Pontifícia Universidade Católica do Rio Grande do Sul
Faculdade de Informática
Programa de Pós-Graduação em Ciência da Computação

Feature-Level Sentiment Analysis Applied to Brazilian Portuguese Reviews

Larissa Astrogildo de Freitas

Thesis presented as partial requirement for obtaining a Ph.D. degree in Computer Science at Pontifícia Universidade Católica do Rio Grande do Sul.

Advisor: Renata Vieira

Porto Alegre
2015

Dados Internacionais de Catalogação na Publicação (CIP)

F866f Freitas, Larissa Astrogildo de

Feature-level sentiment analysis applied to brasilian portuguese reviews / Larissa Astrogildo de Freitas. – Porto Alegre, 2015.
94 p.

Tese (Doutorado) – Fac. de Informática, PUCRS.
Orientador: Renata Vieira.

1. Informática. 2. Ontologia. 3. Processamento da Linguagem Natural. I. Vieira, Renata. II. Título.

CDD 006.35

Ficha Catalográfica elaborada pelo
Setor de Tratamento da Informação da BC-PUCRS

Pontifícia Universidade Católica do Rio Grande do Sul
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

TERMO DE APRESENTAÇÃO DE DISSERTAÇÃO DE MESTRADO

Tese intitulada *"Feature-Level Sentiment Analysis Applied to Brazilian Portuguese Reviews"* apresentada por Larissa Astrogildo de Freitas como parte dos requisitos para obtenção do grau de Doutora em Ciência da Computação, aprovada em 23/03/2015 pela Comissão Examinadora:

Profa. Dra. Renata Vieira– PPGCC/PUCRS
Orientador

Profa. Dra. Vera Lúcia Strube de Lima– PPGCC/PUCRS

Prof. Dr. Adriano César Machado Pereira– UFMG

Prof. Dr. Sandro José Rigo– UNISINOS

Homologada em 29/04/2015, conforme Ata No. 007 pela Comissão Coordenadora.

Prof. Dr. Luiz Gustavo Leão Fernandes
Coordenador

PUCRS
Campus Central
Av. Ipiranga, 6681 – P32– sala 507 – CEP: 90619-900
Fone: (51) 3320-3611 – Fax (51) 3320–3621
E-mail: ppgcc@pucrs.br
www.pucrs.br/facin/pos

"No great discovery was ever made without a bold guess"

Isaac Newton

ACKNOWLEDGEMENTS

To my family Edú, Elaine, Eduardo, Bianca and Ulisses, for their encouragement.

To my advisor Renata, for her patience.

I thank my fellows of Natural Language Processing Research Group at PUCRS, for the stimulating discussions and for all the fun we have had in the last four years.

To FAPERGS, CAPES and PUCRS, for their financial support.

Análise de Sentimento no Nível de Aspecto Aplicada à Opiniões em Português Brasileiro

RESUMO

Análise de sentimento é o campo de estudo que analisa a opinião de pessoas em textos. Na última década, humanos têm compartilhado suas opiniões em mídias sociais na Web (por exemplo, fóruns de discussão e *posts* em sites de redes sociais). Opiniões são importantes porque sempre que necessitamos tomar uma decisão, queremos saber o ponto de vista de outras pessoas. O interesse da indústria e da academia neste campo de estudo se deve a aplicações potenciais, tais como: compra/venda, relações públicas e campanhas políticas. Pesquisas neste campo muitas vezes consideram dados em inglês, enquanto dados em outros idiomas são pouco explorados. É possível realizar a análise dos dados em diferentes níveis, neste trabalho optamos pela análise no nível de aspecto, na qual a granularidade é mais fina. Como ontologias podem ser utilizadas para representar aspectos, que são "parte-de" um objeto ou propriedade de "parte-de" um objeto, propomos um método para análise de sentimento aplicado a comentários em português brasileiro, sob o nível de aspecto usando ontologias. A fim de obter uma análise completa, reconhecemos aspectos explícitos e implícitos usando ontologias. Relativamente poucos trabalhos têm sido feitos sobre identificação de aspectos implícitos. Finalmente determinamos se o sentimento em relação aos aspectos é positivo ou negativo usando léxicos de sentimento e regras linguísticas. Nosso método é composto de quatro etapas: pré-processamento, identificação de aspecto, identificação de polaridade e sumarização. Para avaliar este trabalho, aplicamos o método proposto nos comentários do setor hoteleiro. De acordo com nosso experimento, o melhor resultado obtido foi quando utilizamos o TreeTagger, o *synset* com polaridade do Onto.PT e a regra linguística (posição do adjetivo) na identificação da polaridade negativa e (*baseline*) na identificação da polaridade positiva.

Palavras-chave: Análise de Sentimento; Nível de Característica/Aspecto; Ontologia; Comentários em Português Brasileiro; Mídias Sociais; Web.

Feature-Level Sentiment Analysis Applied to Brazilian Portuguese Reviews

ABSTRACT

Sentiment Analysis is the field of study that analyzes people's opinions in texts. In the last decade, humans have come to share their opinions in social media on the Web (e.g., forum discussions and posts in social network sites). Opinions are important because whenever we need to take a decision, we want to know others' points of view. The interest of industry and academia in this field of study is partly due to its potential applications, such as: marketing, public relations and political campaign. Research in this field often considers English data, while data from other languages are less explored. It is possible realize data analysis in different levels, in this work we choose a finer-grain analysis, at aspect-level. Ontologies can represent aspects, that are "part-of" an object or property of "part-of" an object, we proposed a method for feature-level sentiment analysis using ontologies applied to Brazilian Portuguese reviews. In order to obtain a complete analysis, we recognized features explicit and implicit using ontologies. Relatively less work has been done about implicit feature identification. Finally, determine whether the sentiment in relation to the aspects is positive or negative using sentiment lexicons and linguistic rules. Our method is comprised of four steps: preprocessing, feature identification, polarity identification and summarizing. For evaluate this work, we apply our proposal method to a dataset of accommodation sector. According to our experiments, in general the best results were obtained when using TreeTagger, synsets with polarities from Onto.PT and linguistic rule (adjective position) for negative polarity identification and (baseline) for positive polarity identification.

Keywords: Sentiment Analysis; Feature/Aspect Level; Ontology; Brazilian Portuguese Reviews; Social Media; Web.

List of Figures

List of Tables

List of Abbreviations

CERN	*Conseil Européen pour la Recherche Nucléaire*
WWW	World Wide Web
BBC	British Broadcasting Corporation
LIWC	Linguistic Inquiry and Word Count
POS	Part-Of-Speech
NLP	Natural Language Processing
SVM	Support Vector Machine
NB	Naïve Bayes
ME	Maximum Entropy
OWL	Web Ontology Language
ANEW	Affective Norms for English Words
TEP	*Thesaurus Eletrônico Básico para o Português do Brasil*
OSPM	Ontology Supported Polarity Mining
GATE	General Architecture for Text Engineering
ABSA	Aspect Based Sentiment Analysis
CRF	Conditional Random Fields

List of Fragments

Table of Contents

1. Introduction

This chapter presents a brief introduction about our work, describing its motivation, research problem, objectives, expected contributions, and the thesis organization.

1.1 The Motivation

In the 1990s the Internet ceased to be a resource limited to military organizations and research institutes and became a tool accessible to the general public. However, it only became important due to a British physicist and computer scientist called Tim Berners-Lee, that worked at CERN (from French, *Conseil Européen pour la Recherche Nucléaire*). While working at CERN, Tim Berners-Lee proposed a way to share content on the Internet. Nowadays, we know it as World Wide Web (WWW), or simply Web.

The way of sharing content on the Internet proposed by Tim Berners-Lee became popular, and today the Web is considered the world's biggest data repository. Initially, the content on the Web was available in static systems, usually powered by organizations (companies, universities, governments, *etc.*) or by computer enthusiasts. With the evolution of computer science, Web systems brought interactivity to the general public, turning them from content consumers into possible content generators. Nowadays, it is possible to share information with others individuals around the world, in real time, through several Web tools, such as: forums, blogs, and social networks.

Before the creation of the Web, when people needed to take decisions, they were limited to the opinion of their family and friends. Today, a person can access a much bigger universe of opinions about entities and aspects through the Web.

However, the large volume of opinions generated by these people brings a new problem: how to summarize these data. According to Liu [LIU12] the availability of these large volumes of opinion data encouraged a new research area in computer science called sentiment analysis, which became increasingly prominent in the 2000s.

Still, due to numerous practical applications, both, industry and academy have interests in the area of sentiment analysis. Opinions are available on the Web in an unstructured or in a semi-structured way, so it is very difficult to automatically process it [LIU12]. Moreover, it is time consuming and expensive. Therefore creating tools to automate tasks related to sentiment analysis becomes increasingly important.

The main motivation for this work is that Brazil is the fourth country with more Internet users in the world, according to British Broadcasting Corporation (BBC) [BBC14]. Moreover, according to Semiocast [SEM10], Portuguese is the third most used language on Twitter, after English and Japanese.

Customers identify online reviews as having a significant influence on their purchase in many economic sectors (e.g., hotel 87%, travel 84%, restaurant 79%, legal 79%, automotive 78%, medical

76%, and home 73%)[1].

1.2 The Problem

The area of sentiment analysis has been investigated mainly at three levels of granularity (document, sentence or feature) [LIU12]. According to Liu [LIU10], both document-level and sentence-level analyses do not cover what exactly people liked or not. Studying the opinion text, mainly in feature level, is more challenging because it involves fine-grained sentiment analysis. Research on feature-level sentiment analysis often deals with data in English, while data from other languages are less explored. In Brazil, research on this area is still in its beginning, but an effort is being made to create resources and techniques to be used in this task. For instance, the external resources needed for the process, such as sentiment lexicons, only started to be developed for Brazilian Portuguese in 2011. As far as we know, there are only four lexicons of this kind for Portuguese, OpLexicon [SOU11], SentiLex [SIL12], Brazilian Portuguese Linguistic Inquiry and Word Count (LIWC) dictionary [ALU13] and synsets with polarities of Onto.PT [OLI14]. Sentiment lexicons are necessary but they are not sufficient for sentiment analysis [LIU12]. A positive or negative sentiment word may have opposite orientations depending on the application domain (e.g.: 'cerveja gelada' ['cold beer'] and 'pizza gelada' ['cold pizza']). Still, the same feature may also be expressed with different words (e.g.: 'bebida' ['drink'] and 'refresco' ['beverage']) and, it is possible to recognize relationship between features (e.g.: 'café da manhã' ['breakfast'] is a subcategory of 'refeição' ['meal']). Moreover, another aspect to be considered is that features can be explicit (e.g.: "O **café da manhã** deste hotel é bom." ["The breakfast in this hotel is great."], '**café da manhã**' ['breakfast'] is an explicit feature) or implicit (e.g.: "O **café da manhã** deste hotel é bom." ["The breakfast in this hotel is great."], '**refeição**' ['meal'] is an implicit feature). Most of the existing research focuses on finding explicit features, but only a few studies have been done about extracting implicit features [ASG14]. In order to solve problems like this one about implicit features, external resources, such as domain ontologies, can be used.

Research in Portuguese sentiment analysis is still in its first steps. There is lack of datasets and of more deep studies that consider further levels of analysis, such as feature level and implicit features as we propose in this thesis.

1.3 Objective

1.3.1 General Objective

The general objective of this thesis is to propose (defining, implementing and evaluating) a method of feature-level sentiment analysis for Brazilian Portuguese reviews using ontology. As part of our propose we analysed a set of varied resources and techniques, such as Part-Of-Speech (POS)

[1] http://www.comscore.com

taggers, sentiment lexicons and linguistic rules. For the identification of features, both explicit and implicit, we propose the use of concepts, properties and hierarchy of ontology.

1.3.2 Specific Objectives

- as there are not available resources for this study, we need creating and preparing a corpus from the Web through Web crawler and annotation tool;
- in order to identify explicit features, we employed to the domain ontology concepts as indicator of explicit features;
- in order to identify implicit features, we employed to the domain ontology properties and hierarchy as indicator of implicit features;
- we use sentiment lexicons and linguistic rules to determine the sentiment orientation from reviews;
- we produce a summary of reviews of each entity through different models (textual and graphic);
- we evaluate the proposed approach using some metrics (precision, recall, and f-measure).

1.4 Contributions

We consider the definition, implementation and evaluation of a sentiment analysis method at the feature-level using ontologies in Brazilian Portuguese language as the main contributions of this thesis.

We believe that the development of a tool to capture hotel reviews in the TripAdvisor Web page [TRI15] (Web crawler) and the development of a tool to annotate reviews at the feature-level considering domain ontology are relevant.

Finally, another contribution is the manual evaluation and translation for Brazilian Portuguese of domain ontology, the creation of guidelines for corpus annotation in feature-level (explicit and implicit), the share of the annotated corpus, and the evaluation of several available resources required for sentiment analysis in Portuguese.

1.5 Thesis Organization

The remainder of this thesis is organized as follows. Chapter 2 presents the background of the research field, including sentiment analysis, concepts related to Natural Language Processing (NLP), ontologies, sentiment lexicons, and linguistic rules needed to the reader's understanding of the proposed method. Chapter 3 presents related works. In Chapter 4 we discuss the thesis, focusing on the feature-level sentiment analysis applied to Brazilian Portuguese reviews. Chapter 5 shows an experimental evaluation. In Chapter 6 we present our final remarks, and perspectives for future work. This thesis ends with the bibliographic references.

2. Theoretical Background

This chapter provides several essential concepts to conduct this study. In the following sections we describe about: sentiment analysis (Section 2.1), concepts related to NLP (Section 2.2), ontologies (Section 2.3), sentiment lexicons (Section 2.4), and linguistic rules (Section 2.5).

2.1 Sentiment Analysis

Sentiment analysis is the field of computer science that studies methods to analyze people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities and their aspects. This field is also known as: opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, *etc.* [LIU12].

Sentiment analysis techniques can be grouped into supervised and unsupervised approaches [JOS14]. In the supervised approaches, the sentiment analysis is considered a classification problem. Usually, techniques (e.g., bag-of-words, n-grams, word position) and machine learning (e.g., Support Vector Machine - SVM, Naïve Bayes - NB, Maximum Entropy - ME) are used. In the unsupervised approach, the sentiment analysis depends on external resources (e.g., sentiment lexicon) and a group of heuristics based on linguistic and domain knowledge.

Still, according to Liu [LIU11] exist two main ways to express sentiments: regular and comparative. A regular opinion is often simply called opinion and usually describes some aspects of one entity, and a comparative opinion expresses a relation of similarities or differences between some aspects of two or more entities.

2.1.1 Opinion

According to the Oxford Dictionary for Advanced Learners [HOR10] an opinion represents "feelings or thoughts about somebody/something, rather than a fact".

Quirk *et al.* [QUI85] define an opinion as a subjective statement or thought about an issue or subject, representing a positive or a negative impact.

Liu [LIU12] states that formally an opinion is a quintuple (e_i, a_{ij}, s_{ijkl}, h_k, t_l), where:

- e_i – a given entity: an entity may be a product, person, event, organization, topic, *etc;*
- a_{ij} – j aspect of e_i: an aspect is a set of attributes or components of an entity;
- s_{ijkl} – the sentiment orientation of h_k at t_l about a_{ij}: a sentiment orientation indicates whether the opinion is positive, negative or neutral;
- h_k – an opinion holder: an opinion holder is a person or organization expressing the opinion;
- t_l – an opinion time: an opinion time is a date, important to know changes of opinions over time.

In this work, from the five elements of the quintuple presented above we consider aspect and sentiment orientation.

2.1.2 Levels of Opinion Analysis

Bhuiyan *et al.* [BHU09] classify the research on sentiment analysis in three main directions: (1) document-level, (2) sentence-level, and (3) feature-level (as shown in Figure 2.1). Classifying reviews at the document-level or the sentence-level does not tell exactly what the opinion holder likes and dislikes. Feature-level investigates ways to classify each aspect. On the other hand, the classification of reviews at the feature-level is harder to perform. However, it is the most useful method because it provides detailed results [WES14].

Document level

The task at this level is to determine whether a whole document expresses a positive or negative sentiment. This level of analysis assumes that each document expresses opinions on a single entity [JOS14] as film review.

Sentence level

The task at this level is to analyze the sentences in the document and to determine whether each sentence expresses a positive, negative or neutral opinion. This level of analysis is closely related to subjectivity classification, which distinguishes objective sentences from subjective sentences [JOS14].

Corpus-based and Dictionary-based approach

Corpus-based approaches find co-occurrence patterns in WordNet [FEL98] to determine the sentiments of words or phrases [HAT00] [TUR02]. And, dictionary-based approaches use synonyms and antonyms in WordNet [FEL98] to determine word sentiments based on a set of seed opinion words.

Feature level

At this level, a finer-grained analysis is performed. Instead of looking at language constructs (documents, paragraphs, sentences, clauses or phrases), feature-level analyses directly look at the object of the object of the opinion itself [JOS14].

Using Lists, Taxonomies and Ontologies

The two main families of work in the feature extraction process are: those that extract a simple list of features and that organize them into hierarchy using taxonomies and ontologies (details in Chapter 3).

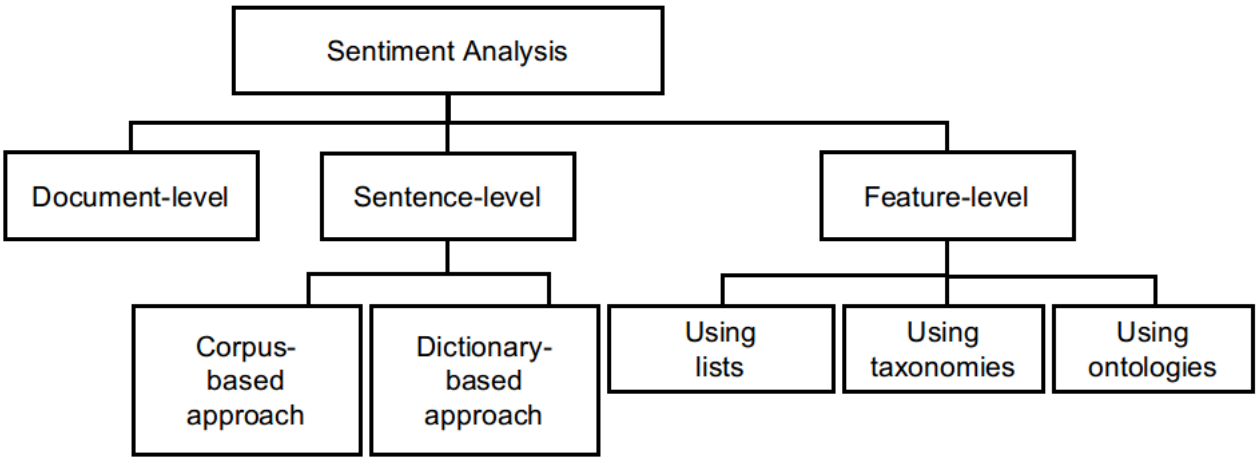

Figure 2.1: Classification of opinion mining research. Adapted from [BHU09]

.

As previously mentioned, sentiment analysis has several application domains, such as marketing, public relations, and political campaigns. In this study, we focus on regular opinions about the accommodation sector. We conducted the following tasks: we identified the entity features in reviews, we decided whether the opinions about the features were positive or negative, and, at last, we summarized the polarity of each feature mentioned in the reviews.

