Departamento de Ciências e Tecnologias da Informação

Análise de Sentimento Baseada em Aspetos

Ricardo Afonso Gomes António

Dissertação submetida como requisito parcial para obtenção do grau de

Mestre em Engenharia Informática

Orientador

Doutor Ricardo Daniel Santos Faro Marques Ribeiro, Professor Auxiliar,

ISCTE-IUL

Coorientador

Doutor Fernando Manuel Marques Batista, Professor Auxiliar,

ISCTE-IUL

Junho 2017

Resumo

A análise de sentimento baseada em aspetos é uma tarefa que tem como objetivo detetar a polaridade de sentimentos e associar os mesmos a entidades. Para alcançar tal objetivo, esta tarefa pode ser dividida em duas sub-tarefas que são o reconhecimento de entidades mencionadas e a análise de sentimentos. A tarefa de reconhecimento de entidades mencionadas tem como objetivo identificar as entidades e o respetivo tipo de entidade, enquanto a tarefa de análise de sentimento tem como objetivo identificar a polaridade dos sentimentos expressos.

O presente documento analisa e compara várias abordagens adotadas para resolver estas duas tarefas, sendo aprofundadas abordagens baseadas em modelos de Markov não observáveis (HMM) e Campo Aleatório Condicional (CRF) para realizar a tarefa de reconhecimento de entidades mencionadas e, também, abordagens baseadas em classificadores Naive Bayes e modelos de Regressão Logística para realizar a tarefa de análise de sentimento baseada em aspetos.

As experiências realizadas no âmbito da tarefa de reconhecimento de entidades mencionadas apresentadas neste documento classificam as entidades em 4 tipos (Pessoa, Organização, Local ou Evento), e os resultados obtidos foram avaliados segundo os critérios apresentados na conferência CoNLL2002.

As experiências de análise de sentimentos classificam a polaridade dos sentimentos em 4 tipos (Positivo, Negativo, Neutro ou Conflito) e os resultados obtidos foram avaliados segundo os critérios da competição SemEval-2014.

Abstract

The Aspect-based Sentiment Analysis (ASBS) is a task that is designed to detect the polarity of sentiments and associate those sentiments to entities. To achieve this goal this task can be divided into two sub-tasks that are the Named-entity Recognition (NER) and Sentiment Analysis (SA). The goal of Named-entity Recognition task is to identify named entities and identify the type of entity. The goal of Sentiment Analysis task is to identify the polarity of expressed sentiments.

This paper analyzes and compares several approaches adopted in these tasks and deepens approaches based on Hidden Markov Models (HMM) and Conditional Random Fields (CRF) to perform the task of Named-entity Recognition, as well as approaches based on Naive Bayes classifiers and Logistic Regression models to perform the task of Aspect-based Sentiment Analysis.

The experiments performed in the scope of the task Named-entity Recognition presented in this document, classify the entities into 4 types (Person, Organization, Place and Event), and the results obtained were evaluated according to the criteria presented at the CoNLL2002 conference.

The experiments of aspect-based sentiment analysis, classify the sentiment into 4 types (Positive, Negative, Neutral or Conflict), and the results obtained were evaluated according to the criteria of SemEval-2014 competition.

Palavras Chave
Keywords

Palavras chave

Análise de Sentimento Baseada em Aspetos

Reconhecimento de Entidades Mencionadas

Análise de Sentimento

Modelo de Markov não observáveis

Campo Aleatório Condicional

Naive Bayes

Regressão Logística

Keywords

Aspect-based Sentiment Analysis

Named-Entity Recognition

Sentiment Analysis

Hidden Markov Model

Conditional Random Field

Naive Bayes

Logistic Regression

Agradecimentos
Acknowledgements

Gostaria de agradecer...

... ao meu orientador e coorientador, Professor Ricardo Ribeiro e Professor Fernando Batista, por toda a disponibilidade e apoio.

... aos meus pais, Maria Alice Afonso Gomes António e Vitor Manuel Gomes António, pela educação e valores transmitidos.

... ao meu irmão, Pedro Afonso Gomes António, e à minha namorada, Sofia de Almeida Carvalho, pelo apoio e incentivo.

Gostaria ainda de dedicar este estudo à memória dos amigos, Ana Maria Henriques de Almeida Carvalho e Aníbal dos Santos Diz.

