAVRAS provides together with its analysis the semantic category of some words. This information is tied to its grammar used for sentential analysis.

Table 9 shows a compilation of the most frequent semantic labels presented in the ReLi corpus. The first column displays the parsed semantic tag; The second shows the explanation of what the label represents; The third shows the absolute frequency of this label in the corpus; The fourth shows the relative frequency in relation to the other tags; The fifth and last shows examples of words that were annotated with the respective label and the percentage of these examples within the words that received the same label. For a detailed

[2] A detailed explanation of the syntax functions provided by the PALAVRAS can be verified at <https://visl.sdu.dk/visl/en/info/syntaxmanual_0.html>

knowledge of the labels and their representations it is advisable to consult the detailed parser manual available in Bick (2000) [3]

Table 9 – Semantic tags present in the ReLi corpus annotated by the parser PALAVRAS

Label	Explanation	Freq.	Relative Freq.	Exemples
sem-r	que pode ser lido	5573	9.34%	livro(50.3%), história(13.6%), leitura(5.9%), romance(3.9%)
am	abstrato	3634	6.09%	amor(6.4%), tempo(6.2%), partido(5.9%), poder(3.1%), atenção(2.2%)
ac	abstrato contável	3338	5.60%	amor(7.0%), coisa(10.3%), parte(4.9%), verdade(4.2%)
per	período do tempo	2889	4.84%	história(26.3%), vida(14.8%), ano(7.9%), tempo(7.8%), dia(3.5%)
sem-c	produto da cognição	2075	3.48%	obra(10.2%), fim(5.8%), visão(4.4%), trama(4.0%), opinião(3.7%)
HH	grupo de humanos	1830	3.07%	sociedade(9.9%), parte(9.0%), grupo(7.2%), família(6.3%), governo(4.5%)
H	humano	1790	3.00%	pessoas(22.3%), amor(13.0%), criança(4.9%), tipo(4.7%)
temp	temporal	1710	2.87%	ano(13.4%), tempo(13.1%), final(8.9%), vez(8.4%), fim(7.1%)
percep-f	que pode ser sentido	1667	2.79%	forma(15.2%), verdade(8.5%), realidade(7.7%), nome(4.9%), pena(4.7%)
act	ação	1298	2.18%	ação(3.2%), carinho(2.5%), geração(2.3%), manipulação(1.8%), prisão(1.8%)

Table 10 shows the results for variations of the feature sets in the RELI corpus. In all experiments, we used a size 1 context window (one word after and one before). From experiment 3, we begin to include syntactic features and, from experiment 4, we also include semantic characteristics.

Table 10 – Results of machine learning using CRF in the ReLi corpus

Experiment	Features	Precision	Recall	F-score
1	Word	57,30%	15,50%	24,40%
2	Word+PoS	56,90%	15,40%	24,20%
3	Lemma+PoS+Head	58,50%	20,10%	29,90%
4	Lemma+PoS+Head+Sem	**60,40%**	**24,70%**	**35,10%**
5	Lemma+Pos+Head+Sem+ clusters+Word2Vec	**60,40%**	**24,70%**	**35,10%**

We observed that the fourth and fifth experiments had the best result (35.1% of f-measure). It is observed that the improvement of result in relation to experiment 3 is due to the addition of the semantic information obtained by the parser PALAVRAS. No impact was observed in the inclusion of other semantic features (clusters and Word2Vec). Also, the inclusion of the syntactic function was not promising.

In order to obtain clusters and representation vectors using the Word2Vec technique, a corpus of book reviews was automatically built by crawling the Skoob.com website, which was the source for the elaboration of the ReLi corpus. This collection consists of 343,000 reviews representing the entire collection of reviews of the site as of November 20, 2015.

This result, in comparison with the other scenarios, highlights the importance of adding syntactic and semantic information in the process of extracting aspects. In particular, it is observed that the semantic information obtained through PALAVRAS parser can replace more complex features such as the use of clusters and Word2Vec, which is very common in NLP.

[3] An explanatory summary of the semantic tags is also available at <http://beta.visl.sdu.dk/semantic_prototypes_overview.pdf>

We believe that these tags were important in improving the extraction results of aspects, because they could relate the words to the semantic information they represent. It is interesting to note, for example, that the first label shown in the table (without-r), which is the most frequent, groups the aspects related to the reading of a book, hence its relevance in the extraction of aspects. This semantic information can also be evidenced in the given graphic example of Figure 5.

In the table Table 11, we display the results with the set of characteristics provided by the corpus annotated with the syntactic dependencies trained with the Universal Treebank (UTB), which, as commented earlier in this report, was developed based on a simpler set of so-called "universal" labels. It is emphasized here that the presence of semantic information is not present, as was previously the case with the parser PALAVRAS.

Although UTB adoption is being considered by several state-of-the-art systems such as the future of PoS and syntax in NLP (MANNING *et al.*, 2014), it did not perform better in this work (29.7% of f-score versus 29.9% for the experiment 3 in the corpus analyzed by the PALAVRAS). However, it is seen that linguistic knowledge (syntax) has also improved the performance of the aspect extraction.