2.2 Concepts Related to Natural Language Processing (NLP)

NLP is defined as a computer's ability to process the same language that human beings use every day. Words are basic expressions of language.

The language is a means of communication that is organized in a level system with complex rules. Each level of this system deals with a specific aspect of the communication process and is composed by its own elements and combination rules. According to Sowa [SOW84], the language levels are:

- Prosody: the rhythm patterns and intonation of the language;
- Phonology: sounds or phonemes of the language;
- Morphology: the high level of significant elements or morphemes, which build the words;
- Syntax: the rules to combine words into clauses/phrases or sentences;
- Semantics: the meaning and expression;
- Pragmatics: the use of language and its effect on the listener.

An overview on these levels is presented in Figure 2.2. In this thesis, we analyze opinion texts for the Portuguese language in the Morphosyntactic (Morphology + Syntax) and Semantics levels.

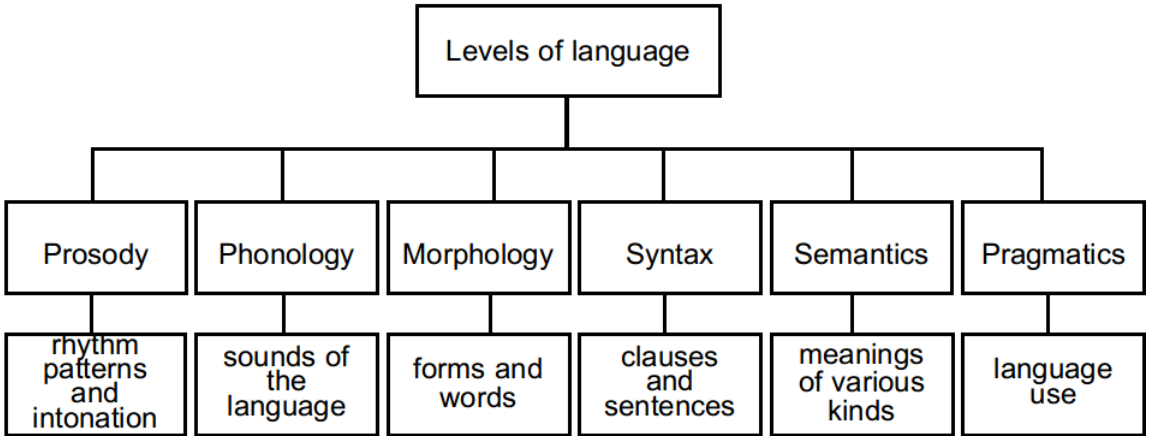

Figure 2.2: Levels of language. Adapted from [HIC05]

.

2.3 Ontologies

A common definition in literature is that ontologies are explicit specifications of conceptualizations, where conceptualizations are simplified summaries of the world we want to represent for some purpose [GRU95]. According to Nicola Guarino [GUA98], ontologies can be developed in different levels (Figure 2.3), such as top-level, domain, task, and application. Top-level ontologies describe general concepts such as time, space and others. Domain and task ontologies specialize the terms introduced in the top-level ontology and describe the vocabulary related to a generic domain or task. Application ontologies describe concepts depending on a particular domain and task.

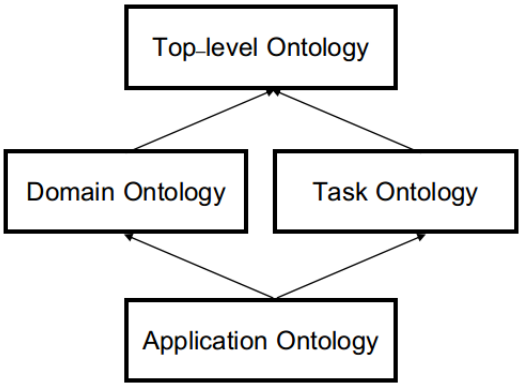

Figure 2.3: Categorization of ontologies according to Guarino. Source: [GUA98]

.

Lassila and McGuinness [LAS01], in turn, classify Web ontologies according to the richness of their structures. The following categories can be seen in the spectrum (Figure 2.4): catalog/id, terms/glossary, thesaurus, informal hierarchy *is-a*, formal hierarchy *is-a*, formal hierarchy with instances, frames/properties, value constraint, disjoint, reverse, and *part-of* and logical constraint.

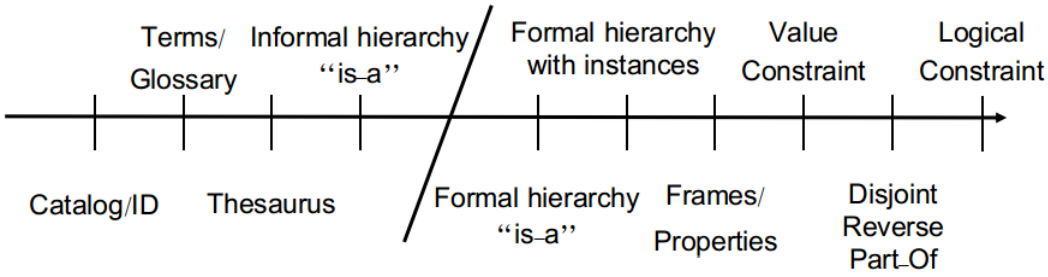

Figure 2.4: Categorization of ontologies according to Lassila. Source: [LAS01]

.

In this study, we use domain ontologies that describe vocabularies in the Web Ontology Language (OWL). OWL [SMI04] is a stable specification developed by the Web Ontology Working Group. It is considered a Web standard for industry and academy.

Below, we describe some of the elements that constitute the OWL ontologies (concepts, properties, and instances) which are important for this work.

2.3.1 Concepts

In OWL ontologies created with the tool Protégé [PRO15], each class defined by users is a subclass of $owl:Thing$. Classes are defined simply when we name the classes. Empty classes are defined using the syntax $owl:Nothing$.

Fragment 2.1: Example of classes.

$<owl:Class\ rdf:ID="Serviço"/>$
$<owl:Class\ rdf:ID="Funcionário"/>$

For HOntology [CHA12b] (Fragment 2.1), we can see the existence of the classes $Serviço$ and $Funcionário$, which are defined using the syntax $rdf:ID$.

Fragment 2.2: Example of subclasses.

$<owl:Class\ rdf:about="\#Recepcionista">$
$<rdfs:subClassOf\ rdf:resource="\#Funcionário"/>$
$</owl:Class>$

The syntax $rdf:ID="Funcionário"$ is used to introduce a name as part of the definition. We reference a class using the character $\#$, such as, $rdf:about=\#Recepcionista$ and $rdf:resource=\#Funcionário$, shown in Fragment 2.2.

The hierarchy constructor for classes is $rdfs:subClassOf$ which relates a more specific class to a more general class. In this case, $Recepcionista$ is a subclass of $Funcionário$. The relation $rdfs:subClassOf$ is transitive.

The tags $rdfs:comment$ and $rdfs:label$ (support multiple languages) are optional comments, that intend to provide a readable name to classes and not contribute to the logical interpretation of an ontology.

2.3.2 Properties

In OWL, there are two types of properties: object properties, which are the relations between instances of two classes ($ObjectProperty$) and data properties, which are relations between instances of classes and literals ($DatatypeProperty$).

The property can be defined as a specialization of an existing property ($subPropertyOf$). The properties can be organized in hierarchies.

Fragment 2.3: Example of properties (relations).

$< owl : ObjectProperty\ rdf : ID = \text{“}temPaísDeOrigem\text{”} >$
$< rdfs : domain\ rdf : resource = \text{“}\#TipoDeHóspede\text{”} / >$
$< rdfs : range\ rdf : resource = \text{“}\#País\text{”} / >$
$< / owl : ObjectProperty >$

In Fragment 2.3 the property $temPaísDeOrigem$ has a domain $TipoDeHóspede$ and a range $País$. In other words, the property relates the instances of the class $TipoDeHóspede$ to the instances of the class $País$.

Fragment 2.4: Example of properties (attributes).

$< owl : DatatypeProperty\ rdf : ID = \text{“}temArCondicionado\text{”} >$
$< rdfs : domain\ rdf : resource = \text{“}\#Quarto\text{”} / >$
$< rdfs : range\ rdf : resource = \text{“}http://www.w3.org/2001/XMLSchema\#boolean\text{”} / >$
$< / owl : DatatypeProperty >$

The property $temArCondicionado$, in Fragment 2.4, refers to the relation between instances of classes and data type ($string$, $boolean$ and $integer$), such as the relation between $Quarto$ and $boolean$.

Some characteristics of the properties are: Symmetric ($owl : SymmetricProperty$), Inverse ($owl : inverseOf$), Transitive ($owl : TransitiveProperty$), Functional ($owl : FunctionalProperty$) and Inverse Functional ($owl : InverseFunctionalProperty$).

Constraints can also be imposed on the properties. They can be of value type ($allValuesFrom$, $someValuesFrom$ and $hasValue$) or of the cardinality type ($maxCardinality$ and $minCardinality$).

2.3.3 Instances

Fragment 2.5: Example of instances.

$< Hotel\ rdf : ID = \text{“}HotelContinentalBusiness\text{”} / >$
$< Hotel\ rdf : ID = \text{“}HotelMinuanoExpress\text{”} / >$

In OWL, every instance is a class member of $owl : Thing$. In Fragment 2.5, the instances $HotelContinentalBusiness$ and $HotelMinuanoExpress$ belong to the class $Hotel$.

2.4 Sentiment Lexicons

Sentiment words are words that indicate sentiments. Good, excellent and correct are examples of positive sentiment words, in the same way that bad, poor and wrong are examples of negative sentiment words. In the literature, sentiment words are also called opinion words, polar words, or opinion-bearing words. A sentiment lexicon is a collection of sentiment words. Moreover, sentiment lexicons can contain idiomatic expressions with the same capability to indicate sentiment [LIU12].

In the literature, there are three main approaches to compile sentiment words. It is possible to collect sentiment words manually (manual approach) or to automate this process through a dictionary-based or a corpus-based approach.

In the dictionary-based approach, a small set of sentiment words (seeds) with known positive or negative orientation is used. Then, this set grows by searching for their synonyms and antonyms in the WordNet, for example. In the corpus-based approach, a general-purpose sentiment lexicon is adapted to a new one using a domain corpus.

In English, there are some well-known sentiment lexicons, such as SentiWordNet 3.0 [BAC10], WordNetAffect [STR04], Affective Norms for English Words (ANEW) [NIE11], and Liu's English Opinion Lexicon [HU04], among others. Most sentiment analysis approaches use SentiWordNet 3.0 [BAC10], a fragment of WordNet 3.0, which is manually annotated for positivity, negativity and neutrality. This resource has nearly 117,000 words.

In Portuguese, as far as we know, there are just four sentiment lexicons: OpLexicon [SOU11], Brazilian Portuguese LIWC Dictionary [ALU13], SentiLex [SIL12], and synsets with polarity of Onto.PT [OLI14].

OpLexicon [SOU11] has 30,322 words (23,433 adjectives and 6,889 verbs) and was built based on a Brazilian Portuguese corpus (composed of 346 movie reviews and 970 journalistic texts), a thesaurus TEP [DIA03] (from the Portuguese, *Thesaurus Eletrônico Básico para o Português do Brasil*) and the translated Liu's English Opinion Lexicon [HU04]. The results of each of these techniques are combined to create a large lexicon for Brazilian Portuguese.

Brazilian Portuguese LIWC Dictionary [ALU13] was built from the original English LIWC Dictionary [PEN01] and has 127,149 entries, where each entry can be assigned to one or more categories.

SentiLex [SIL12] finds which adjectives can be used as human modifiers and then assigns them a polarity attribute. The resource is available in two files, one where the word entries are inflected and other where all entries are lemmatized. The first file covers 82,347 lemmas, of which 16,863 are adjectives, 1,280 are nouns, 29,504 are verbs and 34,700 are idiomatic expressions. The second file covers 7,014 lemmas (5,473 manual and 1,541 automatic; 4,596 negative, 1,548 positive and 860 neutral), of which 4,779 are adjectives, 1,081 are nouns, 489 are verbs, and 666 are idiomatic expressions.

Synsets with polarity of Onto.PT [OLI14] contain 10,318 synsets with assigned polarity and tries to cover the entire language and not just a specific domain. The resource was constructed in two steps. Initially, the polarity of Onto.PT synsets was assigned using SentiLex as the polarity reference.

After, the polarity was propagated through semantic relations.

2.5 Linguistic Rules

Negation is a very common linguistic construction that affects polarity. In the literature, we found some surveys about negation in sentiment analysis, such as Shah and Rekh [SHA14], Wiegand *et al.* [WIE10], and the precursors of this area: Polanyi and Zaenen [POL06], Kennedy and Inkpen [KEN06], and Wilson *et al.* [WIL05].

According to Shah and Rekh [SHA14], the negation is present in all human languages and changes text polarity. However, the presence of a negation word in a sentence does not mean that all words conveying sentiments will be inverted. Polanyi and Zaenen [POL06] describe how the base attitudinal valence of a lexical item is modified by sentence and discourse contexts. By combining words with positive valence with negation words ('*não*' [not], '*nunca*' [never], '*nenhum*' [none], 'ninguém' [nobody], '*nada*' [nowhere, nothing] and '*nem*' [neither]), the authors flip the positive valence to the negative valence.

The effect of valence shifters (negations, intensifiers and diminishers) on polarity when classifying the reviews at the document level is examined in Kennedy and Inkpen's [KEN06] work.

Wilson *et al.* [WIL05] carry out more advanced negation modeling in the expression-level polarity classification. The study uses supervised machine learning where negation modeling is encoded as features using polar expressions.

Some studies have hypothesized that adjectives separated by 'and' have same polarity, while those separated by 'but' have opposite polarity [HAT97]. Therefore, conjunction rules (additives and adversatives) could be used to determine the sentiment orientation.

Still, the position the adjective occupies in the sentence has a relevant role. In Portuguese, a qualifier adjective, can also be found on the left side of the noun, and this position implies a more subjective interpretation [NEV11].

3. Related Work

This chapter presents feature-based sentiment analysis. Works in English and Brazilian Portuguese. After, we show studies on the accommodation domain.

3.1 Feature-Level Sentiment Analysis in English

In the following subsections, we present some works that assist fine-grained (feature-level, aspect-level, phrase-level, and word-level) sentiment analysis. The first uses lists (Subsection 3.1.1), and the others use taxonomies (Subsection 3.1.2) or ontologies (Subsection 3.1.3).

3.1.1 Using Lists

The main related studies concerning feature-level sentiment analysis using lists are Hu and Liu's [HU04] and Popescu and Etzioni's [POP05].

Hu and Liu [HU04] focus on finding features that appear explicitly in the reviews applying the association rule algorithm. To identify features (nouns and noun phrases) in the reviews, the authors used the linguistic parser NLProcessor [NLP15]. In their study, only nouns and noun phrases with more than a determined threshold are kept. The recognition of opinion orientation of each sentence is based on the dominant orientation of the opinion words in the sentence. Although this method is very simple, it is actually quite effective [LIU12].

To improve the approach by Hu and Liu, Popescu and Etzioni [POP05] suggest removing noun phrases that may not be features of entities (OPINE system). The OPINE system uses an unsupervised classification technique (relaxation labeling) to find the opinion orientation of potential opinion words in the context of the extracted features and specific review sentences.

The main limitation of these approaches is that a great number of features is extracted [CAR05].

3.1.2 Using Taxonomies

The use of taxonomies is proposed in order to improve the results obtained in feature-level sentiment analysis using lists by Carenini *et al.* [CAR05] and Gamon *et al.* [GAM05].

Carenini *et al.* [CAR05] use taxonomy of features developed by domain experts. The method was based on some similarity metrics defined using string similarity, synonyms, and lexical distances measured using WordNet [FEL98] [WOR15]. It merges each discovered aspect expression with a node in the taxonomy based on the similarities. The main limitation of the approach is that it is very dependent on the effectiveness of the similarity metrics used. In addition, taxonomies are difficult to construct, time consuming and expensive [MIA10].

Gamon *et al.* [GAM05] automatically extract a taxonomy from the reviews. After, the authors extract sentences of reviews for each leaf node. To train the sentiment classifier, a small random

set of sentences is label as positive, negative or other (PULSE system), the sentences are assigned to clusters and, in the end, a summary that can be more or less detailed is produced. The main limitation of this approach is that supervised learning is dependent on the training dataset.

3.1.3 Using Ontologies

To improve the results obtained in feature-level sentiment analysis using taxonomies, the use of ontologies is proposed. The main related studies using ontologies are Zhou and Chaovalit [ZHO08], Zhao and Li [ZHA09], and, Peñalver-Martínez *et al.* [PEN14].

In Zhou and Chaovalit [ZHO08], the Ontology Supported Polarity Mining (OSPM) architecture is proposed. The use of ontologies has the potential to refine and improve the process of sentiment analysis by identifying properties and relations between concepts. In the architecture, reviews extracted from the Internet are preprocessed. After, each review is parsed in order to extract and map text segments according to the ontology. Finally, a polarity orientation is generated for each text segment and for the text as a whole. The advantage of the use of domain ontology is that it provides detailed topic-specific information.

The method proposed by Zhao and Li [ZHA09] integrates preprocessing, feature identification, polarity identification, and sentiment classification. The preprocessing step includes word segmentation and POS tagging. Ontology terminologies are used in the feature identification step. In the polarity identification step, the polarity measurement relies on a sentiment lexicon (SentiWordNet [SEN15]). Finally, the polarity obtained in the previous step is converted to a more precise polarity by analyzing the context, such as negation and conjunction rules. The authors use ontologies, because ontologies provide knowledge about specific domains that are understandable by both developers and computers.

The main goals of Peñalver-Martínez *et al.*'s [PEN14] study is to improve feature-level sentiment analysis by employing ontologies in the selection of features and to provide a new method based on vector analysis. This approach consists of four modules: the NLP module that uses toolkit GATE; the ontology-based feature identification module where each feature receives a different importance, for instance, the features that are more often cited by users in their opinions will be more relevant; the polarity identification module that calculates the polarity of the feature using SentiWordNet and the summarization module. The main drawback of the proposal is that an ontology has to be provided in order to model the features of a predefined domain.

In summary, the feature identification is guided by a domain ontology, built manually in [ZHO08] and semi-automatically in [ZHA09] and [PEN14]. The ontologies of the latter case are enriched by a process of automatic term extraction. The accuracy obtained by Peñalver-Martínez *et al.* [PEN14], Zhao and Li [ZHA09], and Zhou and Chaovalit [ZHO08] are 84.8%, 78.7%, 72.7% for the positive category and 87.1%, 80.6%, 74.1% for the negative category, respectively. All experiments were performed in the movie domain.

3.1.4 SemEval 2014 (Task 4)

SemEval is an ongoing series of evaluations of computational semantic analysis systems. The evaluations intend to explore the nature of meaning in language. One of the tasks of SemEval 2014 (Task 4), which took place in Dublin, Ireland, was Aspect Based Sentiment Analysis (ABSA).

The goal of this to identify the aspects of given target entities and the sentiments expressed towards each aspect. The task consisted of the following subtasks: aspect term extraction, aspect term polarity, aspect category detection, and aspect category polarity.

Datasets consisting of customer reviews with human annotations identifying the mentioned aspects of the target entities and the sentiment polarity of each aspect were provided. The datasets consisted of over 6,500 sentences about laptops and restaurants.