Lisboa, Junho de 2017
Ricardo António

Índice

Lista de Figuras

Lista de Tabelas

Abreviaturas

ABSA Aspect-Based Sentiment Analysis

CoNLL Computational Natural Language Learning

CRF Conditional Random Field

HMM Hidden Markov Model

NER Named-Entity Recognition

POS Part of Speech

SA Sentiment Analysis

1 Introdução

1.1 Motivação

Hoje em dia, dado o acesso fácil à Internet, quer por dispositivos móveis quer por computadores pessoais, o volume de informação presente na Internet aumenta cada vez mais. A cada minuto que passa são publicados cerca de 350 mil *tweets* na rede social Twitter, partilhados 3,3 milhões de conteúdos e 6,9 milhões de mensagens enviadas na rede social Facebook.

Cada vez mais utilizadores expressam as suas opiniões em redes sociais *online*. Contudo, verifica-se este aumento da quantidade de informação não só nas redes sociais, mas também no jornalismo *online* cada vez mais popular (por exemplo, na área de finanças: Bloomberg, Reuters, Diário Económico, etc.).

O impacto da análise de sentimentos na análise de documentos e das opiniões aí presentes é cada vez maior. Por exemplo, aplicações como o Google Product Search ou o Microsoft Bing Shopping atribuem pontuações gerais aos produtos e aos vários aspetos identificados para cada produto, com base em comentários de utilizadores.

A análise de sentimento baseada em aspetos permitem a construção de uma opinião global de uma determinada entidade ou aspeto, através de um conjunto de textos (i.e., informação partilhada por utilizadores em redes sociais, ou até mesmo notícias).

Com esta análise será possível obter respostas para perguntas como as seguintes:

- Qual a satisfação global para um determinado produto?
- Satisfação das pessoas em relação a um determinado acontecimento?
- Opinião global de acionistas em relação a uma determinada ação?

1.2 Enquadramento

A tarefa de análise de sentimento, igualmente conhecida como extração de opiniões, tem vindo a ter cada vez mais importância, quer a nível académico quer a nível organizacional. Esta tarefa permite identificar o sentimento expresso por um ou mais autores em qualquer meio de divulgação social (e.g., Facebook, Twitter, Jornais,...). Esta informação é importante se o sentimento da pessoa influenciar as pessoas que a rodeiam. O objetivo da tarefa é detetar a polaridade (isto é, se o sentimento é positivo ou negativo) de um determinado texto (Pang et al., 2002). Num texto onde só existe um sentimento positivo ou negativo (e.g., "O iPhone é fantástico!") a tarefa de análise de sentimento consiste em identificar a polaridade desse sentimento. Em outros casos, onde são identificados sentimentos com polaridade oposta ou são expressos sentimentos especificamente em relação a diferentes aspetos (e.g., "O iPhone é fantástico, mas a bateria não dura mais de um dia.") torna-se necessária uma abordagem baseada em aspetos. A análise de sentimento baseada em aspetos tem como objetivos identificar as entidades (e.g., iPhone), os seus aspetos (e.g., Bateria, Ecrã,...) e os sentimentos expressos (Liu, 2012).

1.3 Questões de Investigação

A tarefa de análise de sentimento baseada em aspeto pode ser realizada recorrendo a várias abordagens. Pretende-se estudar algumas abordagens e identificar as que permitem alcançar a classificação de entidades e sentimentos para uma determinada língua com um valor aceitável.

Pretende-se estudar e implementar mais do que uma abordagem de forma a ser possível identificar qual das abordagens se comporta melhor com um determinado conjunto de textos e língua.

Através deste estudo pretende-se implementar um sistema que seja capaz de:

- Identificar entidades e aspetos, incluindo entidades e aspetos compostos por uma (e.g., Google) ou mais palavras (e.g., João Jardim).
- Identificar a polaridade dos aspetos em pelo menos duas classes (i.e., Sentimento Positivo e Sentimento Negativo).

Ambas as tarefas apresentam vários desafios. Por exemplo, na tarefa de identificação de entidades, distinguir se esta é composta por uma ou mais palavras. Adicionalmente, pretende-se

identificar a que classe corresponde a entidade (i.e., Localidade, Pessoa, Evento, Organização,...). Na tarefa de identificação da polaridade, um dos desafio é identificar a polaridade de um aspeto sendo expressos vários sentimentos na mesma frase.