Table 11 – Results of machine learning using CRF in the ReLi corpus with syntactic annotations of Universal TreeBank

Experiment	Features	Precision	Recall	F-score
1	Word	**59,00%**	15,10%	24,10%
2	Word+PoS	54,40%	16,10%	24,80%
3	Lemma+PoS+Head	57,10%	**20,00%**	**29,70%**

Table 12 and Table 13 show the results of the machine learning method for the SemEval ABSA 2015 and 2016 corpus, respectively. In the corpus 2015, we also present the results of the two state-of-the-art systems, which were better scored in said competition, so that one has an idea of how far the methods investigated are in relation to the state of the art.

Table 12 – Results of machine learning using CRF in the corpus SemEval ABSA 2015

Experiment	Features	Precision	Recall	F-score
1	Word	**79,00%**	42,90%	55,60%
2	Word+PoS	74,60%	52,80%	61,80%
3	Lemma+PoS+Head	75,80%	51,90%	61,60%
4	NLANGP (TOH; SU, 2015)	70,50%	64,00%	67,10%
5	EliXa (VICENTE; SARALEGI; AGERRI, 2015)	68,90%	**71,20%**	**70,00%**

As discussed earlier, the EliXa (VICENTE; SARALEGI; AGERRI, 2015) system, the best scored, used machine learning based on the Averaged Perceptron algorithm (COLLINS, 2002). The machine learning features used were: n-grams; PoS label; n-grams of prefixes and suffixes; and word cluster (Brown and Clark clusters; and word2vec). The NLANGP (TOH; SU, 2015) system, the second highest score, reached the score of 67.11% of f-score. This system was based on the machine learning algorithm CRF (LAFFERTY; MCCALLUM; PEREIRA,

2001) with the following characteristics: the word itself; head of dependency (obtained from a parser of dependencies); lists of names (extracted based on frequency in corpus); and Brown clusters (BROWN *et al.*, 1992). In terms of f-score, our methods did not produce better results than the two state-of-the-art systems, but they are not far away either. With regard to precision, our method was more assertive in its answers compared with other systems (79,00%).

Table 13 – Results of machine learning using CRF in the corpus SemEval ABSA 2016

Experiment	Features	Precision	Recall	F-score
1	Word	77,20%	46,80%	58,30%
2	Word+PoS	75,30%	55,40%	63,80%
3	Lemma+PoS+Head	**78,00%**	57,50%	66,20%
4	AUEB (XENOS *et al.*, 2016)	71,82%	69,12	70,44%
5	NLANGP (TOH; SU, 2016)	75,49%	**69,44%**	**72,34%**

Again, in the SemEval 2016 corpus, it is seen that the use of more linguistic information (syntax, in this case) has produced better results. The AUEB (XENOS *et al.*, 2016) system, second in the OTE task in the 2016 edition, is based on the *Conditional Random Fields* algorithm with the following set of features: morphological and lexical of (KARAMPATSIS; PAVLOPOULOS; MALAKASIOTIS, 2014), list aspects and *word embeddings*. The best-placed, NLANGP (TOH; SU, 2016) system is an enhancement of the same system submitted in 2015 with the addition of a new learning feature based on the strained probability of a recurrent neural network. Again, we could not overcome the f-measure, but we overcame the state-of-the-art systems in precision.

6.1. Final Remarks

The methods based on machine learning allowed us to achieve results close to the state-of-the-art for SemEval and to offer a good benchmark for ReLi. In specific, the usage of syntatic dependencies and semantic labels showed as expected benefitial to the both corpora. Results for the ReLi show the frequency methods are extremely efficient, possibly due to the annotation criteria and the domain/genre of the reviews.

The next section shows the final considerations.

CHAPTER

7

CONCLUSIONS

7.1. Final considerations

In general, with the work reported, we believe that we could prove our main hypothesis that linguistic knowledge (syntax and semantics) has a significant impact on the task of extracting aspects. It was possible to demonstrate this for Portuguese and also for English. For Portuguese, this was the first systematic investigation of varied methods of aspect extraction for sentiment analysis. For English, although we confirm our hypothesis, we do not overcome the state of the art.

Of course, as can be seen from the various experiments and reported results, the hypothesis that the different paradigm methods have varied predictive power is also confirmed. Finally, there are differences in performance among the languages investigated, but it is not possible to fully confirm this hypothesis, since the linguistic knowledge available for the languages involved were not equivalent and this may have influenced the results. This research question therefore persists for future work in the area.

This research, unprecedented for the Portuguese language, has resulted in **relevant contributions** to the area of research, both theoretical and practical. Some of these are:

- The exploration of methods based on the frequency and proposal of a variation that overcame the classical methods of this approach;
- The exploration of a classic method based on relation and the proposal of automation of its application, by learning automatic patterns of occurrence of aspects;
- The research of linguistic standards in aspect-based sentiment analysis in Portuguese;
- The exploration of methods based on machine learning and its enrichment with linguistic information of a syntactic and semantic nature, producing better results than the original

methods;

- Linguistic characterization of semantic nature of the most frequent aspects in Portuguese language;
- UTB-based syntactic parser training, providing a new tool for the research area;
- Pre-processing and availability of the ReLi corpus with syntactic and semantic information.

7.2. Publications

This section list directly and indirectly publications originated from this research project.