The ABSA task [PON14] attracted 163 submissions from 32 teams. The evaluation ran in two phases. In Phase A, the participants were asked to return the aspect terms and aspect categories for the provided test datasets. In Phase B, the participants were given the gold aspect terms and aspect categories for the sentences of Phase A and were asked to return the polarities of the aspect terms and the polarities of the aspect categories of each sentence.

In Phase A (the aspect terms), the IHS_RD [CHE14] obtained the best result for the laptop domain, with 74.55% of f-measure, and the DLIREC [TOH14] has the best result for the restaurant domain, with 84.01% of f-measure. IHS_RD relied on Conditional Random Fields (CRF) with features extracted using named entity recognition, POS tagging, parsing, and semantic analysis. DLIREC, also uses a CRF, along with POS tagging and dependency tree.

In Phase B (the polarities of the aspect terms), the DCU [WAG14] and the NRC-Canada [KIR14] were identical: 70.48% of f-measure for the laptop domain, and the DCU performed slightly better: 80.95% of f-measure for the restaurant domain. DCU and NRC-Canada relied on SVM with features mainly based on n-grams, parse trees, and sentiment lexicons.

3.2 Feature-Level Sentiment Analysis in Brazilian Portuguese

In the literature, some approaches applied to Portuguese are: Siqueira and Barros [SIQ10], Ribeiro *et al.* [RIB12], Chaves *et al.* [CHA12a], and Baracho *et al.* [BAR12].

Siqueira and Barros [SIQ10] present a domain free process for feature extraction based on frequent nouns identification, relevant nouns identification, feature indicators mapping, and unrelated nouns removal. This process receives a text containing an opinion as input and returns the extracted features. The system uses a manually compiled lists of 20 feature indicators for the chosen domain to identify features that are implicitly mentioned by using any of these indicators. The experiments were applied on the services domain, and the corpus used was collected from E-bit [EBI15] (200 opinions on Brazilian online stores). Portuguese TreeTagger was used to perform the POS tagging of the corpus.

In Ribeiro *et al.* [RIB12], an adaptation of Hu and Liu's study [LIU12] using feature-based propagation, simple propagation and a general opinion lexicon (OpLexicon) was proposed. Furthermore,

a new approach based on supervised learning algorithms (SVM and NB) was also proposed, and the authors compared these two approaches. The results produced by the lexicons were significantly worse than the results produced by the classification-based methods. The experiments were applied in the vehicle domain, and the corpus used was collected from Carrosnaweb [CAR15] and blogs about cars. The feature extraction step was performed using a method based on grammar dependency trees. FreeLing [FRE15] was used to generate the parse tree and DepPattern [DEP15] was used to generate the dependence tree.

In Chaves *et al.* [CHA12a], the PIRPO algorithm received as input, a set of reviews that were pre-processed in order to extract sentences and to detect which reviews were split into positive and negative segments. This approach used external resources, such as the list of adjectives from OpLexicon and HOntology [HON15]. The polarity was identified based on the values of the list of adjectives.

Baracho *et al.* [BAR12], proposed a methodology that combined text-processing technologies. Morphological, syntactic and semantic analyses guided by the target domain terms supplied by vehicle ontologies and by an opinion lexicon (SentiStrenght [THE12]) translated into Portuguese, were used to detect and classify sentiments. The software PALAVRAS [BIC00], an automatic parser for Portuguese, was used to do a semantic analysis of the text.

Siqueira and Barros [SIQ10] obtained a 77.24% precision and a 90.94% recall in the extraction of implicit and explicit features. In Ribeiro *et al.*'s [RIB12] study, although SVM showed good performance for overall sentiment analysis, NB performed significantly better for feature-based analysis. Preliminary results in Chaves *et al.* [CHA12a] indicate an average f-measure of 32%. In Baracho *et al.* [BAR12], the results obtained about the cars Palio, Gol and Corsa are 63%, 47%, 38% positive and 37%, 53%, 62% negative, respectively. Gol showed more mixed opinion results, Palio had a more positive opinion, and Corsa presented a more negative opinion.

Siqueira and Barros [SIQ10] use a simple list of words to explicit and implicit feature identification, we choose to use a more formal resource, domain ontologies. We propose to use ontology concepts as explicit features, in the same way that Baracho *et al.* [BAR12] and Chaves *et al.* [CHA12a], and ontology properties as implicit features, which as far as we known was not done yet for Portuguese language.

In the same way that most of literature works we also use a single POS tagger tool in our proposed approach, however we analysed three different POS tagger tools to choose the best one for our experiments. Likewise, we analysed four different sentiment lexicons to choose the best one for our experiments.

We evaluated the TreeTagger (used by Siqueira and Barros [SIQ10]), FreeLing (used by Ribeiro *et al.* [RIB12]), and CitiusTagger [CIT15], and we choose TreeTagger. Moreover, we evaluated the OpLexicon (used by Ribeiro *et al.* [RIB12] and Chaves *et al.* [CHA12a]), SentiLex, LIWC-PT and synsets with polarities from Onto.PT, and we choose synsets with polarities from Onto.PT.

Finally, in Table 3.1, we show an overview of the related work (technique, kind of features, domain, tool, resource, and language). We applied our proposal in accommodation domain, likewise

Chaves *et al.* [CHA12a].

It is important to note that have few researches in Portuguese sentiment analysis. There is lack of studies that consider implicit features using ontology as we propose in this thesis. Our proposal is a single exploring different POS tagger tools and sentiment lexicons. We choose the accommodation domain because it is a sector that has a significant influence on customers.

3.3 Sentiment Analysis in the Accommodation Domain

In the literature, we find some previous research on Sentiment Analysis applied to the accommodation domain, such as Ortiz *et al.* [ORT10], Haruechaiyasak *et al.* [HAR10], Kasper and Vela [KAS11], and Chaves *et al.* [CHA12a].

In Ortiz *et al.* [ORT10], the Sentitext tool was presented. The system is entirely based on linguistic knowledge and is independent of any domain. The authors tested their tool in a set of Spanish hotel reviews from Tripadvisor (100 reviews about hotels in London).

In Haruechaiyasak *et al.* [HAR10], a framework for feature-based opinion mining was proposed. To evaluate the proposed framework, the authors performed some tests in hotel reviews. A set of Thai hotel reviews from Agoda [AGO15] (8,436 reviews) was used. The accommodation sector is ranked as one of the top industries in tourism in Thailand.

Kasper and Vela [KAS11] presented a system that captures comments from the Web and creates a structured overview of such comments. The authors evaluated 1,559 German hotel reviews crawled from the Web and manually classified considering their polarity (positive, negative or neutral).

Chaves *et. al.* [CHA12a] dealt with reviews about small and medium hotels in the Lisbon area. The information sources were Tripadvisor and Booking. The dataset consisted of 1,500 reviews from January 2010 to April 2011 in Portuguese, English and Spanish, 180 of which were in Portuguese.

The number of reviews vary according to language (e.g.: Spanish, Thai, German, and Portuguese). We created our dataset in accommodation domain with 194 Brazilian Portuguese reviews from TripAdvisor because this number it is considered satisfactory to perform a manual annotation.

Table 3.1: Overview of the related work.

Work	Technique	Kind of Features	Domain	Tool	Resource	Language
[HU04]	List	Explicit	Product	NLProcessor		English
[POP05]	List	Explicit	Hotel and Product			English
[CAR05]	Taxonomy	Explicit	Product		WordNet	English
[GAM05]	Taxonomy	Explicit	Car			English
[ZHO08]	Ontology	Explicit	Movie		General Inquirer	English
[ZHA09]	Ontology	Explicit	Movie		SentiWordNet	English
[PEN14]	Ontology	Explicit	Movie	GATE	SentiWordNet	English
[SIQ10]	List	Explicit and Implicit	Services	TreeTagger		Portuguese
[RIB12]	List	Explicit	Car	FreeLing DepPattern	OpLexicon	Portuguese
[CHA12a]	Ontology	Explicit	Hotel		OpLexicon	Portuguese
[BAR12]	Ontology	Explicit	Car	PALAVRAS	SentiStrenght PT	Portuguese
Our Proposal	Ontology	Explicit and Implicit	Hotel	TreeTagger FreeLing CitiusTagger	OpLexicon SentiLex LIWC PT Synsets with polarity from Onto.PT	Portuguese

4. Proposed Method

This chapter presents the proposed method for sentiment analysis based on features using ontologies (Figure 4.1).

Similar to approaches [ZHA09] and [PEN14], our study is organized into four main steps. Initially, the method receives as input a set of reviews, which are preprocessed. After, features are identified in the preprocessed reviews using domain ontology. The polarity is identified in the preprocessed reviews containing features using sentiment lexicons and linguistic rules. Finally, a summary with features and their respective polarities is generated.

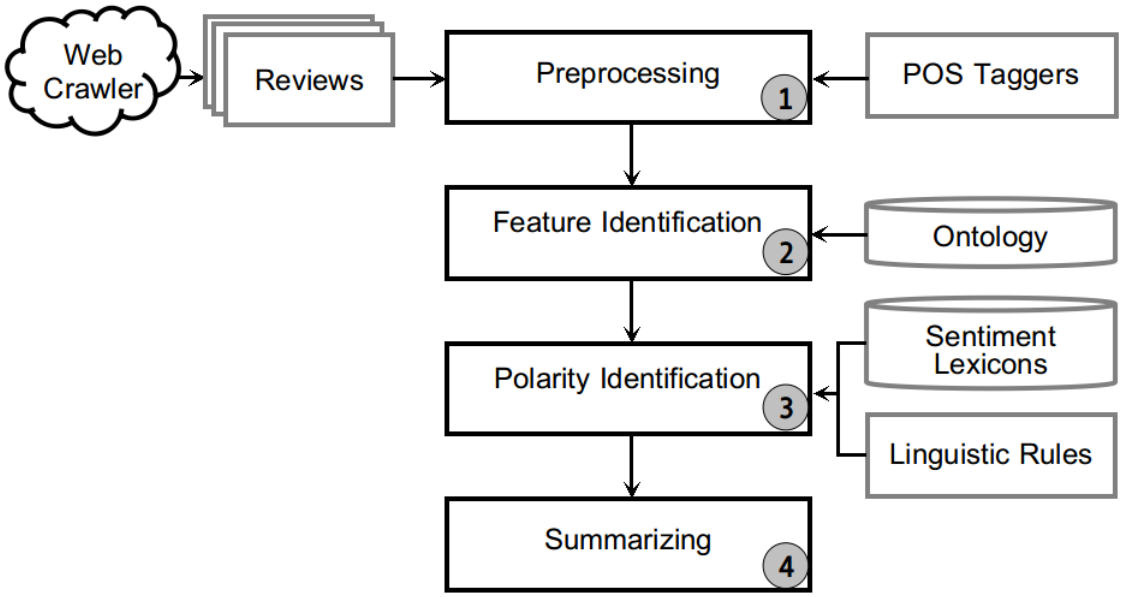

Figure 4.1: Overview of the method.

The steps of preprocessing, feature identification, polarity identification and summarizing are described in detail below.

4.1 Preprocessing

Initially, we collected data through a Web crawler. According to Olston and Najork [OLS10], a Web crawler is a system for the bulk downloading of Web pages. Web crawlers are one of the main components of Web search engines. After, a set of reviews is preprocessed. The main objective of this step is to obtain the grammatical categories and lemma using POS taggers (Figure 4.2).

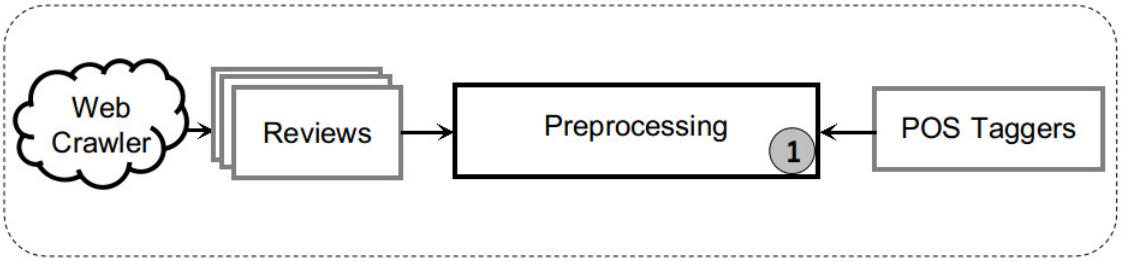

Figure 4.2: Preprocessing.

POS taggers divide words into grammatical categories, based on the role they play in the context in which they appear. Most tools make use of the same basic grammatical categories (noun, verb, adjective, adverb, *etc.*) [FEL07]. Some systems contain a much more elaborate set of tags, such as Portuguese TreeTagger [TRE15]. For instance, adjectives may be marked with the following classification: AQ0, AQA, AQC and AQS (Adjective; Qualifier; and their degree ['0' for default form; 'A' for Augmentative; 'C' for Diminutive; 'S' for Superlative]), nouns may be marked with the following classification: NCCP, NCCS and NCCI (Noun; Common; Common; and their number ['P' for Plural; 'S' for Singular; 'I' for Invariable]).

Figure 4.3 shows the sentence "As suítes são boas e baratas." ["The suites are good and cheap."] marked with Portuguese TreeTagger, where: DA0 is a determinant article, NCCP is a plural common noun, VMI is an indicative main verb, AQ0 is a qualifier adjective, CC is a coordinated conjunction, Fp a is punctuation.

Token	As	suítes	são	boas	e	baratas	.
Grammatical Category	DA0	NCCP	VMI	AQ0	CC	AQ0	Fp
Lemma	o	suíte	ser	bom	e	barato	.

Figure 4.3: Example of preprocessing.

4.2 Feature Identification

The features are identified in the preprocessed reviews using domain ontology (Figure 4.4). Ontologies are used because they provide a formal and structured knowledge representation, with the advantage of being reusable. The literature shows that there are different levels of knowledge representation; some are complex structures [ZHO08] [PEN14] and some are simple structures [ZHA09]. It is important to notice that the same review can be presented at different levels of knowledge. For instance, at the high level an opinion may refer to a 'refeição' ['meal'] whereas at the low level the same review may refer to a 'café da manhã' ['breakfast'].

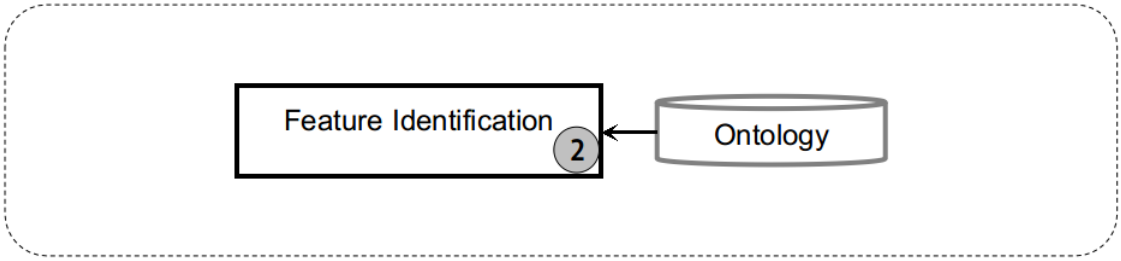

Figure 4.4: Feature identification.

Our work focuses on obtaining an ontology-based feature level method in the same line of [PEN14], using previously-developed ontologies but considering Portuguese language. We divide the feature identification focusing on the extraction of explicit and implicit features in reviews.

POS taggers are efficient for explicit feature extraction in terms of accuracy [HU04]. There are four main approaches to extract explicit features: based on frequent nouns and noun phrases, exploiting opinion and target relations, using supervised learning and using topic modeling [LIU12]. In the first moment, we use concepts of domain ontology in order to identify explicit features (e.g.; 'As **suítes** são boas e baratas.' ["The suites are good and cheap."], 'suíte' ['suite'] is a concept of ontology, 'suítes' ['suites'] is an explicit feature).

Implicit features are the features which are not apparent in review [PAN08]. For instance, in review "O hotel é caro" ["The hotel is expensive"], user is referring to 'preço' ['price'] although word 'preço' ['price'] is not explicitly mentioned [ASG14]. We use the concept hierarchy of the domain ontology in order to identify implicit features (e.g.; 'As **suítes** são boas e baratas.' ["The suites are good and cheap."], 'suíte' ['suite'] is-a 'quarto' ['room'], 'quarto' ['room'] is an implicit feature). Properties are also considered. If two concepts are related by a property, one concept is considered as implicit to the other as in the case of 'Localização possuiEndereço Endereço' ['Location hasAddress Address'], where 'Endereço' ['Address'] is then considered as implicit to 'Localização' ['Location'].

4.3 Polarity Identification

The polarity is identified in the preprocessed reviews containing features using sentiment lexicons and linguistic rules (Figure 4.5). We search the words in the list of adjectives extracted from Portuguese sentiment lexicons, close to the concept of the ontology. Thus, we applied rules that consider the position of the adjective and negation words. At least one word of the list of adjectives must be close to the concept for the feature polarity to be calculated.

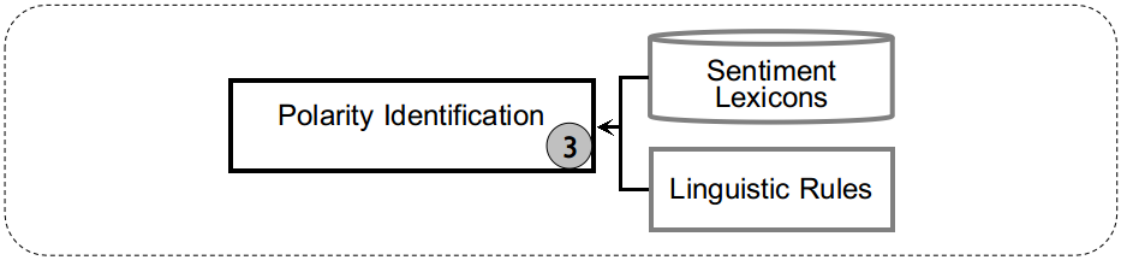

Figure 4.5: Polarity identification.

The complex scope of the negation model in sentiment analysis depends on the language. In Brazilian Portuguese, there are at least three ways of verbal negation. They are: one standard pre-verbal form, in which the negative particle appears before the verb (Figure 4.6) and two nonstandard forms, one post verbal, in which the particle appears after the verb, and one in which there is double negation. In this case the verb is surrounded by two negation particles, one before and one after the verb [SCH05]. Note that, differently from logical rules, double negation in natural language does not make the sentence positive. In the experiments conducted for this thesis, all three forms of negation found in Portuguese were considered. We chose to apply a simple rule that consists of inverting the polarity of the opinion word if the negation particle ('não' ['no'], 'nunca' ['never'], 'nada' ['nothing'], 'nem' ['neither'], 'nenhum' ['none'], 'ninguém' ['nobody']) appears in any of the three verbal negation forms.

... **localização** para sair à noite *não* é *legal* .

-1

Figure 4.6: Example of a negative particle appearing before the verb.

The position the adjective occupies within the noun phrase in Portuguese has a relevant role. Neves [NEV11] affirms that, in Portuguese, adjectives that are to the right of the noun are in default position, which means that this is the less marked position. Classifier adjectives are mostly found in this position. A qualifier adjective, on the other hand, can also be found on the left side of the noun, and this position implies a more subjective interpretation. The author points out that this position is associated to a more specific and restrictive interpretation. The left-head position is more marked, that is, it is not so usual, and that is why it is more subjective and triggers some special meaning effects. In the tests conducted for this thesis, this model searches for adjectives before the aspect attributing polarity to it (Figure 4.7). If this does not happen, the adjectives are sought to the right of the aspect (Figure 4.8). Depending on the case, this operation is performed until another aspect is found or until the end of the sentence.