1.4 Objetivos

Tendo em conta a quantidade de informação presente na Internet (Tweets, Jornais, etc.), pretende-se desenvolver um sistema que analise o sentimento em relação a uma entidade e seus aspetos. Para atingir este fim adotar-se-á uma abordagem que começa por identificar as entidades (e.g., iPhone) e os seus aspetos (e.g., bateria, ecrã, ...) e após esta fase associar a cada aspeto mencionado o sentimento expresso.

Adicionalmente, pretende-se analisar um determinado conjunto de dados de modo a criar uma estruturação de entidades versus aspetos (*Sentiment Ontology Tree*, apresentada por Wei and Gulla, 2011) e gerar um mapa com uma avaliação geral dos sentimentos em relação aos aspetos das várias entidades encontradas.

1.5 Método de Investigação

Através do estudo realizado pretende-se desenvolver um artefacto recorrendo a modelos, trazendo rigor ao desenvolvimento deste artefacto. A fim de atingir este objetivo serão tidos em conta três ciclos. O ciclo de rigor (Estado da Arte) para ser possível obter conhecimento sobre os fundamentos, experiências e artefactos já desenvolvidos. O ciclo de relevância (Requisitos) que permite determinar se as iterações adicionais são necessárias neste projeto de pesquisa científica, visto que os artefactos tecnológicos criados podem possuir deficiências em termos de funcionalidades (e.g., desempenho, usabilidade, entre outros), podendo condicionar a sua utilização. E, por fim, o ciclo do Projeto (Implementação), constituído por várias atividades de pesquisa repetitivas e de construção do artefacto tecnológico. Neste último ciclo são igualmente realizadas avaliações de forma a atingir um nível satisfatório.

Este projeto irá seguir as boas práticas apresentadas no *Management Information Systems Quarterly* (Hevner and Chatterjee, 2010):

1. Elaborar como um artefacto: Um sistema que, dado um conjunto de dados, permita extrair as entidades, os seus aspetos e a polaridade associada;

2. Relevância do problema abordado: Este tipo de abordagem tem tido um impacto crescente na análise de documentos e das opiniões, como referido nas subsecções anteriores. A motivação principal para a realização deste trabalho é a possibilidade de construir uma opinião global de uma determinada entidade ou aspeto, através de um conjunto de textos (i.e., informação partilhada por utilizadores em redes sociais ou notícias);

3. Avaliação do projeto: Para avaliar os resultados obtidos, pretende-se participar numa competição a fim de que seja possível comparar as características do projeto realizado com outros projetos.

4. Contribuições da pesquisa: Um sistema que, dado um documento, permita: identificar as entidades e seus aspetos ai presentes; detetar a polaridade em relação aos aspetos encontrados; opcionalmente, detetar a categoria de cada aspeto e sua polaridade.

5. Rigor da pesquisa: Pretende-se comparar com outras abordagens através de avaliações constantes.

6. Elaborar como um processo de investigação: Realizar estudos sobre outras abordagens conhecendo assim o trabalho que tem vindo a ser feito ao longo dos anos neste domínio.

7. Comunicação da pesquisa: Pretende-se escrever e publicar artigos para comunicar os resultados obtidos e obter opiniões de pessoas envolvidas na área.

1.6 Avaliação

A avaliação da performance realiza-se calculando a Precisão (*Precision*), Cobertura (*Recall*) e Medida F (*F-Measure*), que neste tipo de sistemas são calculada da seguinte forma:

$$\text{Precisão} = \frac{TruePositives}{TruePositives + FalsePositives}$$

$$\text{Cobertura} = \frac{TruePositives}{TruePositives + FalseNegatives}$$

$$\text{Medida F} = 2 * \frac{Precisão * Cobertura}{Precisão + Cobertura}$$

A variável *TruePositives* corresponde ao número de vezes que foi registado um resultado correto, *FalsePositives* corresponde à quantidade de resultados não esperados e *FalseNegatives* corresponde ao número de resultados em falha.

Por outras palavras, no que diz respeito ao reconhecimento de entidades mencionadas, a precisão corresponde à percentagem de entidade mencionadas encontradas pelo sistema que estão corretamente classificadas. A cobertura corresponde à percentagem de entidade mencionadas presentes no conjunto de dados que são encontradas pelo sistema.