- Pedro Paulo Balage Filho, Thiago Pardo, e Sandra Aluísio. 2013a. **An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis.** In Sandra Maria Aluísio e Valéria Delisandra Feltrim, editores, Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology (STIL). Sociedade Brasileira de Computação, Fortaleza-CE, Brazil, pages 215–219.
- Pedro Paulo Balage Filho e Thiago Alexandre Salgueiro Pardo. 2013b. **NILC_USP: A hybrid system for sentiment analysis in twitter messages.** In Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). Association for Computational Linguistics, Atlanta, Georgia, USA, pages 568–572
- Pedro Paulo Balage Filho e Thiago Alexandre Salgueiro Pardo. 2014a. **NILC_USP: Aspect extraction using semantic labels.** In Preslav Nakov e Torsten Zesch, editores, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Association for Computational Linguistics and Dublin City University, Dublin, Ireland, pages 433–436
- Pedro Paulo Balage Filho e Thiago Alexandre Salgueiro Pardo. 2014b. **BuscaOpinioes: Searching for opinions over the internet.** In Proceedings of the 11th International Conference on Computational Processing of Portuguese Language. Software Demonstration. São Carlos-SP, Brazil, pages 1–3
- Nathan Siegle Hartmann, Lucas Vinicius Avanço, Pedro Paulo Balage Filho, Magali Duran, Maria das Graças Volpe Nunes, Thiago Alexandre Salgueiro Pardo, e Sandra Maria Aluísio. 2014. **A large opinion corpus in portuguese: Tackling out-of-vocabulary words.** In Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC). Reykjavik, Iceland, pages 3865–3871

- Roque Lopez, Thiago Pardo, Lucas Avanço, Pedro Paulo Balage Filho, Alessandro Bokan, Paula Cardoso, Márcio Dias, Fernando Nóbrega, Marco Cabezudo, Jackson Souza, Andressa Zacarias, Eloize Seno, e Ariani Di Felippo. 2015. **A qualitative analysis of a corpus of opinion summaries based on aspects.** In Proceedings of The 9th Linguistic Annotation Workshop. Association for Computational Linguistics, Denver, Colorado, USA, pages 62–71

BIBLIOGRAPHY

AFONSO, B.; BATISTA, D.; BOVNJAK, M.; CARVALHO, P.; CORREIA, P.; COUTO, F.; DIAS, J.; LABOREIRO, G.; MARTINS, J.; MOREIRA, S.; OLIVEIRA, E.; RAMALHO, J. a.; RODRIGUES, E. M.; SARMENTO, L.; SILVA, M. J.; TEIXEIRA, J. **Notas sobre a Realização e Qualidade do Twitómetro**. [S.l.], 2011. Citations on pages 34 and 40.

ALVA-MANCHEGO, F.; ROSA, J. L. Towards Semi-supervised Brazilian Portuguese Semantic Role Labeling: Building a Benchmark. In: CASELI, H.; VILLAVICENCIO, A.; TEIXEIRA, A.; Perdigão, F. (Ed.). **PROPOR 2012**. [S.l.]: Springer, Heidelberg, 2012. (LNAI, v. 7243), p. 210–217. Citation on page 45.

AMANCIO, D. R.; FABBRI, R.; JR, O. N. O.; NUNES, M. G.; COSTA, L. da F. Distinguishing between positive and negative opinions with complex network features. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing**. [S.l.], 2010. p. 83–87. Citation on page 34.

ANDREEVSKAIA, A.; BERGLER, S. Mining WordNet for fuzzy sentiment: Sentiment tag extraction from WordNet glosses. In: **Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)**. Trento, IT: Association for Computational Linguistics, 2006. v. 6, p. 209–216. Available: <http://acl.ldc.upenn.edu/eacl2006/main/papers/13_3_andreevskaiab_262.pdf>. Citation on page 31.

ANNEX, S. **The New Wave of User Generated Content: 7 Concepts for 2017 and Beyond**. [S.l.], 2016. Citation on page 19.

BAILEY, A. Consumer awareness and use of product review websites. **Journal of Interactive Advertising**, 2005. Citation on page 19.

BAUM, L. E.; PETRIE, T. Statistical inference for probabilistic functions of finite state markov chains. **The annals of mathematical statistics**, JSTOR, v. 37, n. 6, p. 1554–1563, 1966. Citation on page 53.

BICK, E. **The parsing system PALAVRAS: automatic grammatical analysis of Portuguese in a constraint grammar framework**. [S.l.]: University of Arhus, 2000. Citations on pages 45 and 55.

BLAIR-GOLDENSOHN, S.; NEYLON, T.; HANNAN, K.; REIS, G. A.; MCDONALD, R.; REYNAR, J. Building a sentiment summarizer for local service reviews. In: **NLP in the Information Explosion Era**. [S.l.: s.n.], 2008. Citation on page 37.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **Journal of machine Learning research**, v. 3, n. Jan, p. 993–1022, 2003. Citation on page 38.

BROWN, P. F.; DESOUZA, P. V.; MERCER, R. L.; PIETRA, V. J. D.; LAI, J. C. Class-based n-gram models of natural language. **Computational linguistics**, MIT Press, v. 18, n. 4, p. 467–479, 1992. Citations on pages 42, 55, and 57.

BURKE, R. R. Technology and the customer interface: What consumers want in the physical and virtual store. **Journal of the Academy of Marketing Science**, v. 30, n. 4, p. 411, 2002. ISSN 1552-7824. Available: <http://dx.doi.org/10.1177/009207002236914>. Citation on page 19.