... e um *bom* **atendimento** na **recepção** .

+1

Figure 4.7: Example of adjective position before feature.

A **localização** é *boa* e o **preço** é bem adequado .

$+1$

Figure 4.8: Example of adjective position after feature.

4.4 Summarizing

Finally, a summary with features and their respective polarities is generated. A summary may be shown through different models e.g., textual and graphic as in 4.9 and 4.10. In the example we have neutral evaluation for 'Localização' ['Location'], positive for 'Atendimento' ['Service'] and no evaluation for 'Quartos' ['Rooms'], 'Custo-benefício' ['Value'] and 'Limpeza' ['Cleanliness'].

Feature	Localização	Quartos	Atendimento	Custo-benefício	Limpeza
Polarity	0	-	$+1$	-	-

Figure 4.9: Example of textual summarizing.

Figure 4.10: Example of graphic summarizing.

5. Experimental Evaluation

This chapter presents the experimental evaluation of the proposed method regarding f-measure. The traditional f-measure (5.3) is the harmonic mean of precision (5.1) and recall (5.2). The precision reflects the ratio of accuracy of classified features and opinions to the number of all reviews, while recall reflects the ratio of completeness of all reviews classified correctly. In our case study we focused on a single domain: accommodation sector.

$$Precision = \frac{correct\ matches}{(correct\ matches + wrong\ matches)} \tag{5.1}$$

$$Recall = \frac{correct\ matches}{(correct\ matches + missed\ matches)} \tag{5.2}$$

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{5.3}$$

In order to validate our proposed method the following activities had to be performed:

- evaluate the capability of identification of explicit features based on concepts of domain ontology (HOntology);
- evaluate the capability of implicit feature identification for the five TripAdvisor features: ('Localização' ['Location'], 'Quarto' ['Rooms'], 'Atendimento' ['Service'], 'Custo-benefício'['Value'] and 'Limpeza' ['Cleanliness']);
- evaluate the results of our sentiment analysis method using different POS taggers (Portuguese TreeTagger, CitiusTagger and FreeLing);
- evaluate the results of our sentiment analysis method using different sentiment lexicons (OpLexicon, SentiLex, LIWC-PT and synsets with polarity from Onto.PT);
- evaluate the results of our sentiment analysis method using different linguistic rules (baseline and adjective position).

Hereafter we present the dataset and the ontology applied in our experiments, as well as the obtained results.

5.1 Dataset

We collected data from TripAdvisor through a Web crawler. TripAdvisor is the world's largest travel site, reaching more than 190 million reviews and opinions covering more than 4.4 million accommodations, restaurants and attractions. The data collection contains 194 Brazilian Portuguese reviews published from March, 2010 to May, 2014.

In the first moment, we identify features considering the following terms: 'Localização' ['Location'], 'Quarto' ['Rooms'], 'Atendimento' ['Service'], 'Custo-benefício'['Value'] and 'Limpeza' ['Cleanliness'], which are those presented in the TripAdvisor rating summary. For these five features we also identify their related implicit features. In the second moment, we identify features considering all concepts of domain ontology (HOntology).

The manual annotation of those reviews was conducted by two annotators, both native speakers of Portuguese, one linguist and one computer scientist using the tool developed in the context of this thesis. To perform the annotation task, annotators need to know the annotation guide and how to use the annotation tool (Appendix A).

The TripAdvisor dataset was made available to the annotators in text files. When a TripAdvisor review file is open in the annotation tool, explicit features are automatically highlighted in the text, and a box list all of them. Another box has fields to assign polarity to each automatically identified feature, where the default value is neutral. After reading and interpreting the review annotators can modify the automatically generated data, assigning polarities, inserting explicit features that where not automatically identified, or even removing automatically identified features. Yet, annotators can insert implicit features based on domain ontology concepts and assign polarities to them.

As a result of the annotation we had for the five features from TripAdvisor 269 explicit mentions and 71 implicit mentions, and for the 62 other features (occurring in the ontology and in the reviews) we had 987 explicit mentions, all annotated regarding polarity (Table 5.1).

Table 5.1: Summary of annotated dataset.

	No. of Features	Explicit Mentions		Kappa Explicit	Implicit Mentions		Kappa Implicit
TripAdvisor	5	269	Pos 111 Neg 93 Neu 12	0.67	71	Pos 53 Neg 13 Neu 0	0.79
HOntology	62	987	Pos 323 Neg 330 Neu 60	0.58	-		-

The agreement between annotators was measured with Kappa Statistics [LAN77]. The Kappa Statistics is a metric that evaluates concordance level classification tasks. The annotators agreement about sentiment analysis of the five features from TripAdvisor using Kappa was 0.67 for explicit feature and 0.79 for implicit feature, which is considered a substantial agreement (in a scale consisting of 'poor', 'fair', 'moderate', 'substantial', and 'almost perfect'). The annotators agreement about sentiment analysis of features from domain ontology using Kappa was 0.58 for explicit feature, which is considered a moderate agreement, using the same scale (Table 5.1). We believe that the annotation has an acceptable Kappa value.

It is also important to note that only in a few cases the annotators disagreed between negative and positive polarities, the majority of disagreements was about positive and neutral polarities, or negative and neutral polarities. Table 5.2 shows the number of disagreements cited above, and

the last column shows the percentage of annotators disagreement. As we can see, the biggest disagreement percentage was in polarity annotation of explicit HOntology features (27.76%).

The implicit features annotation disagreement was smaller than the explicit features annotation disagreement, 7.04% and 19.70%, respectively. Although the expected behaviour was that the major disagreements would occur on the implicit feature annotation.

Table 5.2: The disagreement between annotators.

	$+1/-1$	$+1/0$ or $-1/0$	% of disagreement
Explicit TripAdvisor Features	5	48	19.70%
Explicit HOntology Features	29	245	27.76%
Implicit TripAdvisor Features	3	2	7.04%

5.2 Ontology

In the feature identification step, we extracted the features using a multilingual ontology for accommodation sector, HOntology. This ontology reuses concepts of others vocabularies such as: QALL-ME, Schema.org and Dbpedia.org. According to Chaves and Trojahn [CHA10], the HOntology was developed in seven phases: (i) identify existing ontologies on related domains; (ii) select the main concepts and properties; (iii) organize concepts and properties hierarchically into categories; (iv) manually translate of the ontology to some languages (Portuguese, Spanish and French); (v) expand concepts and properties based on online reviews; (vi) translate the new concepts and properties; (vii) export the ontology in OWL format.

HOntology contains 282 concepts categorized into 16 top-level concepts ('Acomodação' ['Accommodation'], 'Aparência' ['Design'], 'Instalações' ['Facility'], 'Tipo de Hóspede' ['GuestType'], 'Hospitalidade' ['Hospitality'], 'Categoria de Hotéis' ['HotelCategory'], Rede de Hotéis ['HotelChain'], 'Localização' ['Location'], 'Refeição' ['Meal'], 'Pontos de Interesse' ['Points of Interest'], 'Preço' ['Price'], 'Avaliação da Acomodação' ['Rating'], 'Quarto' ['Room'], 'Atendimento' ['Service'], 'Funcionários' ['Staff'] and 'Horário' ['Timetable']). The concept hierarchy has a maximum depth of 5. Preliminary experiences in sentiment analysis using HOntology were performed for Portuguese in [CHA12a]. The authors analysed the features 'Localização' ['Location'], 'Quarto' ['Room'] and 'Funcionários'['Staff'].

We can highlight at least two limitations of the HOntology: (i) HOntology is composed by only a small set of data and object properties because the main requirements of the applications do not consider usage scenarios with properties; (ii) it is necessary to validate HOntology with accommodation managers, domain experts [CHA12b].

For this reason, we made a revision of HOntology concepts, hierarchy, and properties. Some hierarchy changes were made, concepts were removed, and new properties were inserted. It is worth mentioning that all concepts and properties were renamed for Brazilian Portuguese.

Among the hierarchy changes that we made in HOntology, we can list the change of the subclass

'Motorista' ['Driver'] of superclass 'Instalações' ['Facility'] to the superclass 'Funcionários' ['Staff']. The changes were made in view of the consensus of three evaluators.

The concepts removal was based on a single criterion: concepts that do not fit in the accommodation domain. From these concepts we can highlight: 'Salgados' ['Snacks'] (subclass of 'Instalações do Quarto' ['Room Facility']) and 'Animador' ['Animator'] (subclass of 'Funcionários' ['Staff']).

The nine properties inserted are related to the five TripAdivor's features, and they were created to support in the implicit feature identification. These properties are: 'acarretaCustoBenefício' ['entailsCostBenefit'], 'ofereceServiçoDeLimpeza' ['offersCleaningService'], 'ofereceAtendimento' ['offersService'], 'éPróximoDe' [isNear], 'pertenceAoQuarto' ['belongsToRoom'], 'temLocalização' ['hasLocation'], 'temPreçoDoQuarto' ['hasRoomPrice'], 'temServiçoDeLimpeza' ['hasCleaningService'], and 'temServiçoDeQuarto' ['hasRoomService'].

Besides that, in the first version of HOntology some concepts were written in European Portuguese and others in Brazilian Portuguese. For instances, 'Breakfast' in European Portuguese is 'Pequeno Almoço' and in Brazilian Portuguese is 'Café da Manhã'. However, both were labeled as $< rdfs : label\ xml : lang = \text{“}pt\text{”} >$. As our dataset is in Brazilian Portuguese we modify all concepts and properties of HOntology to Brazilian Portuguese. Table 5.3 presents some metrics about the HOntology revised. A deeper description of our modifications on HOntology can be found in Appendix B.

Table 5.3: Metrics HOntology revised.

Metric	Value
Number of Concepts	274
Number of Object Properties	16
Number of Data Properties	31

5.3 Results

In this section we present the results of the proposed approach over the accommodation reviews dataset.

Recalling our proposed method, initially, the reviews were preprocessed (POS and lemmatization), then the lemma of the words in the reviews was compared to the five features of TripAdvisor and to the concepts extracted from HOntology. At last, for those reviews that make mentions to features, we look for opinion words in the sentences.

Next, we compute a sentiment orientation score for each feature in the sentence. For that we use sentiment lexicons (adjectives and their polarities) and linguistic rules. The linguist rules used are: (1) Baseline, we identify the presence of adjectives three positions before feature and three positions after feature, apply the negation rules, and attribute as feature polarity the sum of polarities of all the adjectives found; (2) Adjectives Position, we identify the presence of adjectives immediately before feature, if this is not found, then we look for adjectives after feature (this operation is performed until

another feature is found or until the end of the sentence), apply the negation rules, and attribute as feature polarity the polarity of just one adjective found. The negative particles usually reverses the opinion expressed in a sentence.

Finally, tuples containing the features about the hotels and their polarity are shown.

We show the f-measure for polarity recognition using the proposed method in the dataset of 10 Porto Alegre hotel reviews randomly selected (Appendix C).

Polarity Evaluation Configurations

We present a comparison of f-measure scores, obtained for positive and negative polarities for three different POS taggers. The three compared configurations are describe below:

- configuration #1: this configuration uses Portuguese TreeTagger, a Union of PT Sentiment Lexicons, and Baseline;
- configuration #2: this configuration uses FreeLing, a Union of PT Sentiment Lexicons, and Baseline;
- configuration #3: this configuration uses CitiusTagger, a Union of PT Sentiment Lexicons, and Baseline.

Based on the best tagger resulting from this first evaluation, next we present a comparison of f-measure scores for the four different sentiment lexicons, all combined with Portuguese TreeTagger and Baseline. The next four configurations are describe below:

- configuration #4: Portuguese TreeTagger, a OpLexicon, and Baseline;
- configuration #5: Portuguese TreeTagger, a SentiLex, and Baseline;
- configuration #6: Portuguese TreeTagger, a LIWC-PT, and Baseline;
- configuration #7: Portuguese TreeTagger, a synsets with polarities from Onto.PT, and Baseline.

Finally we present a comparison of f-measure scores for the four different sentiment lexicons, now combined with Portuguese TreeTagger and Adjectives Position. The last four configurations are describe below:

- configuration #8: Portuguese TreeTagger, a OpLexicon, and Adjectives Position;
- configuration #9: Portuguese TreeTagger, a SentiLex, and Adjectives Position;
- configuration #10: Portuguese TreeTagger, a LIWC-PT, and Adjectives Position;
- configuration #11: Portuguese TreeTagger, a synsets with polarities from Onto.PT, and Adjectives Position.

Evaluation of TripAdvisor Features

We present a comparison of f-measure scores, obtained for positive and negative polarities of TripAdvisor's features, for three different POS taggers (Table 5.4). This table shows that using Portuguese TreeTagger we found the best positive f-measure for the following features: 'Localização' ['Location'] and 'Custo-benefício' ['Value']. And, when using FreeLing, we found the best negative f-measure for the following features: 'Quarto' ['Rooms'] and 'Atendimento' ['Service']. For 'Limpeza' ['Cleanliness'] and 'Custo-benefício' ['Value'] the negative f-measure has the same value for all POS taggers. Similarly, for the feature 'Atendimento' ['Service'] the positive f-measure has the same value for all tried POS taggers.

As shown in Table 5.4 we obtained 0.539 as the best average between positive and negative f-measure, this result was obtained for configuration #1.

Based on results presented in Table 5.4, we selected the Portuguese TreeTagger as the POS tagger to be applied in next analysis, presented in Tables 5.5 and 5.6.

Table 5.4: Polarity recognition of features (TripAdvisor) using different Portuguese POS taggers.

Explicit Features	No. of Features	#1		#2		#3	
		Pos	Neg	Pos	Neg	Pos	Neg
Quarto [Rooms]	105	0.68	0.29	**0.69**	**0.31**	0.67	0.24
Localização [Location]	55	**0.79**	0.17	0.71	0.33	0.71	0.33
Atendimento [Service]	45	0.90	0.36	0.90	**0.43**	0.90	0.31
Limpeza [Cleanliness]	7	0.80	0.40	0.80	0.40	0.67	0.40
Custo-benefício [Value]	4	**1.00**	0.00	0.67	0.00	0.80	0.00
Avg.		0.834	0.244	0.754	0.294	0.750	0.256
		0.539		0.524		0.503	

Table 5.5 shows that using synsets with polarities from Onto.PT and Baseline we found the best positive and negative f-measures for the most features.

As shown in Table 5.5 we obtained 0.626 as the best average between positive and negative f-measure, this result was obtained for configuration #7.

Table 5.5: Polarity recognition of features (TripAdvisor) using different Portuguese sentiment lexicons with baseline.

Explicit Features	No. of Features	#4		#5		#6		#7	
		Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
Quarto [Rooms]	105	0.69	0.29	0.57	0.44	0.46	0.18	**0.70**	**0.61**
Localização [Location]	55	**0.79**	0.17	0.77	0.31	0.50	0.00	0.77	0.31
Atendimento [Service]	45	0.90	0.36	0.90	0.36	0.76	0.20	0.90	**0.50**
Limpeza [Cleanliness]	7	0.67	0.40	0.80	0.67	0.80	0.40	0.80	0.67
Custo-benefício [Value]	4	1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Avg.		0.810	0.244	0.608	0.356	0.504	0.156	0.834	0.418
		0.527		0.482		0.330		**0.626**	

Table 5.6 shows that using synsets with polarities from Onto.PT and Adjectives Position we found the best negative f-measure for most features, except for the 'Localização' ['Location'] feature.

As shown in Table 5.6 we obtained 0.632 as the best average between positive and negative f-measure, this result was obtained for configuration #11.

Table 5.6: Polarity recognition of features (TripAdvisor) using different Portuguese sentiment lexicons with adjective position.

Explicit Features	No. of Features	#8		#9		#10		#11	
		Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
Quarto [Rooms]	105	**0.69**	0.24	0.46	0.40	0.41	0.12	0.66	**0.67**
Localização [Location]	55	**0.79**	0.29	0.75	**0.37**	0.49	0.17	0.78	0.35
Atendimento [Service]	45	0.84	0.50	0.84	0.36	0.71	0.36	0.84	**0.62**
Limpeza [Cleanliness]	7	0.57	0.40	0.67	0.67	**0.80**	0.40	0.67	0.67
Custo-benefício [Value]	4	1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Avg.		0.778	0.286	0.544	0.360	0.482	0.210	0.790	0.474
		0.532		0.456		0.346		**0.632**	

If we compare Table 5.5 and 5.6, we note that the best averages between positive and negative f-measure were 0.532, 0.346 and 0.632 using configurations #8, #10 and #11. In these configurations, we use the Adjectives Position and Brazilian or European Portuguese sentiment lexicon.

Table 5.7: Polarity recognition of features (HOntology concepts) using different Portuguese POS taggers.

Explicit Features	No. of Features	#1		#2		#3	
		Pos	Neg	Pos	Neg	Pos	Neg
Hotel [Hotel]	84	0.53	0.29	0.53	0.28	**0.59**	**0.33**
Café da manhã [Breakfast]	64	0.81	0.33	0.81	0.33	**0.83**	0.14
Preço [Price]	30	**0.88**	0.00	0.85	0.00	0.76	0.00
Recepção [Reception]	27	0.67	0.15	**0.76**	0.27	0.70	**0.43**
Cama [Bed]	26	0.81	0.00	0.81	0.00	0.79	0.00
Chuveiro[Shower]	21	0.33	**0.12**	0.29	0.11	**0.80**	0.11
Elevador [Lift]	18	0.00	0.00	**0.67**	0.00	0.00	0.00
Internet [Internet]	13	0.00	**1.00**	0.00	0.36	0.00	0.20
Serviço [Service]	13	0.00	0.00	0.50	0.33	**1.00**	0.33
Toalha [Towel]	9	0.50	0.57	**1.00**	0.57	0.67	**0.75**
Estabelecimento [Establishment]	7	0.00	**1.00**	0.00	**1.00**	0.00	**1.00**
Apartamento [Apartment]	6	1.00	**1.00**	0.89	**1.00**	1.00	**1.00**
Corredor [Hall]	6	0.00	0.29	0.00	0.29	0.00	**0.50**
Ar condicionado [AC]	4	0.67	**1.00**	0.67	**1.00**	0.00	**1.00**
Portaria [Lobby]	4	0.80	0.00	0.80	0.00	0.80	0.00
Serviço de Quarto [Service Room]	4	0.00	0.67	0.00	0.67	0.00	0.67
Shopping [Shopping]	4	**0.67**	0.00	0.40	0.00	0.40	0.00
Frigobar [Refrigerator]	3	0.00	0.00	1.00	0.00	1.00	0.00
Aeroporto [Airport]	2	**1.00**	0.00	**1.00**	0.00	**1.00**	0.00
Aparência [Design]	2	0.00	0.67	0.00	0.67	0.00	0.67
Ducha [Douche]	2	0.00	0.67	0.00	0.67	0.00	0.00
Colchão [Mattress]	1	0.00	**1.00**	0.00	**1.00**	0.00	**1.00**
Cortina [Curtain]	1	0.00	**1.00**	0.00	**1.00**	0.00	**1.00**
Gerência [Management]	1	**1.00**	0.00	**1.00**	0.00	**1.00**	0.00
Travesseiro [Pillow]	1	0.00	**1.00**	0.00	**1.00**	0.00	**1.00**

As we can see in Tables 5.4, 5.5 and 5.6, in general the f-measures for sentiment orientation recognition were better for positive than for negative cases. This may be explained because the reviews in the website were mostly marked as positive, against a low number of negative reviews.