Outra métrica usada na avaliação é a taxa de acerto (*Accuracy*):

$$\text{Taxa de acerto} = \frac{TruePositives+TrueNegatives}{TruePositives+TrueNegatives+FalseNegatives+FalsePositives}$$

1.7 Estrutura do Documento

Este documento encontra-se estruturado em quatro partes essenciais: estado da arte, reconhecimento de entidades mencionadas, análise de sentimento baseada em aspetos, e a conclusão. Na primeira parte pretende-se apresentar o trabalho que tem vindo a ser desenvolvido na área de processamento computação da língua para realizar a tarefa de análise de sentimento baseada em aspetos. Na segunda parte são descritos os modelos usados e as experiências realizadas no âmbito da tarefa de reconhecimento de entidades mencionadas. Na terceira parte são descritos os modelos usados e as experiências realizadas no âmbito da tarefa de análise de sentimento baseada em aspetos. Na última parte são discutidos os resultados obtidos nas duas tarefas, identificadas e apresentadas possíveis melhorias para o sistema desenvolvido ao longo das experiências.

2 *Estado da Arte*

Neste capítulo é apresentado o estado da arte da tarefa de análise de sentimento baseada em aspetos. Este capítulo começa por apresentar algumas abordagens usadas para realizar a tarefa de análise de sentimentos. De seguida é apresentada a tarefa de análise de sentimento baseada em aspectos, que é decomposta por duas sub-tarefas, reconhecimento de entidades mencionadas e identificação da polaridade. Ainda neste capítulos são apresentados modelos probabilísticos e não probabilísticos usados para realizar estas sub-tarefas.

2.1 Análise de Sentimento

O objetivo da análise de sentimento, tal como referido anteriormente, é detetar a polaridade de um determinado texto. A polaridade pode ser classificada como positiva, negativa ou neutra, embora existam abordagens com mais níveis de classificação (e.g., o projeto OpeNER 2014 que usa as classificações positiva, negativa, neutra e fortemente positiva e fortemente negativa).

Esta tarefa tem como objetivo identificar automaticamente os sentimentos expressos em textos de determinadas áreas como política, cinema, hotelaria, restauração, etc., o que pode ser usado para melhorar a relação com clientes, a satisfação dos clientes em relação aos serviços ou produtos, aplicações como extração de sentimentos de mensagens e extração do ânimo de acionistas em relação a uma determinada ação (Bollen et al., 2010).

Um exemplo de uma aplicação de extração de sentimentos é a aplicação apresentada na competição SemEval-2013 por Kiritchenko, Zhu e Mohammad (2014b), na tarefa de *Sentiment Analysis in Twitter*, que extrai sentimentos de mensagens curtas (i.e., SMS e *tweets*). Esta aplicação usou uma abordagem probabilística supervisionada para a análise de sentimento.

Esta tarefa pode basear-se num léxico de sentimentos (Castellucci et al., 2014), isto é, uma lista de palavras e frases com um sentimento atribuído (e.g., Fantástico - Positivo). Alguns léxicos permitem adicionar uma pontuação ao sentimento (e.g., Fantástico, Positivo, 2.1), permitindo atribuir um peso e avaliar a intensidade do sentimento (e.g., a palavra “Fantástico”

deveria ter um peso superior comparada com a palavras “Bom”). Assim o léxico de sentimentos fornece informação sobre a polaridade da palavra. Em casos particulares, a polaridade pode sofrer alterações. Um caso particular desta alteração é a negação (e.g., a palavra “gosto” que expressa um sentimento positivo, caso seja negada, ou seja, a existência da palavra “não”, passa a transmitir um sentimento negativo).

2.2 Análise de Sentimento Baseada em Aspetos

A tarefa de análise de sentimento baseada em aspetos tem como objetivos, identificar as entidades, os seus aspetos e os sentimentos expressos.

A primeira tarefa a ser efetuada numa abordagem baseada em aspetos consiste em identificar todos os aspetos presentes em cada frase (e.g., preço, comida, empregado, ...). Esta tarefa deve ser aplicada a todo o conteúdo mesmo na ausência de qualquer sentimento expresso (i.e, polaridade neutra).

As etapas desta tarefa são:

- Identificar entidades e aspetos;
- Identificação da polaridade.

2.2.1 Reconhecimento de Entidades Mencionadas

A tarefa de identificação de entidades e aspetos pode ser realizada recorrendo ao reconhecimento de entidades mencionadas. O reconhecimento de entidades mencionadas tem como objetivo localizar as entidades que são mencionadas no texto, como por exemplo: Organizações, Pessoas, Locais e Eventos.