CARVALHO, P.; SARMENTO, L.; SILVA, M. J.; OLIVEIRA, E. de. Clues for detecting irony in user-generated contents: oh...!! it's so easy;-). In: **Proceeding of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion**. ACM, 2009. p. 53–56. Available: <http://dl.acm.org/citation.cfm?id=1651471>. Citation on page 34.

CARVALHO, P.; SARMENTO, L.; TEIXEIRA, J.; SILVA, M. J. Liars and saviors in a sentiment annotated corpus of comments to political debates. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2**. Portland, Oregon: Association for Computational Linguistics, 2011. p. 564–568. Available: <http://xldb.lasige.di.fc.ul.pt/xldb/publications/Carvalho:LiarsAndSaviors:2011_document.pdf>. Citations on pages 40 and 44.

CHAUDHURI, A. **Emotion and reason in consumer behavior**. [s.n.], 2006. 56–71 p. ISBN 9780750679763. Available: <http://onlinelibrary.wiley.com/doi/10.1002/cbdv.200490137/abstract$\delimiter"026E30F$nhttp://www.lavoisier.fr/livre/notice.asp?id=OLAWOXAL6LXOWV>. Citation on page 26.

CHAVES, M. S.; FREITAS, L. A. de; SOUZA, M.; VIEIRA, R. Pirpo: An algorithm to deal with polarity in portuguese online reviews from the accommodation sector. In: SPRINGER. **International Conference on Application of Natural Language to Information Systems**. [S.l.], 2012. p. 296–301. Citation on page 40.

CHURCH, K. W.; HANKS, P. Word association norms, mutual information, and lexicography. **Computational Linguistics**, MIT Press, v. 16, n. 1, p. 22–29, Mar. 1990. ISSN 0891-2017. Available: <http://dl.acm.org/citation.cfm?id=89086.89095>. Citation on page 33.

COLLINS, M. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10**. [S.l.], 2002. p. 1–8. Citations on pages 42 and 57.

DAS, S.; CHEN, M. Yahoo! for Amazon: Extracting market sentiment from stock message boards. In: **Proceedings of the Asia Pacific Finance Association Annual Conference (APFA)**. Barcelona, Spain: [s.n.], 2001. v. 35. Available: <http://scholar.google.com/scholar?hl=en&q=Yahoo!+for+amazon:+Extracting+market+sentiment+from+stock+message+boards&btnG=Search&as_sdt=1,5&as_ylo=&as_vis=0#0>. Citation on page 25.

DING, X.; LIU, B.; YU, P. A holistic lexicon-based approach to opinion mining. In: **Proceedings of the international conference on Web search and web data mining**. ACM, 2008. p. 231–240. Available: <http://dl.acm.org/citation.cfm?id=1341561>. Citations on pages 32 and 39.

EKMAN, P. Are there basic emotions? **Psychological review**, v. 99, p. 550–553, 1992. ISSN 0033-295X. Citations on pages 11, 26, and 27.

ESULI, A.; SEBASTIANI, F. Determining Term Subjectivity and Term Orientation for Opinion Mining. In: **Proceedings of the 11th Conference of the European Chapter of the Association**

for Computational Linguistics (EACL'06). Trento, IT: [s.n.], 2006. p. 193–200. Available: <http://tcc.itc.it/projects/ontotext/Publications/EACL-2006-Esuli-Sebastiani.pdf>. Citation on page 31.

____. Sentiwordnet: A publicly available lexical resource for opinion mining. In: **Proceedings the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2006)**. Citeseer, 2006. v. 6, p. 417–422. Available: <http://gandalf.aksis.uib.no/lrec2006/pdf/384_pdf.pdfhttp://acl.ldc.upenn.edu/eacl2006/main/papers/13_1_esulisebastiani_192.pdf>. Citations on pages 31 and 40.

FELLBAUM, C. **WordNet: An electronic lexical database**. [S.l.]: The MIT press, 1998. Citations on pages 31 and 39.

FERNANDES, F. M. d. M. **Um Framework para Análise de Sentimento em Comentários sobre Produtos em Redes Sociais**. Recife-PE, Brasil: [s.n.], 2010. Citation on page 40.

FILHO, P. P. B.; BRUN, C.; RONDEAU, G. A graphical user interface for feature-based opinion mining. In: **Proceedings of the NAACL-HLT 2012: Demonstration Session**. Montréal, Canada: Association for Computational Linguistics, 2012. p. 5–8. Citation on page 39.

FILHO, P. P. B.; PARDO, T.; ALUÍSIO, S. An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In: ALUÍSIO, S. M.; FELTRIM, V. D. (Ed.). **Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology (STIL)**. Fortaleza-CE, Brazil: Sociedade Brasileira de Computação, 2013. p. 215–219. Citation on page 34.

FINKEL, J. R.; GRENAGER, T.; MANNING, C. Incorporating non-local information into information extraction systems by gibbs sampling. In: **Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)**. Ann Arbor, Michigan: Association for Computational Linguistics, 2005. p. 363–370. Available: <http://www.aclweb.org/anthology/P05-1045>. Citation on page 53.

FONSECA, E. R.; ROSA, J. L. An Architecture for Semantic Role Labeling on Portuguese. In: CASELI, H.; VILLAVICENCIO, A.; TEIXEIRA, A.; Perdigão, F. (Ed.). **PROPOR 2012**. [S.l.]: Springer, Heidelberg, 2012. (LNAI, v. 7243), p. 204–209. Citation on page 45.