Evaluation of HOntology Features

Tables 5.7, 5.8 and 5.9 show 29 HOntology concepts from the 62 identified in the manual annotation. They are the only features for which some polarity was identified by the system.

Table 5.8: Polarity recognition of features (HOntology concepts) using different Portuguese sentiment lexicons with baseline.

Explicit Features	No. of Features	#4		#5		#6		#7	
		Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
Hotel [Hotel]	84	0.51	0.31	0.43	0.41	0.51	0.25	0.59	0.55
Café da manhã [Breakfast]	64	0.69	0.33	0.75	0.42	0.60	0.32	0.79	0.36
Preço [Price]	30	0.85	0.00	0.85	0.00	0.48	0.00	0.88	0.00
Recepção [Reception]	27	0.59	0.15	0.59	0.15	0.43	0.15	**0.67**	**0.40**
Cama [Bed]	26	0.77	0.00	0.44	0.29	0.44	0.00	**0.88**	**0.44**
Chuveiro[Shower]	21	0.50	0.12	0.50	0.24	0.50	0.12	0.33	0.42
Elevador [Lift]	18	0.00	0.00	0.00	0.14	0.00	0.00	0.00	0.27
Internet [Internet]	13	0.00	1.00	0.00	1.00	0.00	0.00	0.00	1.00
Toalha [Towel]	9	0.67	0.57	0.67	0.89	0.00	0.33	0.67	0.67
Rua [Street]	8	0.00	0.50	0.00	0.50	0.00	0.00	0.00	0.50
Estabelecimento [Establishment]	7	0.00	1.00	0.00	1.00	0.00	0.00	0.00	1.00
Apartamento [Apartment]	6	1.00	1.00	0.89	1.00	0.33	1.00	1.00	1.00
Corredor [Hall]	6	0.00	0.29	0.00	0.67	0.00	0.00	0.00	0.80
Ar condicionado [AC]	4	0.67	1.00	0.00	1.00	0.00	0.00	0.67	1.00
Portaria [Lobby]	4	0.50	0.00	0.00	0.00	0.00	0.00	0.80	0.00
Serviço de Quarto [Service Room]	4	0.00	0.67	0.00	0.67	0.00	0.67	0.00	0.67
Shopping [Shopping]	4	0.00	0.00	0.00	0.00	0.67	0.00	1.00	0.00
Aeroporto [Airport]	2	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Aparência [Design]	2	0.00	0.67	0.00	0.67	0.00	0.67	0.00	1.00
Ducha [Douche]	2	0.00	0.67	0.00	1.00	0.00	0.67	0.00	1.00
Estacionamento [Parking]	2	0.00	0.33	0.00	0.33	0.00	0.00	0.00	0.33
Colchão [Mattress]	1	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
Cortina [Curtain]	1	0.00	1.00	0.00	1.00	0.00	0.00	0.00	1.00
Gerência [Management]	1	1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Travesseiro [Pillow]	1	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00

The most frequent HOntology explicit concepts were 'Hotel' ['Hotel'], 'Café da manhã' ['Breakfast'], 'Preço' ['Price'], 'Recepção' ['Reception'], 'Cama' ['Bed'], Chuveiro ['Shower'], Elevador ['Lift'] and 'Internet' ['Internet'].

Table 5.7 shows that using Portuguese TreeTagger we found the best positive f-measure for the following features: 'Preço' ['Price'] and 'Shopping' ['Shopping'].

For 'Aeroporto' ['Airport'] and 'Gerência' ['Management'] the positive f-measure has the same value for all POS taggers. Still, when using TreeTagger, we found the best negative f-measure

for the following features: 'Chuveiro' ['Shower'] and 'Internet' ['Internet']. For the feature 'Estabelecimento' ['Establishment'], 'Apartamento' ['Apartment'], 'Ar-condicionado' ['AC'], 'Colchão' ['Mattress'], 'Cortina'['Curtain'], and 'Travesseiro' ['Pillow'] the negative f-measure has the same value for all tried POS taggers.

Table 5.9: Polarity recognition of features (HOntology concepts) using different Portuguese sentiment lexicons with adjective position.

Explicit Features	No. of Features	#8		#9		#10		#11	
		Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
Hotel [Hotel]	84	0.64	0.29	0.52	0.41	0.51	0.29	0.68	0.57
Café da manhã [Breakfast]	64	0.71	0.35	0.76	0.35	0.57	0.24	0.81	0.30
Preço [Price]	30	0.77	0.00	0.82	0.29	0.37	0.00	0.82	0.25
Recepção [Reception]	27	0.47	0.37	0.40	0.37	0.29	0.15	0.50	0.35
Cama [Bed]	26	0.80	0.00	0.44	0.29	0.44	0.00	**0.83**	**0.50**
Chuveiro[Shower]	21	0.80	0.12	0.50	0.24	0.80	0.12	0.44	0.24
Elevador [Lift]	18	0.00	0.14	0.00	0.14	0.00	0.14	0.00	0.38
Internet [Internet]	13	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00
Toalha [Towel]	9	0.50	0.57	0.67	0.89	0.00	0.33	0.67	0.89
Rua [Street]	8	0.00	0.67	0.00	0.67	0.00	0.00	0.00	0.67
Estabelecimento [Establishment]	7	0.00	1.00	0.00	1.00	0.00	0.00	0.00	1.00
Apartamento [Apartment]	6	0.89	1.00	0.75	1.00	0.33	1.00	0.75	1.00
Corredor [Hall]	6	0.00	0.29	0.00	0.67	0.00	0.00	0.00	0.67
Ar condicionado [AC]	4	0.67	1.00	0.00	1.00	0.00	0.00	0.67	1.00
Portaria [Lobby]	4	0.40	0.00	0.00	0.00	0.00	0.00	0.40	0.00
Serviço de Quarto [Service Room]	4	0.00	0.67	0.00	0.67	0.00	0.67	0.00	0.67
Shopping [Shopping]	4	0.00	0.00	0.00	0.00	0.67	0.00	0.67	0.00
Cozinha [Kitchen]	3	1.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00
Aeroporto [Airport]	2	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Aparência [Design]	2	0.00	0.67	0.00	0.67	0.00	0.67	0.00	1.00
Ducha [Douche]	2	0.00	0.67	0.00	1.00	0.00	0.67	0.00	1.00
Estacionamento [Parking]	2	1.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00
Isolamento acústico [Soundproofing]	2	0.00	0.67	0.00	0.67	0.00	0.00	0.00	1.00
Colchão [Mattress]	1	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
Cortina [Curtain]	1	0.00	1.00	0.00	1.00	0.00	0.00	0.00	1.00
Gerência [Management]	1	1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Travesseiro [Pillow]	1	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00

Table 5.8, shows that using synsets with polarities from Onto.PT we found the best positive and negative f-measure for the following features: 'Recepção' ['Reception'] and 'Cama' ['Bed']. Also using synsets with polarities from Onto.PT (Table 5.9), we found the best positive and negative f-measure for 'Cama' ['Bed']. For 'Cama' ['Bed'], the best positive f-measure was obtained by the Baseline and the best negative f-measure was obtained using Adjectives Position.

There are other 28 features identified in the manual annotation but for which no polarity was assigned by the system. Most of these features for which no polarity was identified has low frequency in the dataset (Appendix D), they are:

'Funcionários', 'Instalações', 'Televisão', 'Lençóis', 'Tapete', 'Lavanderia', 'Torneira' ,'Calefação', 'Casal', 'Horário', 'Iluminação', 'Janta', 'Motel', 'Padrão', 'Telefone', 'Tomada', 'Pensão', 'Arena',

'Centro da Cidade', 'Cidade', 'Conforto', 'Escada', 'Gerente', 'Luxo', 'Móveis', 'País', 'Porteiro', and 'Quarto Duplo';

['Staff', 'Facility', 'TV', 'Bed Linen', 'Carpet', 'Laundry', 'Faucet', 'Heating System', 'Couple', 'Timetable', 'Lamp', 'Dinner', 'Motel', 'Standard', 'Telephone', 'Socket', 'Hostel', 'Arena', 'Downtown', 'City', 'Comfort', 'Stair', 'Manager', 'Luxury', 'Furniture', 'Country', 'Doorman', and 'Double Room'].

Implicit features

Our annotated data contains 66 mentions of implicit features related to the 5 features of the TripAdvisor (55 for 'Localização' ['Location'], 10 for 'Limpeza' ['Cleanliness'], 1 for 'Quarto' ['Rooms']). No implicit features were annotated for 'Custo-benefício' ['Value'], and 'Atendimento' ['Service'], see Table 5.10. For the system's identification of implicit features, we consider all the subclasses in the hierarchy of the target feature and other relations provided by HOntology. For instance, 'Suíte' ['Suite'] is a subclass of 'Quarto' ['Rooms'], if a review contains an occurrence of 'Suíte' ['Suite'] (or any other subclass of 'Quarto' ['Rooms']) we consider it as an implicit feature for 'Quarto' ['Rooms']. An example of the use of property is the mention of 'Endereço' ['Address'] as a implicit feature related to 'Localização' ['Location'], since they are linked by the property 'possuiEndereço' ['hasAddress'].

In the automatic implicit feature identification based on HOntology, we obtained 85 features (36 for 'Localização' ['Location'], 1 for 'Quarto' ['Rooms'], 33 for 'Custo-benefício' ['Value'], and 15 for 'Atendimento' ['Service']).

Table 5.10: Number of mentions: manual versus system.

Implicit Features	No. of Mentions (Manual)	No. of Mentions (System)	#11	
			Pos	Neg
Quarto [Rooms]	1	1	-	-
Localização [Location]	55	36	0.50	-
Atendimento [Service]	0	15	-	-
Limpeza [Cleanliness]	10	0	-	-
Custo-benefício [Value]	0	33	-	-
Sum	66	85		

However, we obtained polarity results just for the implicit features of 'Localização' ['Location'], for which we got 0.50 of positive f-measure.

Although the system has found implicit features related to 'Custo-benefício' ['Value'] and 'Atendimento' ['Service'], they were not considered in the manual annotation. For instance, the relation between 'Preço' ['Price'] and 'Custo-benefício' ['Value'] ('acarretaCustoBenefício' ['entailsCostBenefit']) occurs in the HOntology and therefore was identified by the system, but the annotators did not identify it. Perhaps in future versions of the annotation tool, the cases considered as implicit by the system could be shown to the annotators. These cases might have been missed because the difficulty of the task of finding any and all implicit features.

5.4 Error Analysis and Proposed Method Limitations

Here we discuss the main problems that may have happened in the proposed method regarding feature identification and polarity identification.

Feature identification errors occur when features in reviews are not recognized, i.e., a word representing a feature in a review is not detected as a feature (from TripAdvisor or HOntology).

Due to the different POS taggers methodologies each tagger can behaviour differently for the same review. Even the best found Tagger presented some problems. For instance, from the four times that the feature 'Custo-benefício' ['Value'] appears in our dataset, TreeTagger labels it three times as <unknown> (lemma). However the feature actually is a noun and its lemma should be 'Custo-benefício' ['Value']. CitiusTagger and FreeLing tag this feature correctly. It is extremely important to our sentiment analysis method, because we use the lemma to look for features in the preprocessed reviews.

Thus, depending on the employed POS tagger and the analysed feature the results of the proposed approach can be directly affected. For instance, TreeTagger labels the feature 'Móveis' ['Furnitures'] as Noun Common Masculine Plural (NCMP) with lemma 'móveis'. In its turn, FreeLing labels it as Verb Main Indicative Present Second Person Plural (VMIP2P0) with lemma 'mover'. At last, CitiusTagger labels the same feature as Noun Common Masculine Plural (NCMP000) with lemma as 'móvel'. As we can see, the only POS tagger that correctly tagged the feature was CitiusTagger, since the lemma must be always in singular and 'móveis' is a noun, not a verb. All HOntology concepts are in the noun masculine singular form, then for this example we only identify the feature if we use the CitiusTagger.

The following review from TripAdvisor do not follow the written language standards required for correctly processing the text: "... ruim.**Café da manhã** ok**Atendimento** sem reclamações ..." ["... bad.**Breakfast** ok**Service** no complains ..."]. The absence of a white space between the words 'ok' and 'Atendimento' results that POS taggers can not identify the 'Atendimento' ['Service'] feature. However, when a punctuation mark is between two words, even without white spaces, the POS taggers FreeLing and CitiusTagger are able to identify the punctuation mark and the feature, 'Café da manhã' ['Breakfast'] in this example.

Polarity identification errors occur when the polarity obtained by our proposed method differs from the polarity manually annotated in our dataset. This may occur when: a qualifier adjective is not tagged as such, a word is incorrectly tagged as a qualifier adjective, or the sentiment lexicon does not contain a polarity for this qualifier adjective. Any of these cases lead to an incorrect polarity count.

The review "Em relação à Limpeza, achei REGULAR ..." ["In relation to the **cleanliness**, I thought *REGULAR* ..."] shows a case where we detect the 'Limpeza' ['Cleanliness'] feature, but as the TreeTagger incorrectly tagged 'REGULAR' ['REGULAR'] as a noun instead of an adjective.

Other polarity identification error occurs in the following review: "... **internet** que todo dia tem que pedir para *resetar* ..." ["... **internet** that every day we have to ask for *reset* ..."]. In this

review the proposed approach is able to identify the 'Internet' ['Internet'] feature, however 'resetar' ['reset'] is incorrectly tagged by TreeTagger as an adjective, but it is an verb. Thus, we look in the sentiment lexicon and can not found this adjective. It is important to notice that, even if the tagger correctly tags 'resetar' ['reset'] our method will not get the correct polarity because it currently only looks in the adjective list from sentiment lexicons. Among the limitations are language pattern that the proposed method does not deal yet are adversative and additive conjunctions. In the following reviews we can see the use of these types of conjunctions: (1) "A **localização** e **custo-benefício** são *ótimos* para quem tem negócios" ["The **location** and **value** are *great* for those who have business"] (2) ' '*Próximo* do **aeroporto** mas com **instalações** *precárias* ..." ["*Close* to the **airport** but with *poor* **facilities** ..."]. In review (1) the additive conjunction 'e' ['and'] between 'Localização' ['Localization'] and 'Custo-benefício' ['Value'] should assign polarity to both features, but as the proposed method does not deals with this language pattern only 'Custo-benefício' ['Value'] is assigned with positive polarity. In review (2) the adversative conjunction 'mas' ['but'] between 'Aeroporto' ['Airport'] and 'Instalações' ['Facility'] could help in the polarity assigning if one of the qualifier adjectives, 'próximo' ['close'] or 'precárias' ['poor'], is not in the sentiment lexicon, since the adversative conjunction implies in an inversion of polarities.

5.5 Summarization

Our data set comprises reviews about ten hotels. Each hotel has its own set of reviews. In TripAdvisor, users write reviews about the hotels, and after that they evaluate five predefined features which they rate with grades from 1 to 5. In this way, the TripAdvisor's reviewers rate all 5 features, even though in some cases they are not all referenced in the review's text. The hotel final rating is based on the ratings of its users.

Our system, on the other hand, produces a summary which is entirely based on the review's texts. So the comparison we present is not an evaluation against a reference. Indeed, it may indicate that the textual opinion of the users rate the hotel in a different way.

Table 5.11 shows the TripAdvisor's rating and our system outputs for the same features, considering the whole set of hotels. Note that our system is able to detect implicit and explicit references to the features for all hotels in reviews, however not all reviews make reference to all features. So while the TripAdvisor's users rated all features for the ten hotels, our system, based on the textual reviews, may not rate all of them.

In the last column we can see that for all hotel's reviews only eight had any reference to 'Atendimento' ['Service']. Similarly, for four of the hotels the feature 'Limpeza' ['Cleanliness'] was not detected. And, at last, the feature 'Custo-benefício' ['Value'] was referenced in the reviews of only one hotel.

TripAdvisor's reviewers use the rating tool as a complement of the textual review, i.e., the textual review not always expresses some opinion about all TripAdvisor's features. Our system could be used to request further explanations when a rating is not verified in the opinion expressed in the

Table 5.11: Reviewers rating versus system outputs.

	Reviewers rating				System outputs			
	Pos	Neg	Neu	Sum	Pos	Neg	Neu	Sum
Quarto [Rooms]	1	9	0	10	8	2	0	10
Localização [Location]	7	2	1	10	7	1	2	10
Atendimento [Service]	3	3	4	10	7	0	1	8
Limpeza [Cleanliness]	1	3	6	10	2	2	2	6
Custo-benefício [Value]	4	1	5	10	1	0	0	1

text. Some questions, such as: 'Why you didn't like the room?' or 'What is the advantage of this location?', could help to clarify the reasons for some of the ratings. Still, our system could be used to propose other features for the summary, such as the most frequent features that are found. In the case of our experiments, the system could suggest 'Café da manhã' ['Breakfast'] and 'Internet' ['Internet'].

The original data, summarized in Table 5.11, can be found in E.

6. Final Remarks

This chapter presents the achievements of this work and the perspectives for future works based on the current status.

6.1 Conclusions

In this work we propose a method for feature-level sentiment analysis for Brazilian Portuguese reviews using POS taggers, ontologies, sentiment lexicons, and linguistic rules. This is the first approach considering a set of Portuguese linguistic rules and using a Portuguese ontology on identification of explicit and implicit features.

We consider as the main contributions of this thesis: the definition, implementation and evaluation of the proposed method, which was applied to accommodation reviews, written in Brazilian Portuguese.

Among the specific contributions of this thesis we can list:

- the development of a tool to capture hotel reviews in the TripAdvisor Web page [TRI15] (Web crawler);
- the development of a tool to annotate reviews at the feature-level considering a domain ontology;
- the manual evaluation and translation for Brazilian Portuguese of domain ontology (HOntology);
- the creation of guidelines for corpus annotation in feature-level (explicit and implicit);
- the evaluation of currently available Portuguese resources (e.g.: POS tagger and sentiment lexicon) for sentiment-analysis;
- the sharing of the annotated corpus.

During this work development we evaluated the impact of three different POS taggers on the feature-level sentiment analysis proposed method. TreeTagger, FreeLing, and CitiusTagger were integrated in the implementation of the proposed method. According to our experiments (considering our accommodation reviews dataset and our proposed method), in general the best results were obtained when using TreeTagger.

Besides that, we also evaluated four different Portuguese sentiment lexicons: OpLexicon, SentiLex, LIWC-PT, and synsets with polarities from Onto.PT. From the experiments realized the best results, for our dataset and proposed method, were obtained using synsets with polarities from Onto.PT.

We also tried different linguist rules, such as: baseline and adjective position. In our experiments, the baseline produced the best f-measure results for positive polarity identification (70%) for 'Quarto' ['Room'], and the adjective position produced the best f-measure results for negative polarity identification (67%) for 'Atendimento' ['Service'] and (62%) for 'Quarto' ['Rooms'].

The quality of the proposed method results can be better noticed when we consider the results of other works. For English, Peñalver-Martínez *et al.* [PEN14] obtained 84.8% of accuracy for positive polarity identification and 87.1% of accuracy for negative polarity identification in the movies domain. For Portuguese, Baracho *et al.* [BAR12] obtained 63% of positive f-measure and 37% of negative f-measure for a car model (Palio) and 38% of positive f-measure and 63% of negative f-measure for another model (Corsa).