A tarefa de reconhecimento de entidades mencionadas, também conhecida por NER (*Named-Entity Recognition*), apresenta bastante complexidade. Vejamos as seguintes frases:

1. João Jardim é eleito presidente.
2. Vou passar no jardim.

Como podemos ver nas frases 1 e 2, encontra-se presente a palavra “jardim” com significados diferentes, o que pode levar a dificuldades na classificação do tipo de entidade. Além deste

tipo de problemas, existem outros desafios na tarefa de reconhecimento de entidades, como encontrar as entidades, identificar entidades que possuam mais do que uma palavra (e.g., João Jardim) e classificar essas mesmas entidades. Para resolver este problema, precisamos de extrair informações do texto, as entidades, e classificá-las corretamente.

A maioria das abordagens usadas na tarefa de identificação de entidades baseiam-se em técnicas de extração de informação (Hu and Liu, 2004), embora existam outras abordagens como abordagens baseadas em regras, apresentada por Popescu e Etzioni (Popescu and Etzioni, 2005), e mais recente as abordagens baseadas em modelos de tópico usando o algoritmo Latent Dirichlet Allocation (Brody and Elhadad, 2010).

As abordagens mais comuns recorrem a vários modelos e sistemas como as seguintes:

- Adaboost (Carreras et al., 2002);
- Modelos de Markov não Observáveis (*Hidden Markov Model*) (Malouf, 2002);
- Modelo de Markov Condicional (*Conditional Markov Model*) (Jansche, 2002);
- Modelo de Campo Aleatório Conditional (*Conditional Random Field*) (McCallum and Li, 2003).

Em 2002, no contexto da conferência CoNLL (*Computational Natural Language Learning*), decorreu uma competição que propôs aos participantes desenvolver sistemas de reconhecimento de entidades para um conjunto de dados em Espanhol (326 mil palavras) e Holandês (292 mil palavras) (Sang, 2002).

No CoNLL-2002 foram apresentadas várias abordagens, como por exemplo a de Xavier Carreras, Lluís Márques e Lluís Padró (2002) que dividia a tarefa em duas fases, reconhecimento (NER) e classificação (NEC) usando o algoritmo de aprendizagem automática Adaboost para classificação. A primeira fase da abordagem, o reconhecimento, tinha como objetivo classificar se uma palavra pertencia a uma classe entidade ou não, tendo como segunda fase, classificar as entidades em quatro classes: Organizações, Pessoas, Locais e Eventos. Na classificação das palavras são tidas em conta algumas características, como por exemplo: saber se uma palavra contém dígitos, possui hífen, entre outras.

Ainda nesta competição foram apresentadas outras abordagens, como a de Robert Malouf (2002) que recorreu a modelos de Markov não observáveis e a abordagem de Martin Jansche (2002) que também recorreu a modelos de Markov mas apenas para classificar se a palavra era candidata a ser entidade ou não. Ainda na abordagem de Martin Jansche foi usado um classificador Naive Bayes para classificar o tipo de entidade.

2.2.2 Identificação da Polaridade

A análise de sentimento consiste em associar um sentimento a uma frase ou texto, este sentimento pertence a um conjunto de valores (e.g., Positivo, Negativo ou Neutro) como foi abordado anteriormente.

Esta tarefa pode ser realizada com diferentes níveis de complexidade, desde um nível baixo de complexidade como identificar se uma frase ou texto possui um sentimento positivo ou negativo, até um nível elevado de complexidade como identificar o sentimento num conjunto superior a duas classes. Na identificação do sentimento pode ser usado um léxico de palavras ou expressões que expressam sentimentos positivos ou negativos.

Em 2012, realizou-se a competição TASS *(Taller de Análisis de Sentimientos en la SEPLN*) que propôs aos seu participantes duas tarefas para um conjunto de dados de 68 mil *tweets* em Espanhol. A primeira tarefa correspondia a uma análise de sentimentos, ou seja, determinar a polaridade dos textos divulgados na competição. A segunda tarefa correspondia à identificação do tópico do texto (Villena-Román et al., 2012).