FREITAS, C.; MOTTA, E.; MILIDIÚ, R.; CÉSAR, J. Vampiro que brilha... rá! desafios na anotaçao de opiniao em um corpus de resenhas de livros. **ENCONTRO DE LINGUÍSTICA DE CORPUS**, v. 11, p. 3, 2012. Citations on pages 15, 22, 43, and 44.

GUO, H.; ZHU, H.; GUO, Z.; ZHANG, X.; SU, Z. Product feature categorization with multilevel latent semantic association. In: ACM. **Proceeding of the 18th ACM conference on Information and knowledge management**. ACM Press, 2009. (CIKM '09), p. 1087–1096. ISBN 9781605585123. Available: <http://portal.acm.org/citation.cfm?doid=1645953.1646091>. Citation on page 38.

GWAVA. **How Much Data is Created on the Internet Each Day?** 2016. Online. Available: <https://www.gwava.com/blog/internet-data-created-daily>. Citation on page 19.

HATZIVASSILOGLOU, V.; MCKEOWN, K. R. Predicting the semantic orientation of adjectives. In: **Proceedings of the 35th Annual Meeting of the Association for Computational**

Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 1997. (ACL '98), p. 174–181. Citation on page 32.

HOFMANN, T. Probabilistic latent semantic indexing. In: ACM. **Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval**. [S.l.], 1999. p. 50–57. Citation on page 38.

HSIEH, W.-t.; WU, C.-M.; KU, T.; CHOU, S.-c. T. Social Event Radar : A Bilingual Context Mining and Sentiment Analysis Summarization System. In: **Proceedings of the ACL 2012 System Demonstrations**. Jeju Island, Korea: Association for Computational Linguistic, 2012. p. 163–168. Citation on page 39.

HU, M.; LIU, B. Mining and summarizing customer reviews. In: ACM. **Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining**. [S.l.], 2004. p. 168–177. Citations on pages 31 and 37.

JAKOB, N.; GUREVYCH, I. Extracting opinion targets in a single-and cross-domain setting with conditional random fields. In: **Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing**. [s.n.], 2010. p. 1035–1045. Available: <http://portal.acm.org/citation.cfm?id=1870759>. Citation on page 38.

JIANG, L.; YU, M.; ZHOU, M.; LIU, X.; ZHAO, T. Target-dependent Twitter Sentiment Classification. **Computational Linguistics**, Association for Computational Linguistics, v. 1, n. June 19-24, 2011, p. 151–160, 2011. Available: <http://www.aclweb.org/anthology/P11-1016>. Citation on page 39.

JIN, W.; HO, H. H.; SRIHARI, R. K. OpinionMiner: A Novel Machine Learning System for Web Opinion Mining and Extraction. In: ACM. **Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining**. New York, New York, USA: ACM New York, NY, USA, 2009. p. 1195–1203. ISBN 9781605584959. Available: <http://portal.acm.org/citation.cfm?id=1557019.1557148>. Citation on page 38.

KAPLAN, A. M.; HAENLEIN, M. Users of the world, unite! the challenges and opportunities of social media. **Business horizons**, Elsevier, v. 53, n. 1, p. 59–68, 2010. Citation on page 20.

KARAMPATSIS, R. M.; PAVLOPOULOS, J.; MALAKASIOTIS, P. Aueb: Two stage sentiment analysis of social network messages. In: **Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)**. Dublin, Ireland: Association for Computational Linguistics and Dublin City University, 2014. p. 114–118. Available: <http://www.aclweb.org/anthology/S14-2015>. Citations on pages 42 and 58.

KESSLER, J.; NICOLOV, N. Targeting Sentiment Expressions through Supervised Ranking of Linguistic Configurations. In: **Proceedings of the Third International AAAI Conference on Weblogs and Social Media (ICWSM-09)**. San Jose, California, US: [s.n.], 2009. p. 90 – 97. Citation on page 37.

KIM, S.; HOVY, E. Determining the sentiment of opinions. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 20th international conference on Computational Linguistics**. [S.l.], 2004. p. 1367–es. Citation on page 31.

KOBAYASHI, N.; INUI, K.; MATSUMOTO, Y. Opinion Mining from Web Documents: Extraction and Structurization. **Transactions of the Japanese Society for Artificial Intelligence**, v. 22, p. 227–238, 2007. ISSN 1346-0714. Available: <http://www.aclweb.org/anthology/D/D07/D07-1114http://joi.jlc.jst.go.jp/JST.JSTAGE/tjsai/22.227?from=CrossRef>. Citation on page 37.

LAFFERTY, J.; MCCALLUM, A.; PEREIRA, F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: **Proceedings of the eighteenth international conference on machine learning, ICML**. [S.l.: s.n.], 2001. v. 1, p. 282–289. Citations on pages 42, 53, and 57.

LERMAN, K.; BLAIR-GOLDENSOHN, S.; MCDONALD, R. Sentiment summarization: evaluating and learning user preferences. In: **Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. (EACL '09), p. 514–522. Available: <http://dl.acm.org/citation.cfm?id=1609067.1609124>. Citation on page 39.

LI, F.; HUANG, M.; ZHU, X. Sentiment Analysis with Global Topics and Local Dependency. In: FOX, M.; POOLE, D. (Ed.). **AAAI**. AAAI Press, 2010. Available: <http://dblp.uni-trier.de/db/conf/aaai/aaai2010.html#LiHZ10>. Citation on page 38.