Thus, we believe that although the proposed method has many opportunities for improvements, the obtained results were satisfactory. Next we list related published papers developed during this thesis. Then, in section 6.4 we present the following actions to improve the results of obtained with the proposed method.

6.2 Publications

The work being developed in the last four years has achieved important results, described in the following scientific publications:

- One journal paper:

  [FRE13b] **Comparing Portuguese Opinion Lexicons in Feature-Based Sentiment Analysis**.

- Five conference papers (including full papers and abstracts):

  [CHA12a] **PIRPO: An Algorithm to Deal with Polarity in Portuguese Online Reviews from the Accommodation Sector**;

  [CHA12b] **HOntology: A Multilingual Ontology for the Accommodation Sector in the Tourism Industry**;

  [FRE14] **Pathways for irony detection in tweets**;

  [FRE13a] **Ontology-based Feature Level Opinion Mining for Portuguese Reviews**;

  [VAN13] **Some Clues on Irony Detection in Tweets**.

6.3 Resources

We also produced research resources that are available to the scientific community:

- Four systems available at `https://sites.google.com/site/larissaaf/pesquisa`;

  Crawler Tool, **Irony Detection Tool**, **OLARE Tool**, **FAMA**.

- One feature level sentiment annotated corpus.

6.4 Future Work

As future work there are several improvements to be done in the proposed method. One of them is to explore new ways to discover implicit features, besides the ontology relations approach applied in this work. According to Liu [LIU12], there are many types of expressions that refer to implicit features, such as: an occurrence of adjectives, adverbs and verbs, for instance, the adjectives 'barato' ['cheap'] and 'caro' ['expensive'] describe 'preço' ['price']; an occurrence of synonym, for instance, 'vizinhança' ['neighbourhood'] synonymous with 'localização' ['location']; an occurrence of term derived, for instance, 'limpeza' ['cleanliness'] derived from 'limpo' ['clean']; an occurrence of very complicated expressions, for instance, "coube no bolso" ["fits your budget"] indicates 'preço' ['price']. For instance, we explore types of expressions cited by Liu [LIU12] using thesaurus and dictionary of multi-word expressions.

As we discussed in Section 5.4, the linguistic rules can be extended to use other types of words than adjectives, such as: adverbs, verbs, nouns, *etc.* Besides that, we also have to adapt the proposed method to deal with additive and adversatives conjunctions.

Other interesting future works are to implement ways to dealing with sarcasm/irony in reviews. Furthermore, many challenges in sentiment analysis involve issues in NLP, such as: entity recognition, co-reference resolution, word sense disambiguation and, *etc.* [LIU12].

In domain reviews, customers discuss specific features related to their experience. Another problem is that different customers prioritize different features [CAT13]. For instance, a couple on a honeymoon probably does not give much importance to the quality of the Internet connection at the hotel, whereas this feature can be very important for a manager on a business trip [TIT08]. Thus we explore others elements of the quintuple presented by Liu [LIU12] as opinion holder and time.

References

[AGO15] Agoda. "Agoda.com: Smarter Hotel Booking". Available from: http://www.agoda.com, Accessed in: January 2015.

[ALU13] Aluísio, S. and Chechia, R. and Chishman, R. "Brazilian Portuguese LIWC 2007 Dictionary". Available from: http://www.nilc.icmc.usp.br/portlex/index.php/pt/projetos/liwc, October 2013.

[ASG14] Asghar, M. Z. and Khan, A. and Ahmad, S. and Kundi, F. M. "A Review of Feature Extraction in Sentiment Analysis". *Journal of Basic and Applied Scientific Research*, vol. 4, 2014, pp. 181–186.

[BAC10] Baccianella, A. and Esuli, S. and Sebastiani, F. "Sentiwordnet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining". In: 7th International Conference on Language Resources and Evaluation, 2010, pp. 2200–2204.

[BAR12] Baracho, R. and Silva, G. and Ferreira, L. "Sentiment Analysis in Social Networks: A Study on Vehicle". In: 5th Research Seminar of Ontologies in Brazil, 2012, pp. 132–143.

[BBC14] BBC. "Brasil deve fechar 2014 como quarto país com mais acesso à internet, diz consultoria". Available from: http://www.bbc.co.uk/portuguese/noticias/2014/11/141124_brasil_internet_pai, November 2014.

[BHU09] Bhuiyan, T. and Xu, Y. and Josang, A. "State-of-the-Art Review on Opinion Mining from Online Customer's Feedback". In: 9th Asia-Pacific Complex Systems Conference, 2009, pp. 385–390.

[BIC00] Bick, E. "The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework" Ph.D. thesis, Aarhus University, 2000, 505p.

[CAR05] Carenini, G. and Ng, R. T. and Zwart, E. "Extracting Knowledge from Evaluative Text". In: 3rd International Conference on Knowledge Capture, 2005, pp. 11–18.

[CAR15] Carrosnaweb. "Carros na Web - Classificados, catálogo, avaliações, opinião do dono, notícias, lançamentos". Available from: http://www.carrosnaweb.com.br/, Accessed in: January 2015.

[CAT13] Cataldi, M. and Ballatore, A. and Tiddi, I. and Aufaure, M. A. "Good Location, Terrible Food: Detecting Feature Sentiment in User-Generated Reviews". *Journal Social Network Analysis and Mining*, vol. 3, 2013, pp. 1149–1163.

[CHA10] Chaves, M. S. and Trojahn, C. "Towards a Multi-Lingual Ontology for Ontology-Driven Content Mining in Social Web Sites". In: 1st International Workshop on Cross-Cultural and Cross-Lingual Aspects of the Semantic Web, 2010, pp. 1–10.

[CHA12a] Chaves, M. S. and Freitas, L. A. and Souza, M. and Vieira, R. "PIRPO: An Algorithm to Deal with Polarity in Portuguese Online Reviews from the Accommodation Sector". In: 17th International Conference on Applications of Natural Language Processing to Information Systems, 2012, pp. 296–301.

[CHA12b] Chaves, M. S. and Freitas, L. A. and Vieira, R. "HOntology: A Multilingual Ontology for the Accommodation Sector in the Tourism Industry". In: 4th International Conference on Knowledge Engineering and Ontology Development, 2012, pp. 149–154.

[CHE14] Chernyshevich, M. "IHS R&D Belarus: Cross-domain extraction of product features using CRF". In: 8th International Workshop on Semantic Evaluation (SemEval 2014), 2014, pp. 309–313.

[CIT15] CitiusTagger. "ProlnatTagger". Available from: http://gramatica.usc.es/pln/tools/ Citius-Tools.html, Accessed in: January 2015.

[DEP15] DepPattern. "DepPattern". Available from: http://gramatica.usc.es/pln/tools/ deppattern.html, Accessed in: January 2015.

[DIA03] Dias, B. C. and Moraes, H. R. "A Construção de um Thesaurus Eletrônico para o Português do Brasil". *Alfa*, vol. 47, 2003, pp. 101–115.

[EBI15] Ebit. "Compras online com mais segurança ao consumidor - E-bit". Available from: http://www.ebit.com.br/, Accessed in: January 2015.

[FEL07] Feldman, R. and Sange, J. "The Text Mining Handbook, Advanced Approach in Analyzing Unstructured Data". New York: Cambridge University Press, 2007, 424p.

[FEL98] Fellbaum, C. "WordNet: And Electronic Lexical Database". Massachusetts: MIT Press, 1998, 422p.

[FRE13a] Freitas, L. and Vieira, R. "Ontology-based Feature Level Opinion Mining for Portuguese Reviews". In: 22nd International World Wide Web Conference - Doctoral Consortium, 2013, pp. 367–370.

[FRE13b] Freitas, L. and Vieira, R. "Comparing Portuguese Opinion Lexicons in Feature-Based Sentiment Analysis". In: International Journal of Computational Linguistics and Applications, 2013, pp. 147–158.

[FRE14] Freitas, L. and Vanin, A. and Hogetop, D. and Bochernitsan, M. and Vieira, R. "Pathways for irony detection in tweets". In: 29th Symposium On Applied Computing, 2014, pp. 628–633.

[FRE15] FreeLing. "FreeLing". Available from: http://gramatica.usc.es/pln/tools/freeling.html, Accessed in: January 2015.

[GAM05] Gamon, M. and Aue, A. and Corston-Oliver, S. and Ringger, E. "Pulse: Mining Customer Opinions from Free Text". In: 6th International Symposium on Intelligent Data Analysis, 2005, pp. 121–132.

[GRU95] Gruber, T. R. "Toward Principles for the Design of Ontologies used for Knowledge Sharing". *International Journal of Human Computer Studies*, vol. 43, 1995, pp. 907–928.

[GUA98] Guarino, N. "Formal Ontology in Information Systems". In: 1st International Conference on Formal Ontology in Information Systems, 1998, pp. 3–15.

[HAR10] Haruechaiyasak, C. and Kongthon, A. and Palingoon, P. and Sangkeettrakarn, C. "Constructing Thai Opinion Mining Resource: A Case Study on Hotel Reviews". In: 8th Workshop on Asian Language Resources, 2010, pp. 64–71.

[HAT97] Hatzivassiloglou, V. and McKeown, K. R. "Predicting the Semantic Orientation of Adjectives". In: 8th Conference on European Chapter of the Association for Computational Linguistics, 1997, pp. 174–182.

[HAT00] Hatzivassiloglou, V. and Wiebe, J. "Effects of Adjective Orientation and Gradability on Sentence Subjectivity". In: 18th International Conference on Computational Linguistics, 2000, pp. 299–305.

[HIC05] Hickey, R. "Level of Language". Available from: https://www.uni-due.de/SHE/REV_Levels_Chart.htm, December 2005.

[HON15] HOntology. "OntoLP - Portal de Ontologia". Available from: http://ontolp.inf.pucrs.br/Recursos/downloads-Hontology.php, Accessed in: January 2015.

[HOR10] Hornby, A. H. "Oxford Advanced Learner's Dictionary". United Kingdom: Oxford University Press, 2010, 1842 p.

[HU04] Hu, M. and Liu, B. "Mining Opinion Features in Customer Reviews". In: 19th National Conference on Artificial Intelligence, 2004, pp. 755–760.

[JOS14] Joshi, N. S. and Itkat, S. A. "A Survey on Feature Level Sentiment Analysis". *International Journal of Computer Science and Information Technologies*, vol. 5, 2014, pp. 5422–5425.

[KAS11] Kasper, W. and Vela, M. "Sentiment Analysis for Hotel Reviews". In: Computational Linguistics-Applications Conference, 2011, pp. 45–52.

[KEN06] Kennedy, A. and Inkpen, D. "Sentiment Classification of Movie Reviews Using Contextual Valence Shifters". *Computational Intelligence*, vol. 22, 2006, pp. 110–125.

[KIR14] Kiritchenko, S. and Zhu, X. and Cherry, C. and Mohammad, S. "NRC-Canada-2014: Detecting Aspects and Sentiment in Customer Reviews". In: 8th International Workshop on Semantic Evaluation (SemEval 2014), 2014, pp. 437–442.

[LAN77] Landis, J. and Koch, G. "The Measurement of Observer Agreement for Categorical Data". *Biometrics*, vol. 33, 1977, pp. 159–174.

[LAS01] Lassila, O e McGuinness, D. L. "The Role of Frame-Based Representation on the Semantic Web". *Linköping Electronic Articles in Computer and Information Science*, vol. 6, 2001, pp. 78–87.

[LIU10] Liu, B. "Sentiment Analysis and Subjectivity". In: Indurkhya, N. and Damerau, F. (Org.). Handbook of Natural Language Processing, Second Edition. Flórida: Taylor and Francis Group, 2010. p. 627–666.

[LIU11] Liu, B. "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)". New York: Springer-Verlag, 2011, 624p.

[LIU12] Liu, B. "Sentiment Analysis and Opinion Mining (Synthesis Lectures on Human Language Technologies)". California: Morgan & Claypool Publishers, 2012, 180p.

[MIA10] Miao, Q. and Li, Q. and Zeng, D. "Fine-Grained Opinion Mining by Integrating Multiple Review Sources". *Journal of the American Society for Information Science and Technology*, vol. 61, 2010, pp. 2288–2299.

[NEV11] Neves, M. H. "Gramática de usos do português". São Paulo: Editora Unesp, 2011, 1005p.

[NIE11] Nielsen, F. A. "A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs". In: 1st Workshop on Making Sense of Microposts, 2011, pp. 93–98.

[NLP15] NLProcessor. "Infogistics' NLProcessor". Available from: http://www.infogistics.com/textanalysis.html, Accessed in: January 2015.

[OLI14] Oliveira, H. G. and Santos, A. P. and Gomes, P. "Assigning Polarity Automatically to the Synsets of Wordnet-Like Resource". In: 3rd Symposium on Languages, Applications and Technologies, 2014, pp. 169–184.

[OLS10] Olston, C. and Najork, M. "Web Crawling". *Foundations and Trends in Information Retrieval*, vol. 4, 2010, pp. 175–246.

[ORT10] Ortiz, A. M. and Castollo, F. P. and García, R. H. "Analyzing Hotel Reviews with Sentitext: A Domain-Independent, Sentiment Analysis System". *Procesamiento del Lenguaje Natural*, vol. 45, 2010, pp. 31–39.

[PAN08] Pang, B. and Lee, L. "Opinion Mining and Sentiment Analysis". *Foundations and Trends in Information Retrieval*, vol. 2, 2008, pp. 1–135.

[PEN14] Peñalver-Martínez, I. and Valencia-García, R. and García-Sánchez, F. and Rodríguez-García, M. and Moreno, V. and Fraga, A. and Sánchez-Cervantes, J. "Feature-Based Opinion Mining Through Ontologies". *Expert Systems with Applications*, vol. 41, 2014, pp. 5995–6008.

[PEN01] Pennerbaker, J. W. and Francis, M. E. and Booth, R. J. "Linguistic Inquiry and Word Count". Mahwah, NJ: Erlbaum Publishers, 2001, pp. 1–13.

[POL06] Polanyi, L. and Zaenen, A. "Contextual Valence Shifters". In: Computing Attitude and Affect in Text: Theory and Applications, 2006, pp. 1–10.

[PON14] Pontiki, M. and Galanis, D. and Pavlopoulos, J. and Papageorgiou, H. and Androutsopoulos, I. and Manandhar, S. "SemEval-2014 Task 4: Aspect Based Sentiment Analysis". In: 8th International Workshop on Semantic Evaluation (SemEval 2014), 2014, pp. 27–35.

[POP05] Popescu, A. M. and Etzioni, O. "Extracting Product Features and Opinions from Reviews". In: Conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005, pp. 9–28.

[PRO15] Protégé. "Protégé". Available from: http://protege.stanford.edu/, Accessed in: January 2015.

[QUI85] Quirk, R. and Greenbaum, S. and Leech, G. and Svartvik, J. "A Comprehensive Grammar of the English Language". Harlow, Essex, England: Longman, 1985, 1779p.

[RIB12] Ribeiro, S. S. and Junior, Z. and Meira, W. and Pappa, G. L. "Positive or Negative? Using Blogs to Assess Vehicles Features". In: Encontro Nacional de Inteligência Artificial, 2012, pp. 1–12.

[SCH05] Schwenter S. A. "The Pragmatics of Negation in Brazilian Portuguese". *Lingua*, vol. 115, 2005, pp. 1427–1455.

[SEM10] Semiocast. "Half of Messages on Twitter are not in English Japanese is the Second Most Used Language". Available from: https://semiocast.com/downloads/Semiocast_Half_of_messages_on_Twitter_are_not_in_English_20100224.pdf, February 2010.

[SEN15] SentiWordNet. "SentiWordNet". Available from: http://sentiwordnet.isti.cnr.it/, Accessed in: January 2015.

[SHA14] Shah, V. and Rekh, P. "A Survey: Importance of Negation in Sentiment Analysis". *International Journal of Emerging Technology and Advanced Engineering*, vol. 4, 2014, pp. 70–73.

[SIL12] Silva, M. J. and Carvalho, P. and Sarmento, L. "Building a Sentiment Lexicon for Social Judgement Mining". In: 10th International Conference Computational Processing of the Portuguese Language, 2012, pp. 218–228.

[SIQ10] Siqueira, H. and Barros, F. "A Feature Extraction Process for Sentiment Analysis of Opinions on Services". In: 3rd International Workshop on Web and Text Intelligence, 2010, pp. 1–10.

[SMI04] Smith, M. K. and Welty, C. and McGuinness, D. L. "OWL Web Ontology Language Guide". Available from: http://www.w3.org/TR/owl-guide/, February 2004.

[SOU11] Souza, M. and Vieira, R. and Chishman, R. and Alves, I. M. "Construction of a Portuguese Opinion Lexicon from Multiple Resources". In: 8th Brazilian Symposium in Information and Human Language Technology, 2011, pp. 59–66.

[SOW84] Sowa, J. F. "Conceptual Structures - Information Processing in Mind and Machine". Boston: Addison-Wesley, 1984, 481p.

[STR04] Strapparava, C. and Valitutti, A. "Wordnet-Affect: An Affective Extension of Wordnet". In: 4th International Conference on Language Resources and Evaluation, 2004, pp. 1083–1086.

[THE12] Thelwall, M. and Buckley, K. and Paltoglou G. "Sentiment Strength Detection for the Social Web". *Journal of the American Society for Information Science and Technology*, vol. 59, 2012, pp. 163–173.

[TIT08] Titov, I. and McDonald, R. "Modeling Online Reviews with Multi-grain Topic Models". In: 17th International World Wide Web Conference, 2008, pp. 111–120.

[TOH14] Toh, Z. and Wang, W. "DLIREC: Aspect Term Extraction and Term Polarity Classification System". In: 8th International Workshop on Semantic Evaluation (SemEval 2014), 2014, pp. 235–240.

[TRE15] TreeTagger. "TreeTagger". Available from: http://gramatica.usc.es/~gamallo/tagger.htm, Accessed in: January 2015.

[TRI15] TripAdvisor. "Dicas, avaliações e comentários de hotéis e pousadas, restaurantes e atrações turísticas - TripAdvisor". Available from: http://www.tripadvisor.com.br/, Accessed in: January 2015.

[TUR02] Turney, P. "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews". In: 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 417–424.

[VAN13] Vanin, A. and Freitas, A. F., Vieira, R. and Bochernitsan, M. N. "Some Clues on Irony Detection in Tweets". In: 22nd International Conference on World Wide Web, 2013, pp. 635–636.

[ZHA09] Zhao, L. and Li, C. "Ontology Based Opinion Mining for Movie Reviews". In: 3rd International Conference Knowledge, Science, Engineering and Management, 2009, pp. 204–214.

[ZHO08] Zhou, L. and Chaovalit, P. "Ontology-Supported Polarity Mining". *Journal of the American Society for Information Science and Technology*, vol. 59, 2008, pp. 98–110.

[WAG14] Wagner, J. and Arora, P. and Cortes, S. and Barman, U. and Bogdanova, D. and Foster, J. and Tounsi, L. "DCU: Aspect-based Polarity Classification for SemEval Task 4". In: 8th International Workshop on Semantic Evaluation (SemEval 2014), 2014, pp. 223–229.

[WES14] Westerski, A. "Sentiment Analysis: Introduction and the State of the Art overview" Technical Report, Universidad Politecnica de Madrid, 2014, pp. 211–218.

[WIE10] Wiegand, M. and Balahur, A. and Roth, B. and Klakow, D. and Montoyo, A. "A Survey on the Role of Negation in Sentiment Analysis". In: Workshop on Negation and Speculation in Natural Language Processing, 2010, pp. 60–68.