LIN, C.; HE, Y. Joint Sentiment/Topic Model for Sentiment Analysis. In: **Proceedings of the 18th ACM Conference on Information and Knowledge Management**. New York, NY, USA: ACM, 2009. (CIKM '09), p. 375–384. ISBN 978-1-60558-512-3. Available: <http://doi.acm.org/10.1145/1645953.1646003>. Citation on page 38.

LIU, B. Sentiment analysis and subjectivity. **Handbook of natural language processing**, v. 2, p. 627–666, 2010. Citations on pages 20, 24, 27, 28, 30, 35, and 36.

____. **Sentiment analysis and opinion mining**. Morgan & Claypool Publishers, 2012. ISSN 1947-4040. Available: <https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf>. Citations on pages 26, 37, and 38.

LIU, B.; HU, M.; CHENG, J. Opinion Observer: Analyzing and comparing opinions on the web. In: **Proceedings of WWW '05, the 14th international conference on World Wide Web**. [S.l.: s.n.], 2005. p. 342–351. Citation on page 39.

LONG, C.; JIE, Z.; XIAOYAN, Z. A Review Selection Approach for Accurate Feature Rating Estimation. **Proceedings of the 23rd International Conference on Computational Linguistics Posters**, Coling 2010 Organizing Committee, n. August, p. 766–774, 2010. Available: <http://portal.acm.org/citation.cfm?id=1944566.1944654>. Citation on page 37.

MANNING, C. D.; SURDEANU, M.; BAUER, J.; FINKEL, J. R.; BETHARD, S.; MCCLOSKY, D. The stanford corenlp natural language processing toolkit. In: **ACL (System Demonstrations)**. [S.l.: s.n.], 2014. p. 55–60. Citation on page 57.

MEI, Q.; LING, X.; WONDRA, M.; SU, H.; ZHAI, C. Topic sentiment mixture: modeling facets and opinions in weblogs. In: ACM. **Proceedings of the 16th international conference on World Wide Web**. [S.l.], 2007. p. 171–180. Citation on page 38.

MIKOLOV, T.; DEAN, J. Distributed representations of words and phrases and their compositionality. **Advances in neural information processing systems**, 2013. Citations on pages 42 and 55.

MOGHADDAM, S. A. **Aspect-based opinion mining in online reviews**. Phd Thesis (PhD Thesis) — School of Computing Science, Faculty of Applied Sciences, Simon Fraser University, 2013. Available: <http://summit.sfu.ca/system/files/iritems1/12790/etd7688_SAbbasiMoghaddam.pdf>. Citation on page 42.

MORINAGA, S.; YAMANISHI, K.; TATEISHI, K.; FUKUSHIMA, T. Mining product reputations on the Web. In: **Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02**. New York, New York, USA: ACM Press, 2002. p. 341. ISBN 158113567X. Available: <http://dl.acm.org/citation.cfm?id=775047.775098>. Citation on page 25.

MUKHERJEE, A.; LIU, B. Aspect Extraction Through Semi-Supervised Modeling. In: **Proceedings of 50th anunal meeting of association for computational Linguistics**. Jeju Island, Korea: Association for Computational Linguistics, 2012. p. 339–348. Available: <http://newdesign.aclweb.org/anthology/P/P12/P12-1036.pdf>. Citation on page 39.

NASUKAWA, T.; YI, J. Sentiment analysis: Capturing favorability using natural language processing. In: ACM. **Proceedings of the 2nd international conference on Knowledge capture**. [S.l.], 2003. p. 70–77. Citations on pages 15, 50, and 51.

OKAZAKI, N. **CRFsuite: a fast implementation of conditional random fields (CRFs)**. 2007. Available: <http://www.chokkan.org/software/crfsuite>. Citation on page 54.

PANG, B.; LEE, L. Opinion mining and sentiment analysis. **Foundations and trends in information retrieval**, Now Publishers Inc., v. 2, n. 1-2, p. 1–135, 2008. Citations on pages 20 and 25.

PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up?: sentiment classification using machine learning techniques. In: **Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10**. Morristown, NJ, USA: Association for Computational Linguistics, 2002. (EMNLP '02), p. 79–86. Available: <http://dx.doi.org/10.3115/1118693.1118704>. Citations on pages 25, 31, and 32.

PARROTT, W. G. **Emotions in Social Psychology: Essentials Readings**. New York, NY, US: Psychology Press, 2001. 378 pp p. ISBN 0-86377-682-5. Citation on page 26.

PASQUALOTTI, P. R. **Reconhecimento de expressões de emoções na interação mediada por computador**. 95 p. Master's Thesis (Master's Thesis) — UNIVERSIDADE DO VALE DO RIO DOS SINOS, 2008. Available: <http://www.dominiopublico.gov.br/download/texto/cp057721.pdfhttp://bdtd.unisinos.br/tde_arquivos/1/TDE-2008-09-02T064908Z-574/Publico/PauloRobertoPasqualotti.pdf>. Citation on page 34.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine Learning in Python . **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. Citation on page 54.

PONTIKI, M.; GALANIS, D.; PAPAGEORGIOU, H.; MANANDHAR, S.; ANDROUTSOPOULOS, I. 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. Available: <http://www.aclweb.org/anthology/S15-2082>. Citations on pages 21, 22, 41, and 53.

PONTIKI, M.; GALANIS, D.; PAPAGEORGIOU, H.; ANDROUTSOPOULOS, I.; MANANDHAR, S.; AL-SMADI, M.; AL-AYYOUB, M.; ZHAO, Y.; QIN, B.; CLERCQ, O. D.; HOSTE, V.; APIDIANAKI, M.; TANNIER, X.; LOUKACHEVITCH, N.; KOTELNIKOV, E.; BEL, N.; JIMéNEZ-ZAFRA, S. M.; ERYIĞIT, G. 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. Available: <http://www.aclweb.org/anthology/S16-1002>. Citations on pages 21, 22, 41, and 53.

PONTIKI, M.; GALANIS, D.; PAVLOPOULOS, J.; PAPAGEORGIOU, H.; ANDROUTSOPOULOS, I.; MANANDHAR, S. 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 and Dublin City University, 2014. p. 27–35. Available: <http://www.aclweb.org/anthology/S14-2004>. Citations on pages 21, 41, and 47.

POPESCU, A.-M.; ETZIONI, O. Extracting product features and opinions from reviews. **Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing**, p. 339–346, 2005. Available: <http://dl.acm.org/citation.cfm?id=1220618>. Citations on pages 37 and 38.

QIU, L.; ZHANG, W.; HU, C.; ZHAO, K. SELC : A Self-Supervised Model for Sentiment Classification. **Information Storage and Retrieval**, ACM, v. 26, n. 6, p. 929–936, 2009. ISSN 00104825. Available: <http://portal.acm.org/citation.cfm?id=1645953.1646072>. Citation on page 32.

RIBEIRO, S.; JUNIOR, Z.; MEIRA, W.; PAPPA, G. L. Positive or negative? using blogs to assess vehicles features. **Encontro Nacional de Inteligência Artificial**, p. 1–12, 2012. Citations on pages 40 and 44.

RILOFF, E.; PATWARDHAN, S.; WIEBE, J. Feature subsumption for opinion analysis. In: **Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing - EMNLP '06**. Morristown, NJ, USA: Association for Computational Linguistics, 2006. p. 440. ISBN 1932432736. Available: <http://portal.acm.org/citation.cfm?doid=1610075.1610137>. Citation on page 26.

SARMENTO, L.; CARVALHO, P.; SILVA, M. J.; OLIVEIRA, E. D. Automatic creation of a reference corpus for political opinion mining in user-generated content. In: ACM. **Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion**. [S.l.], 2009. p. 29–36. Citations on pages 21 and 34.

SCOPIM, D. *et al.* **Estudo de padrões lexicais em textos opinativos**. Master's Thesis (Mestrado em Linguística)) — Universidade Federal de São Carlos, 2012. Citation on page 34.

SEDDAH, D.; CHRUPAŁA, G.; CETINOGLU, O.; GENABITH, J. van; CANDITO, M. Lemmatization and lexicalized statistical parsing of morphologically-rich languages: the case of french. In: **Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages**. Los Angeles, CA, USA: Association for Computational Linguistics, 2010. p. 85–93. Available: <http://www.aclweb.org/anthology/W10-1410>. Citation on page 54.

SHA, F.; PEREIRA, F. Shallow parsing with conditional random fields. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 2003 Conference of the North**

American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. [S.l.], 2003. p. 134–141. Citation on page 53.

SILFVERBERG, M.; RUOKOLAINEN, T.; LINDéN, K.; KURIMO, M. Part-of-speech tagging using conditional random fields: Exploiting sub-label dependencies for improved accuracy. In: **Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**. Baltimore, Maryland: Association for Computational Linguistics, 2014. p. 259–264. Available: <http://www.aclweb.org/anthology/P14-2043>. Citation on page 53.

SILVA, M. J.; CARVALHO, P.; COSTA, C.; SARMENTO, L. **Automatic expansion of a social judgment lexicon for sentiment analysis**. [S.l.], 2010. Citation on page 34.

SILVA, M. J.; CARVALHO, P.; SARMENTO, L.; aES, P. M.; OLIVEIRA, E. The Design of OPTIMISM, an Opinion Mining System for Portuguese Politics. **New Trends in Artificial Intelligence: Proceedings of EPIA 2009 - Fourteenth Portuguese Conference on Artificial Intelligence**, Universidade de Aveiro, p. 565–576, Oct. 2009. Available: <http://paginas.fe.up.pt/~niadr/PUBLICATIONS/2009/epia2009-OPTIMISM-submitted.pdfhttp://epia2009.web.ua.pt/onlineEdition/565.pdf>. Citation on page 40.

SILVA, N. G. R. da. **BestChoice: Classificação de Sentimento em Ferramentas de Expressão de Opinião**. Phd Thesis (Tese de graduação,) — Universidade Federal de Pernambuco, Recife, 2010. 7, 17, 2010. Citation on page 40.

SIQUEIRA, H.; BARROS, F. A feature extraction process for sentiment analysis of opinions on services. In: **Proceedings of International Workshop on Web and Text Intelligence**. São Bernardo do Campo-SP, Brazil: [s.n.], 2010. Citations on pages 40 and 44.

SOMASUNDARAN, S.; NAMATA, G.; GETOOR, L.; WIEBE, J. Opinion graphs for polarity and discourse classification. In: **Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing - TextGraphs-4**. Morristown, NJ, USA: Association for Computational Linguistics, 2009. p. 66. ISBN 9781932432541. Available: <http://www.aclweb.org/anthology/W/W09/W09-3210http://portal.acm.org/citation.cfm?doid=1708124.1708138>. Citation on page 37.