[WIL05] Wilson, T. and Wiebe, J. and Hoffmann, P. "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis". In: Conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005, pp. 347–354.

[WOR15] WordNet. "WordNet". Available from: http://wordnet.princeton.edu/, Accessed in: January 2015.

A. Appendix A

A.1 Guia de Anotação

A finalidade dessa anotação é detectar aspectos e suas polaridades em comentários de hotéis. Como tarefa do anotador podemos citar: identificar a polaridade dos aspectos explícitos, os aspectos implícitos e a polaridade dos aspectos implícitos.

A.1.1 Esquema de Anotação

- **Alvo da Opinião**: Um texto opinativo pode conter vários alvos da opinião. Os alvos correspondem as entidades e seus aspectos. Uma entidade é um produto, serviço, tópico, pessoa, organização ou evento. O termo aspecto ou feature (subfeatures, subsubfeatures) é usado para denotar partes (subpartes, subsubpartes) e atributos de uma entidade [LIU12].

  Por exemplo, hotel é uma entidade; quarto é uma parte; Ibis é um atributo.

- **Polaridade**: A polaridade da opinião é representada através de um valor numérico, as opiniões positivas são classificadas como 1 e as opiniões negativas são classificadas como -1. As opiniões neutras são classificadas como 0 (nem positivas e nem negativas).

  Por exemplo, em "A localização é ótima.", o aspecto localização tem polaridade 1; em "A localização é bem estranha e há barulho o tempo todo.", o aspecto localização tem polaridade -1; em "Em relação a localização, fica no centro.", o aspecto localização tem polaridade 0.

- **Visibilidade (Tipo de Aspecto)**: Quanto à visibilidade os aspectos podem ser explícitos ou implícitos. Consideramos como aspecto explícito a ocorrência direta de um termo (simples ou múltiplo) da ontologia. Neste trabalho utilizamos a HOntology revisada. Consideramos como aspecto implícito uma referência indireta a um dos aspectos explícitos (localização, quarto, atendimento, custo benefício e limpeza).

  Por exemplo, em "A localização é ótima.", o aspecto localização é explícito; em "Achei as imediações do hotel bem desertas e fui aconselhada a não me aventurar sozinha pelas redondezas.", o aspecto localização é implícito.

A.1.2 Ferramenta de Anotação

A ferramenta utilizada para realizar a anotação é a FAMA (Ferramenta de Anotação Manual Automatizada). Para acessá-la é necessário efetuar o login, ou seja, preencher os campos username e password e clicar no botão entrar (Figure A.1).

O próximo passo é abrir um arquivo no formato .txt com codificação utf-8. Para isso, clicar no botão selecionar arquivo... .

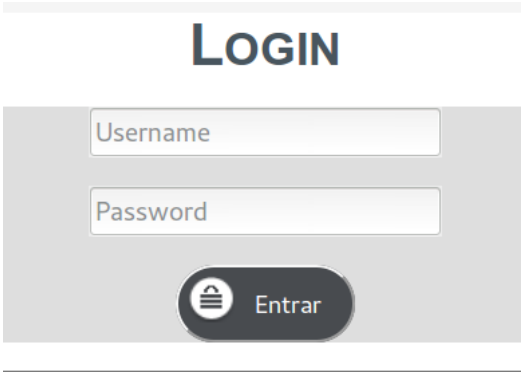

Figure A.1: Página de login.

Após o arquivo ser aberto, as features explícitas identificadas no comentário são listadas (Figura A.2). Cabe salientar que as features, as subfeatures e as subsubfeatures são provindas dos conceitos e da hieraquia da ontologia do setor hoteleiro (HOntology revisada).

Como mencionado anteriormente, uma das tarefas do anotador é identificar a polaridade dos aspectos explícitos e implícitos, ou seja, dependendo do contexto, a quarta coluna da Figura A.2 e da Figura A.3 devem ser classificadas como 1 (positiva), -1 (negativa) ou 0 (neutra). Por padrão, a quarta coluna é classificada como 0 (neutra). Ainda, outra tarefa do anotador é identificar aspectos implícitos, para isso o anotador deve clicar no botão adicionar linha e preencher os campos feature, subfeature, subsubfeature e termo (Figura A.3). Os campos subfeature e subsubfeature são de preenchimento opcional. O campo termo deve remeter a algum aspecto explícito (localização, quarto, atendimento, custo-benefício e limpeza). Por exemplo, a ocorrência de expressões mais complexas, como "ser salgada" remete ao aspecto custo-benefício.

Por fim, a anotação deve ser salva, para isso o anotador deve clicar no botão salvar. Após, uma mensagem irá aparecer na tela informando que a operação foi realizada com sucesso.

Figure A.2: Página de anotação manual com aspectos explícitos extraídos.

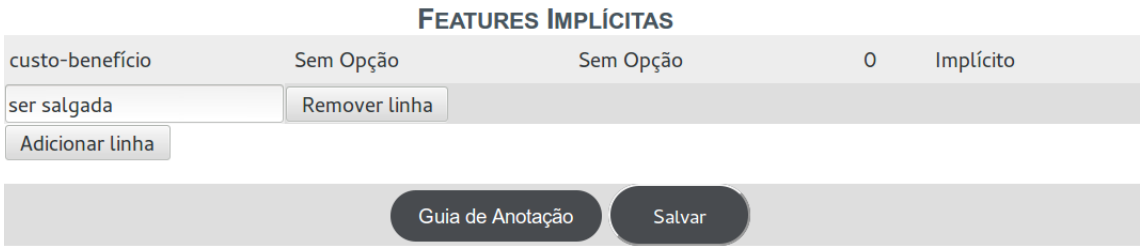

Figure A.3: Página de anotação manual com aspectos implícitos extraídos.

B. Appendix B

B.1 HOntology Revised

Table B.2: Properties HOntology revised.

Class	Property
Location	Establishment **hasLocation** Address
	Points of Interest **hasLocation** Address
	Establishment **isNear** Interest Points
Rooms	Facility/Rooms Facility **belongsToRoom** Rooms
	Facility/Rooms Facility **hasCleaningService** Service/Cleanliness
	Rooms **hasRoomService** Serviço/Service Rooms
	Rooms **hasRoomPrice** Price/Rooms Price
Service	Staff **offersService** Service
	Service **offersService** Service
Value	Price **entailsCostBenefit** Value
Cleanliness	Rooms **hasCleaningService** Service/Cleanliness
	Facility/Rooms Facility **hasCleaningService** Service/Cleanliness
	Facility/Outside Facility **hasCleaningService** Service/Cleanliness
	Facility/Inside Facility **hasCleaningService** Service/Cleanliness
	Facility/Bathroom Facility **hasCleaningService** Service/Cleanliness
	Staff/Cleanliness Staff **offersCleaningService** Service/Cleanliness

Table B.1: Concepts HOntology revised.

	Class	**SubClass**
Accommodation	Establishment (I **equivalence between** Hostel and Inn)	Hotel (I Boutique Hotel, Farm Hotel, Boat Hotel, History Hotel)
Rating	Rating (RM)	
Hospitality	Hospitality (RM)	
Address=Address	Address=Address (RM)	
Timetable	Timetable (I Parking Timetable)	
Facility	Facility (RM Driver, Kitchen)	Bathroom Facility (I Towel Heater) Rooms Facility (I Furniture, Bed Linen; RM Snack) Inside Facility (RM Playground, Meeting, Beauty Saloon)
Points of Interest	Location (I Bar and Restaurant, Night Club, Park; RM Carré-FR)	
Staff	Staff (I Concierge, Valet, Laundry Staff, Baby-sitting, Driver; RM **equivalence between** Check Out and Reception, **equivalence between** Reception and Animator	
Rooms	Accommodation (I **equivalence between** Apartment Room and Hotel Room, **equivalence between** Double Room and Twin; RM all concepts in Apartment Room, Luxury Room, Standard Room)	Hostel Room (I 12 Bed Male Dorm, 10 Bed Male Dorm, 8 Bed Male Dorm, 6 Bed Male Dorm, 4 Bed Male Dorm)
Service	Service (I **hierarchy** In-room Breakfast, Guest-chosen Daily Newspaper Delivery, Management; RM **equivalence between** Check Out and Receptionist, **equivalence between** Reception and Receptionist)	
Guest Type	Group of Friends (I Young Group, Mature Group)	

RM - Remove, I - Insert

C. Appendix C

C.1 Evaluation of Proposed Method

Table C.1: TripAdvisor features, configuration #1 and #2.

Features	#1						Features	#2					
	Pos			Neg				Pos			Neg		
	P	R	F	P	R	F		P	R	F	P	R	F
Rooms	0.71	0.65	0.68	1.00	0.17	0.29	Rooms	0.71	0.67	0.69	1.00	0.18	0.31
Location	0.88	0.72	0.79	1.00	0.09	0.17	Location	0.91	0.58	0.71	1.00	0.20	0.33
Service	1.00	0.82	0.90	1.00	0.22	0.36	Service	0.97	0.85	0.90	1.00	0.27	0.43
Cleanliness	1.00	0.67	0.80	1.00	0.25	0.40	Cleanliness	1.00	0.67	0.80	1.00	0.25	0.40
Value	1.00	1.00	1.00	0.00	0.00	0.00	Value	0.67	0.67	0.67	0.00	0.00	0.00

Table C.2: TripAdvisor features, configuration #3 and #4.

Features	#3						Features	#4					
	Pos			Neg				Pos			Neg		
	P	R	F	P	R	F		P	R	F	P	R	F
Rooms	0.70	0.63	0.67	1.00	0.14	0.24	Rooms	0.83	0.59	0.69	1.00	0.17	0.29
Location	0.91	0.58	0.71	1.00	0.20	0.33	Location	0.88	0.72	0.79	1.00	0.09	0.17
Service	0.97	0.85	0.90	1.00	0.18	0.31	Service	1.00	0.82	0.90	1.00	0.22	0.36
Cleanliness	1.00	0.50	0.67	1.00	0.25	0.40	Cleanliness	0.67	0.67	0.67	1.00	0.25	0.40
Value	1.00	0.67	0.80	0.00	0.00	0.00	Value	1.00	1.00	1.00	0.00	0.00	0.00

Table C.3: TripAdvisor features, configuration #5 and #6.

Features	#5						Features	#6					
	Pos			Neg				Pos			Neg		
	P	R	F	P	R	F		P	R	F	P	R	F
Rooms	0.93	0.41	0.57	1.00	0.28	0.44	Rooms	0.59	0.38	0.46	1.00	0.10	0.18
Location	0.88	0.69	0.77	1.00	0.18	0.31	Location	0.92	0.34	0.50	0.00	0.00	0.00
Service	1.00	0.82	0.90	1.00	0.22	0.36	Service	1.00	0.61	0.76	1.00	0.11	0.20
Cleanliness	1.00	0.67	0.80	1.00	0.50	0.67	Cleanliness	1.00	0.67	0.80	1.00	0.25	0.40
Value	0.00	0.00	0.00	0.00	0.00	0.00	Value	0.00	0.00	0.00	0.00	0.00	0.00

Table C.4: TripAdvisor features, configuration #7 and #8.

Features	#7						Features	#8					
	Pos			Neg				Pos			Neg		
	P	R	F	P	R	F		P	R	F	P	R	F
Rooms	0.81	0.62	0.70	0.88	0.47	0.61	Rooms	0.83	0.59	0.69	1.00	0.13	0.24
Location	0.82	0.72	0.77	1.00	0.18	0.31	Location	0.88	0.72	0.79	0.67	0.18	0.29
Service	1.00	0.82	0.90	1.00	0.33	0.50	Service	0.95	0.75	0.84	1.00	0.33	0.50
Cleanliness	1.00	0.67	0.80	1.00	0.50	0.67	Cleanliness	0.50	0.67	0.57	1.00	0.25	0.40
Value	1.00	1.00	1.00	0.00	0.00	0.00	Value	1.00	1.00	1.00	0.00	0.00	0.00

Table C.5: TripAdvisor features, configuration #9 and #10.

Features	#9						Features	#10					
	Pos			Neg				Pos			Neg		
	P	R	F	P	R	F		P	R	F	P	R	F
Rooms	0.79	0.32	0.46	1.00	0.25	0.40	Rooms	0.55	0.32	0.41	1.00	0.07	0.12
Location	0.88	0.66	0.75	0.60	0.27	0.37	Location	0.85	0.34	0.49	1.00	0.09	0.17
Service	0.95	0.75	0.84	1.00	0.22	0.36	Service	0.94	0.57	0.71	1.00	0.22	0.36
Cleanliness	0.67	0.67	0.67	1.00	0.50	0.67	Cleanliness	1.00	0.67	0.80	1.00	0.25	0.40
Value	0.00	0.00	0.00	0.00	0.00	0.00	Value	0.00	0.00	0.00	0.00	0.00	0.00

Table C.6: TripAdvisor features, configuration #11.

Features	#11					
	Pos			Neg		
	P	R	F	P	R	F
Rooms	0.74	0.59	0.66	0.85	0.55	0.67
Location	0.85	0.72	0.78	0.50	0.27	0.35
Service	0.95	0.75	0.84	1.00	0.44	0.62
Cleanliness	0.67	0.67	0.67	1.00	0.50	0.67
Value	1.00	1.00	1.00	0.00	0.00	0.00

Table C.7: HOntology concepts, configuration #1 and #2.

Features	#1					
	Pos			Neg		
	P	R	F	P	R	F
Hotel	0.55	0.52	0.53	0.89	0.17	0.29
Breakfast	0.94	0.71	0.81	0.75	0.21	0.33
Price	0.95	0.82	0.88	0.00	0.00	0.00
Reception	0.86	0.55	0.67	0.50	0.09	0.15
Bed	0.79	0.85	0.81	0.00	0.00	0.00
Shower	0.33	0.33	0.33	1.00	0.07	0.12
Lift	0.00	0.00	0.00	0.00	0.00	0.00
Internet	0.00	0.00	0.00	1.00	1.00	1.00
Service	0.00	0.00	0.00	0.00	0.00	0.00
Towel	0.50	0.50	0.50	1.00	0.40	0.57
Establishment	0.00	0.00	0.00	1.00	1.00	1.00
Apartment	1.00	1.00	1.00	1.00	1.00	1.00
Hall	0.00	0.00	0.00	1.00	0.17	0.29
AC	1.00	0.50	0.67	1.00	1.00	1.00
Lobby	1.00	0.67	0.80	0.00	0.00	0.00
Service Rooms	0.00	0.00	0.00	1.00	0.50	0.67
Shopping	1.00	0.50	0.67	0.00	0.00	0.00
Refrigerator	0.00	0.00	0.00	0.00	0.00	0.00
Airport	1.00	1.00	1.00	0.00	0.00	0.00
Design	0.00	0.00	0.00	1.00	0.50	0.67
Douche	0.00	0.00	0.00	1.00	0.50	0.67
Mattress	0.00	0.00	0.00	1.00	1.00	1.00
Curtain	0.00	0.00	0.00	1.00	1.00	1.00
Management	1.00	1.00	1.00	0.00	0.00	0.00
Pillow	0.00	0.00	0.00	1.00	1.00	1.00

Features	#2					
	Pos			Neg		
	P	R	F	P	R	F
Hotel	0.58	0.48	0.53	0.89	0.16	0.28
Breakfast	0.94	0.70	0.81	1.00	0.20	0.33
Price	0.94	0.77	0.85	0.00	0.00	0.00
Reception	0.89	0.67	0.76	0.50	0.18	0.27
Bed	0.79	0.85	0.81	0.00	0.00	0.00
Shower	0.25	0.33	0.29	1.00	0.06	0.11
Lift	1.00	0.50	0.67	0.00	0.00	0.00
Internet	0.00	0.00	0.00	1.00	0.22	0.36
Service	1.00	0.33	0.50	1.00	0.20	0.33
Towel	1.00	1.00	1.00	1.00	0.40	0.57
Establishment	0.00	0.00	0.00	1.00	1.00	1.00
Apartment	1.00	0.80	0.89	1.00	1.00	1.00
Hall	0.00	0.00	0.00	1.00	0.17	0.29
AC	1.00	0.50	0.67	1.00	1.00	1.00
Lobby	1.00	0.67	0.80	0.00	0.00	0.00
Service Rooms	0.00	0.00	0.00	1.00	0.50	0.67
Shopping	1.00	0.25	0.40	0.00	0.00	0.00
Refrigerator	1.00	1.00	1.00	0.00	0.00	0.00
Airport	1.00	1.00	1.00	0.00	0.00	0.00
Design	0.00	0.00	0.00	1.00	0.50	0.67
Douche	0.00	0.00	0.00	1.00	0.50	0.67
Mattress	0.00	0.00	0.00	1.00	1.00	1.00
Curtain	0.00	0.00	0.00	1.00	1.00	1.00
Management	1.00	1.00	1.00	0.00	0.00	0.00
Pillow	0.00	0.00	0.00	1.00	1.00	1.00

Table C.8: HOntology concepts, configuration #3 and #4.

Features	#3					
	Pos			Neg		
	P	R	F	P	R	F
Hotel	0.67	0.53	0.59	0.91	0.20	0.33
Breakfast	0.91	0.76	0.83	1.00	0.08	0.14
Price	0.93	0.64	0.76	0.00	0.00	0.00
Reception	0.88	0.58	0.70	1.00	0.27	0.43
Bed	0.73	0.85	0.79	0.00	0.00	0.00
Shower	1.00	0.67	0.80	1.00	0.06	0.11
Lift	0.00	0.00	0.00	0.00	0.00	0.00
Internet	0.00	0.00	0.00	1.00	0.11	0.20
Service	1.00	1.00	1.00	1.00	0.20	0.33
Towel	1.00	0.50	0.67	1.00	0.60	0.75
Establishment	0.00	0.00	0.00	1.00	1.00	1.00
Apartment	1.00	1.00	1.00	1.00	1.00	1.00
Hall	0.00	0.00	0.00	1.00	0.33	0.50
AC	0.00	0.00	0.00	1.00	1.00	1.00
Lobby	1.00	0.67	0.80	0.00	0.00	0.00
Service Rooms	0.00	0.00	0.00	1.00	0.50	0.67
Shopping	1.00	0.25	0.40	0.00	0.00	0.00
Refrigerator	1.00	1.00	1.00	0.00	0.00	0.00
Airport	1.00	1.00	1.00	0.00	0.00	0.00
Design	0.00	0.00	0.00	1.00	0.50	0.67
Douche	0.00	0.00	0.00	0.00	0.00	0.00
Mattress	0.00	0.00	0.00	1.00	1.00	1.00
Curtain	0.00	0.00	0.00	1.00	1.00	1.00
Management	1.00	1.00	1.00	0.00	0.00	0.00
Pillow	0.00	0.00	0.00	1.00	1.00	1.00

Features	#4					
	Pos			Neg		
	P	R	F	P	R	F
Hotel	0.58	0.45	0.51	0.82	0.19	0.31
Breakfast	0.92	0.55	0.69	0.75	0.21	0.33
Price	0.94	0.77	0.85	0.00	0.00	0.00
Reception	0.83	0.45	0.59	0.50	0.09	0.15
Bed	0.77	0.77	0.77	0.00	0.00	0.00
Shower	1.00	0.33	0.50	1.00	0.07	0.12
Lift	0.00	0.00	0.00	0.00	0.00	0.00
Internet	0.00	0.00	0.00	1.00	1.00	1.00
Towel	1.00	0.50	0.67	1.00	0.40	0.57
Street	0.00	0.00	0.00	1.00	0.33	0.50
Establishment	0.00	0.00	0.00	1.00	1.00	1.00
Apartment	1.00	1.00	1.00	1.00	1.00	1.00
Hall	0.00	0.00	0.00	1.00	0.17	0.29
AC	1.00	0.50	0.67	1.00	1.00	1.00
Lobby	1.00	0.33	0.50	0.00	0.00	0.00
Service Rooms	0.00	0.00	0.00	1.00	0.50	0.67
Shopping	0.00	0.00	0.00	0.00	0.00	0.00
Airport	0.00	0.00	0.00	0.00	0.00	0.00
Design	0.00	0.00	0.00	1.00	0.50	0.67
Douche	0.00	0.00	0.00	1.00	0.50	0.67
Parking	0.00	0.00	0.00	1.00	0.20	0.33
Mattress	0.00	0.00	0.00	1.00	1.00	1.00
Curtain	0.00	0.00	0.00	1.00	1.00	1.00
Management	1.00	1.00	1.00	0.00	0.00	0.00
Pillow	0.00	0.00	0.00	1.00	1.00	1.00

Table C.9: HOntology concepts, configuration #5 and #6.