SOUZA, M.; VIEIRA, R. Sentiment analysis on twitter data for portuguese language. In: SPRINGER. **International Conference on Computational Processing of the Portuguese Language**. [S.l.], 2012. p. 241–247. Citation on page 34.

SOUZA, M.; VIEIRA, R.; BUSETTI, D.; CHISHMAN, R.; ALVES, I. M. *et al.* Construction of a portuguese opinion lexicon from multiple resources. In: **8th Brazilian Symposium in Information and Human Language Technology**. [S.l.: s.n.], 2011. p. 59–66. Citations on pages 34 and 40.

TABOADA, M.; BROOKE, J.; TOFILOSKI, M.; VOLL, K.; STEDE, M. Lexicon-Based Methods for Sentiment Analysis. **Computational Linguistics**, MIT Press, v. 37, n. 2, p. 267–307, Jun. 2011. ISSN 0891-2017. Available: <http://www.sfu.ca/~mtaboada/docs/Taboada_etal_SO-CAL.pdfhttp://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00049>. Citation on page 33.

TITOV, I.; MCDONALD, R. T. A joint model of text and aspect ratings for sentiment summarization. In: CITESEER. **ACL**. [S.l.], 2008. v. 8, p. 308–316. Citation on page 38.

TOH, Z.; SU, J. Nlangp: Supervised machine learning system for aspect category classification and opinion target extraction. In: **Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)**. Denver, Colorado: Association for Computational Linguistics, 2015. p. 496–501. Available: <http://www.aclweb.org/anthology/S15-2083>. Citations on pages 42 and 57.

____. Nlangp at semeval-2016 task 5: Improving aspect based sentiment analysis using neural network features. In: **Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)**. San Diego, California: Association for Computational Linguistics, 2016. p. 282–288. Available: <http://www.aclweb.org/anthology/S16-1045>. Citations on pages 42 and 58.

TONG, R. An Operational System for Detecting and Tracking Opinions in On-line Discussion. In: **SIGIR 2001 Workshop on Operational Text Classification**. [S.l.: s.n.], 2001. Citation on page 25.

TURNEY, P. D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: **Proceedings of the 40th Annual Meeting on Association for Computational Linguistics**. Morristown, NJ, USA: Association for Computational Linguistics, 2002. (ACL ’02), p. 417–424. Citations on pages 25, 31, 33, and 40.

VICENTE, I. S.; SARALEGI, X.; AGERRI, R. Elixa: A modular and flexible absa platform. In: **Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)**. Denver, Colorado: Association for Computational Linguistics, 2015. p. 748–752. Available: <http://www.aclweb.org/anthology/S15-2127>. Citations on pages 42, 49, and 57.

WIEBE, J. Learning Subjective Adjectives from Corpora. In: **Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence**. AAAI Press, 2000. p. 735–740. ISBN 0-262-51112-6. Available: <http://dl.acm.org/citation.cfm?id=647288.721121>. Citation on page 25.

WIEBE, J.; WILSON, T.; CARDIE, C. Annotating expressions of opinions and emotions in language. **Language resources and evaluation**, Springer, v. 39, n. 2-3, p. 165–210, 2005. Available: <http://www.springerlink.com/index/E854800108051U61.pdf>. Citations on pages 26 and 32.

WIEBE, J. M.; BRUCE, R. F.; O’HARA, T. P. Development and use of a gold-standard data set for subjectivity classifications. In: **Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics**. Stroudsburg, PA, USA: Association for Computational Linguistics, 1999. (ACL ’99), p. 246–253. ISBN 1-55860-609-3. Available: <http://dx.doi.org/10.3115/1034678.1034721>. Citation on page 31.

WILSON, T.; WIEBE, J.; HOFFMANN, P. Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level Sentiment Analysis. **Computational Linguistics**, MIT Press, v. 35, n. 3, p. 399–433, Sep. 2009. Citation on page 32.

WILSON, T.; WIEBE, J.; HWA, R. Just how mad are you? finding strong and weak opinion clauses. AAAI Press, p. 761–767, Jul. 2004. Available: <http://dl.acm.org/citation.cfm?id=1597148.1597270>. Citation on page 31.

XENOS, D.; THEODORAKAKOS, P.; PAVLOPOULOS, J.; MALAKASIOTIS, P.; ANDROUTSOPOULOS, I. Aueb-absa at semeval-2016 task 5: Ensembles of classifiers and embeddings for 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. 312–317. Available: <http://www.aclweb.org/anthology/S16-1050>. Citations on pages 42 and 58.

ZHAI, Z.; LIU, B.; XU, H.; JIA, P. Clustering Product Features for Opinion Mining. In: **Proceedings of the Fourth ACM International Conference on Web Search and Data Mining**. New York, NY, USA: ACM, 2011. (WSDM '11), p. 347–354. ISBN 978-1-4503-0493-1. Available: <http://doi.acm.org/10.1145/1935826.1935884>. Citation on page 38.

ZHUANG, L.; JING, F.; ZHU, X.-Y. Movie review mining and summarization. In: ACM. **Proceedings of the 15th ACM international conference on Information and knowledge management**. [S.l.], 2006. p. 43–50. Citation on page 37.

ICMC USP
SÃO CARLOS