Features	#5					
	Pos			Neg		
	P	R	F	P	R	F
Hotel	0.55	0.35	0.43	0.81	0.28	0.41
Breakfast	0.96	0.62	0.75	0.80	0.29	0.42
Price	0.94	0.77	0.85	0.00	0.00	0.00
Reception	0.83	0.45	0.59	0.50	0.09	0.15
Bed	0.80	0.31	0.44	1.00	0.17	0.29
Shower	1.00	0.33	0.50	1.00	0.13	0.24
Lift	0.00	0.00	0.00	1.00	0.08	0.14
Internet	0.00	0.00	0.00	1.00	1.00	1.00
Towel	1.00	0.50	0.67	1.00	0.80	0.89
Street	0.00	0.00	0.00	1.00	0.33	0.50
Establishment	0.00	0.00	0.00	1.00	1.00	1.00
Apartment	1.00	0.80	0.89	1.00	1.00	1.00
Hall	0.00	0.00	0.00	1.00	0.50	0.67
AC	0.00	0.00	0.00	1.00	1.00	1.00
Lobby	0.00	0.00	0.00	0.00	0.00	0.00
Service Rooms	0.00	0.00	0.00	1.00	0.50	0.67
Shopping	0.00	0.00	0.00	0.00	0.00	0.00
Airport	0.00	0.00	0.00	0.00	0.00	0.00
Design	0.00	0.00	0.00	1.00	0.50	0.67
Douche	0.00	0.00	0.00	1.00	1.00	1.00
Parking	0.00	0.00	0.00	1.00	0.20	0.33
Mattress	0.00	0.00	0.00	1.00	1.00	1.00
Curtain	0.00	0.00	0.00	1.00	1.00	1.00
Management	0.00	0.00	0.00	0.00	0.00	0.00
Pillow	0.00	0.00	0.00	1.00	1.00	1.00

Features	#6					
	Pos			Neg		
	P	R	F	P	R	F
Hotel	0.65	0.42	0.51	0.78	0.15	0.25
Breakfast	0.90	0.45	0.60	0.60	0.21	0.32
Price	1.00	0.32	0.48	0.00	0.00	0.00
Reception	1.00	0.27	0.43	0.50	0.09	0.15
Bed	0.80	0.31	0.44	0.00	0.00	0.00
Shower	1.00	0.33	0.50	1.00	0.07	0.12
Lift	0.00	0.00	0.00	0.00	0.00	0.00
Internet	0.00	0.00	0.00	0.00	0.00	0.00
Towel	0.00	0.00	0.00	1.00	0.20	0.33
Street	0.00	0.00	0.00	0.00	0.00	0.00
Establishment	0.00	0.00	0.00	0.00	0.00	0.00
Apartment	1.00	0.20	0.33	1.00	1.00	1.00
Hall	0.00	0.00	0.00	0.00	0.00	0.00
AC	0.00	0.00	0.00	0.00	0.00	0.00
Lobby	0.00	0.00	0.00	0.00	0.00	0.00
Service Rooms	0.00	0.00	0.00	1.00	0.50	0.67
Shopping	1.00	0.50	0.67	0.00	0.00	0.00
Airport	0.00	0.00	0.00	0.00	0.00	0.00
Design	0.00	0.00	0.00	1.00	0.50	0.67
Douche	0.00	0.00	0.00	1.00	0.50	0.67
Parking	0.00	0.00	0.00	0.00	0.00	0.00
Mattress	0.00	0.00	0.00	1.00	1.00	1.00
Curtain	0.00	0.00	0.00	0.00	0.00	0.00
Management	0.00	0.00	0.00	0.00	0.00	0.00
Pillow	0.00	0.00	0.00	1.00	1.00	1.00

Table C.10: HOntology concepts, configuration #7 and #8.

Features	#7					
	Pos			Neg		
	P	R	F	P	R	F
Hotel	0.60	0.58	0.59	0.86	0.40	0.55
Breakfast	0.94	0.69	0.79	0.50	0.29	0.36
Price	0.95	0.82	0.88	0.00	0.00	0.00
Reception	0.86	0.55	0.67	0.75	0.27	0.40
Bed	0.92	0.85	0.88	0.67	0.33	0.44
Shower	0.33	0.33	0.33	1.00	0.27	0.42
Lift	0.00	0.00	0.00	1.00	0.15	0.27
Internet	0.00	0.00	0.00	1.00	1.00	1.00
Towel	1.00	0.50	0.67	0.75	0.60	0.67
Street	0.00	0.00	0.00	1.00	0.33	0.50
Establishment	0.00	0.00	0.00	1.00	1.00	1.00
Apartment	1.00	1.00	1.00	1.00	1.00	1.00
Hall	0.00	0.00	0.00	1.00	0.67	0.80
AC	1.00	0.50	0.67	1.00	1.00	1.00
Lobby	1.00	0.67	0.80	0.00	0.00	0.00
Service Rooms	0.00	0.00	0.00	1.00	0.50	0.67
Shopping	1.00	1.00	1.00	0.00	0.00	0.00
Airport	1.00	1.00	1.00	0.00	0.00	0.00
Design	0.00	0.00	0.00	1.00	1.00	1.00
Douche	0.00	0.00	0.00	1.00	1.00	1.00
Parking	0.00	0.00	0.00	1.00	0.20	0.33
Mattress	0.00	0.00	0.00	1.00	1.00	1.00
Curtain	0.00	0.00	0.00	1.00	1.00	1.00
Management	1.00	1.00	1.00	0.00	0.00	0.00
Pillow	0.00	0.00	0.00	1.00	1.00	1.00

Features	#8					
	Pos			Neg		
	P	R	F	P	R	F
Hotel	0.72	0.58	0.64	1.00	0.17	0.29
Breakfast	1.00	0.55	0.71	1.00	0.21	0.35
Price	0.88	0.68	0.77	0.00	0.00	0.00
Reception	0.67	0.36	0.47	0.60	0.27	0.37
Bed	0.83	0.77	0.80	0.00	0.00	0.00
Shower	1.00	0.67	0.80	1.00	0.07	0.12
Lift	0.00	0.00	0.00	1.00	0.08	0.14
Internet	1.00	1.00	1.00	1.00	1.00	1.00
Towel	0.50	0.50	0.50	1.00	0.40	0.57
Street	0.00	0.00	0.00	0.00	0.00	0.00
Establishment	0.00	0.00	0.00	1.00	1.00	1.00
Apartment	1.00	0.80	0.89	1.00	1.00	1.00
Hall	0.00	0.00	0.00	1.00	0.17	0.29
AC	1.00	0.50	0.67	1.00	1.00	1.00
Lobby	0.50	0.33	0.40	0.00	0.00	0.00
Service Rooms	0.00	0.00	0.00	1.00	0.50	0.67
Shopping	0.00	0.00	0.00	0.00	0.00	0.00
Airport	0.00	0.00	0.00	0.00	0.00	0.00
Design	0.00	0.00	0.00	1.00	0.50	0.67
Douche	0.00	0.00	0.00	1.00	0.50	0.67
Parking	0.00	0.00	0.00	0.00	0.00	0.00
Mattress	0.00	0.00	0.00	1.00	1.00	1.00
Curtain	0.00	0.00	0.00	1.00	1.00	1.00
Management	1.00	1.00	1.00	0.00	0.00	0.00
Pillow	0.00	0.00	0.00	1.00	1.00	1.00

Table C.11: HOntology concepts, configuration #9 and #10.

Features	#9					
	Pos			Neg		
	P	R	F	P	R	F
Hotel	0.68	0.42	0.52	1.00	0.26	0.41
Breakfast	1.00	0.62	0.76	1.00	0.21	0.35
Price	0.94	0.73	0.82	1.00	0.17	0.29
Reception	0.75	0.27	0.40	0.60	0.27	0.37
Bed	0.80	0.31	0.44	1.00	0.17	0.29
Shower	1.00	0.33	0.50	1.00	0.13	0.24
Lift	0.00	0.00	0.00	1.00	0.08	0.14
Internet	1.00	1.00	1.00	1.00	1.00	1.00
Service	0.00	0.00	0.00	0.00	0.00	0.00
Towel	1.00	0.50	0.67	1.00	0.80	0.89
Street	0.00	0.00	0.00	0.00	0.00	0.00
Establishment	0.00	0.00	0.00	1.00	1.00	1.00
Apartment	1.00	0.60	0.75	1.00	1.00	1.00
Hall	0.00	0.00	0.00	1.00	0.50	0.67
AC	0.00	0.00	0.00	1.00	1.00	1.00
Lobby	0.00	0.00	0.00	0.00	0.00	0.00
Service Rooms	0.00	0.00	0.00	1.00	0.50	0.67
Shopping	0.00	0.00	0.00	0.00	0.00	0.00
Kitchen	0.00	0.00	0.00	0.00	0.00	0.00
Airport	0.00	0.00	0.00	0.00	0.00	0.00
Design	0.00	0.00	0.00	1.00	0.50	0.67
Douche	0.00	0.00	0.00	1.00	1.00	1.00
Parking	0.00	0.00	0.00	0.00	0.00	0.00
Soundproofing	0.00	0.00	0.00	0.00	0.00	0.00
Mattress	0.00	0.00	0.00	1.00	1.00	1.00
Curtain	0.00	0.00	0.00	1.00	1.00	1.00
Management	0.00	0.00	0.00	0.00	0.00	0.00
Pillow	0.00	0.00	0.00	1.00	1.00	1.00

Features	#10					
	Pos			Neg		
	P	R	F	P	R	F
Hotel	0.65	0.42	0.51	1.00	0.17	0.29
Breakfast	0.94	0.40	0.57	0.67	0.14	0.24
Price	1.00	0.23	0.37	0.00	0.00	0.00
Reception	0.67	0.18	0.29	0.50	0.09	0.15
Bed	0.80	0.31	0.44	0.00	0.00	0.00
Shower	1.00	0.67	0.80	1.00	0.07	0.12
Lift	0.00	0.00	0.00	1.00	0.08	0.14
Internet	1.00	1.00	1.00	0.00	0.00	0.00
Service	0.00	0.00	0.00	0.00	0.00	0.00
Towel	0.00	0.00	0.00	1.00	0.20	0.33
Street	0.00	0.00	0.00	0.00	0.00	0.00
Establishment	0.00	0.00	0.00	0.00	0.00	0.00
Apartment	1.00	0.20	0.33	1.00	1.00	1.00
Hall	0.00	0.00	0.00	0.00	0.00	0.00
AC	0.00	0.00	0.00	0.00	0.00	0.00
Lobby	0.00	0.00	0.00	0.00	0.00	0.00
Service Rooms	0.00	0.00	0.00	1.00	0.50	0.67
Shopping	1.00	0.50	0.67	0.00	0.00	0.00
Kitchen	0.00	0.00	0.00	0.00	0.00	0.00
Airport	0.00	0.00	0.00	0.00	0.00	0.00
Design	0.00	0.00	0.00	1.00	0.50	0.67
Douche	0.00	0.00	0.00	1.00	0.50	0.67
Parking	0.00	0.00	0.00	0.00	0.00	0.00
Soundproofing	0.00	0.00	0.00	0.00	0.00	0.00
Mattress	0.00	0.00	0.00	1.00	1.00	1.00
Curtain	0.00	0.00	0.00	0.00	0.00	0.00
Management	0.00	0.00	0.00	0.00	0.00	0.00
Pillow	0.00	0.00	0.00	1.00	1.00	1.00

Table C.12: HOntology concepts, configuration #11.

Features	#11					
	Pos			Neg		
	P	R	F	P	R	F
Hotel	0.62	0.74	0.68	0.95	0.40	0.57
Breakfast	0.97	0.69	0.81	0.50	0.21	0.30
Price	0.94	0.73	0.82	0.50	0.17	0.25
Reception	0.80	0.36	0.50	0.50	0.27	0.35
Bed	0.91	0.77	0.83	1.00	0.33	0.50
Shower	0.33	0.67	0.44	1.00	0.13	0.24
Lift	0.00	0.00	0.00	1.00	0.23	0.38
Internet	1.00	1.00	1.00	1.00	1.00	1.00
Service	0.00	0.00	0.00	0.00	0.00	0.00
Towel	1.00	0.50	0.67	1.00	0.80	0.89
Street	0.00	0.00	0.00	0.00	0.00	0.00
Establishment	0.00	0.00	0.00	1.00	1.00	1.00
Apartment	1.00	0.60	0.75	1.00	1.00	1.00
Hall	0.00	0.00	0.00	1.00	0.50	0.67
AC	1.00	0.50	0.67	1.00	1.00	1.00
Lobby	0.50	0.33	0.40	0.00	0.00	0.00
Service Rooms	0.00	0.00	0.00	1.00	0.50	0.67
Shopping	1.00	0.50	0.67	0.00	0.00	0.00
Kitchen	0.00	0.00	0.00	0.00	0.00	0.00
Airport	1.00	1.00	1.00	0.00	0.00	0.00
Design	0.00	0.00	0.00	1.00	1.00	1.00
Douche	0.00	0.00	0.00	1.00	1.00	1.00
Parking	0.00	0.00	0.00	0.00	0.00	0.00
Soundproofing	0.00	0.00	0.00	0.00	0.00	0.00
Mattress	0.00	0.00	0.00	1.00	1.00	1.00
Curtain	0.00	0.00	0.00	1.00	1.00	1.00
Management	1.00	1.00	1.00	0.00	0.00	0.00
Pillow	0.00	0.00	0.00	1.00	1.00	1.00

D. Appendix D

D.1 Frequency of HOntology Features in the Accommodation Dataset.

Table D.1: Frequency of HOntology features.

Features	No. de Features	Features	No. de Features
Rooms	105	Faucet	3
Hotel	84	Airport	2
Breakfast	64	Design	2
Location	55	Heating System	2
Service	45	Couple	2
Price	30	Douche	2
Reception	27	Parking	2
Bed	26	Timetable	2
Shower	21	Lamp	2
Lift	18	Soundproofing	2
Staff	16	Dinner	2
Internet	13	Motel	2
Service	13	Standard	2
Facility	11	Telephone	2
TV	11	Socket	2
Bed Linen	9	Hostel	1
Towel	9	Arena	1
Cleanliness	7	Downtown	1
Street	8	City	1
Establishment	7	Mattress	1
Carpet	7	Comfort	1
Apartment	6	Curtain	1
Hall	6	Stair	1
AC	4	Manager	1
Value	4	Management	1
Laundry	4	Luxury	1
Lobby	4	Furnitures	1
Service Room	4	Country	1
Shopping	4	Doorman	1
Kitchen	3	Double Room	1
Refrigerator	3	Pillow	1

E. Appendix E

E.1 Comparing TripAdvisor and Proposed Method Summary of 10 Porto Alegre Hotels

This Appendix presents a comparison between the summaries generated by our experimental results and TripAdvisor's users rating for all the ten hotels analized.

Note that, features 'Limpeza' ['Cleanliness'], 'Atendimento' ['Service'] and 'Custo-benefício' ['Value'] do not appear in the textual reviews of some Hotels (1,3,4,6 - 1,7- 0,1,3,4,5,6,7,8,9, respectively).

This type of summary, besides serving as a general evaluation of a specific hotel, may be useful for specific features recommendations. For instance, Hotel 2 (Figure E.3) could be recommended when users look for better 'Localização' ['Location'], 'Atendimento' ['Service'], and 'Custo-benefício' ['Value'] ratings and Hotel 8 (Figure E.9) could be recommended when users look for better 'Localização ['Location'], 'Atendimento' ['Service'], and 'Limpeza' ['Cleanliness'].

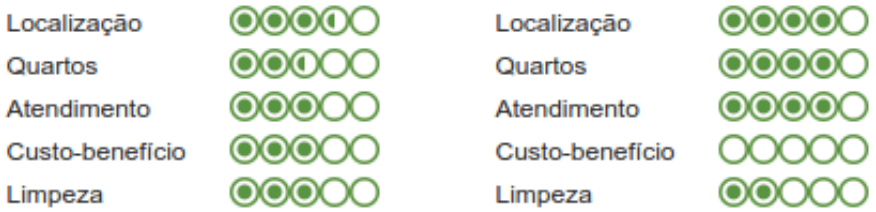

Figure E.1: TripAdvisor and Proposed Method Summary of Hotel 0.

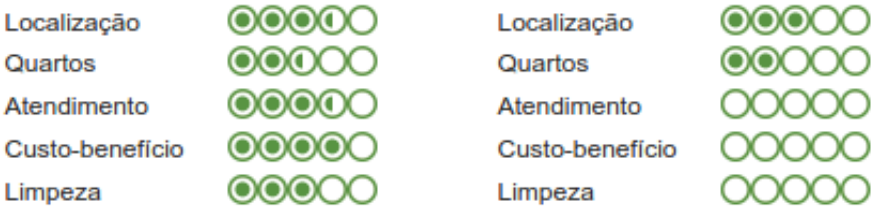

Figure E.2: TripAdvisor and Proposed Method Summary of Hotel 1.

Figure E.3: TripAdvisor and Proposed Method Summary of Hotel 2.

Figure E.4: TripAdvisor and Proposed Method Summary of Hotel 3.

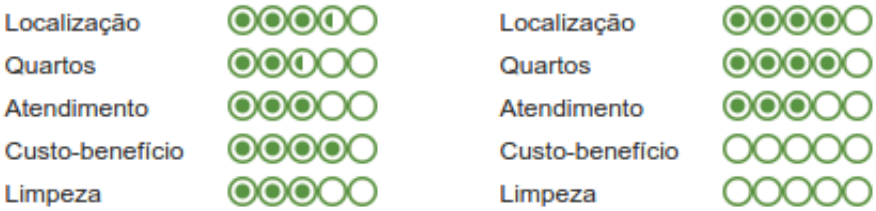

Figure E.5: TripAdvisor and Proposed Method Summary of Hotel 4.

Figure E.6: TripAdvisor and Proposed Method Summary of Hotel 5.

Figure E.7: TripAdvisor and Proposed Method Summary of Hotel 6.

Figure E.8: TripAdvisor and Proposed Method Summary of Hotel 7.

Figure E.9: TripAdvisor and Proposed Method Summary of Hotel 8.

Figure E.10: TripAdvisor and Proposed Method Summary of Hotel 9.