Esta competição juntou nove participantes. Um dos participantes foi o grupo L2F-INESC-ID (Batista and Ribeiro, 2013), que obteve na primeira tarefa 63,37 % de Precisão para 5 níveis de polaridade (Fortemente Positiva, Positiva, Fortemente Negativa, Negativa e Neutra) e 69,05 % para 3 Níveis de polaridade (Positiva, Negativa e Neutra). Na tarefa de deteção do tópico obteve 65,37 %, alcançado a primeira posição na tarefa de deteção do tópico e a segunda posição em análise de sentimentos. Foi apresentada por este grupo uma abordagem baseada em classificadores binários de máxima entropia (*Binary Maximum Entropy*) para análise automática de sentimento e classificação de tópico.

Nesta competição também foram apresentadas abordagens supervisionadas baseadas em classificadores Support Vector Machines (SVM) e Multinomial Naive Bayes (MNB).

Pela oitava vez, em 2014, a competição SemEval (Pontiki et al., 2014) com a intenção de incentivar a investigação na tarefa de análise de sentimentos baseado em aspetos (*Aspect-based Sentiment Analysis*), colocou em desafio quatro tarefas, que são as seguintes: extração de aspetos (Tarefa 1), identificação da polaridade dos aspetos (Tarefa 2), identificação da categoria a que pertencia o aspeto (Tarefa 3) e a identificação da polaridade das categorias (Tarefa 4). Esta competição onde participaram 32 equipas fornecia dois conjuntos de dados (Computadores e Restaurantes), com cerca de 8 mil palavras (6 mil para o conjunto de dados de treino e 2 mil para o conjunto de dados de teste).

Foram apresentadas nesta competição abordagens como as seguintes:

- Abordagem baseada no modelo Campo Aleatório Conditional, apresentada pela equipa JU_CSE (Patra et al., 2014);
- Abordagem baseada num sistema de votação, apresentada pela equipa SINAI (Jiménez-Zafra et al., 2014);
- Abordagem recorrendo ao uso de métodos supervisionados para análise de sentimento baseada em aspetos, apresentada pela equipa lsis_lif (Hamdan et al., 2014) e pela equipa NRC-Canada-2014 (Kiritchenko et al., 2014a);
- Abordagem recorrendo a aprendizagem estruturada (*Structured Learning*), apresentada pela equipa UNITOR (Castellucci et al., 2014).

A abordagem aqui mencionada que se destacou mais na competição foi a abordagem da equipa UNITOR. Esta abordagem recorreu a um classificador Support Vector Machines (SVM) para realizar as quatro tarefas propostas e obteve os resultados presentes na Tabela 2.1.

UNITOR	Conjunto de dados	
	Computadores	Restaurantes
Tarefa 1: Extração de aspetos	67,95 %	80,09 %
Tarefa 2: Identificação da polaridade dos aspetos	62,99 %	74,95 %
Tarefa 3 e 4: Deteção de categoria e sua polaridade	85,26 %	76,29 %

Tabela 2.1: Resultados obtidos pela equipa UNITOR na competição SemEval-2014. Os valores apresentados correspondem à Medida F.

No ano 2015, realizou-se a competição SemEval-2015 (Androutsopoulos et al., 2015), que dado um conjunto de revisões de clientes, teve como desafio as seguintes tarefas:

- Identificar o alvo da opinião, isto caso seja expresso (e.g., "Wonderful restaurant, especially the fajitas" -> {"restaurant","fajitas"});
- Identificar a categoria dos aspetos (e.g., "The fajitas were delicious, but the salad was awful" -> {"fajitas":FOOD , "salad":FOOD})
- Detetar a polaridade dos sentimentos (e.g., "The fajitas were delicious, but expensive" -> {"fajitas":FOOD:positive, "fajitas":PRICE:negative})

Pretendia-se com estas três tarefas extrair um quíntuplo por cada revisão feita, como por exemplo para a frase "*Avoid this place!*" devolver {target:"place" category:"RESTAURANT#GENERAL" polarity:"negative" from:"11" to:"16"}).

2.3 Tipologia das Abordagens

Nesta secção são apresentadas inúmeras abordagens com base em modelos probabilísticos e modelos não probabilísticos.

2.3.1 Modelos Probabilísticos

O modelo Campo Aleatório Conditional (CRF) é um modelo probabilístico usado no reconhecimento de padrões, aprendizagem automática e várias tarefas de processamento de língua natural, como reconhecimento de entidades e extração de informação.

Em 2001, John Lafferty, Andrew McCallum e Fernando Pereira apresentaram este modelo para construir modelos probabilísticos que segmentam e classificam sequências de dados. Também foi apresentado como uma alternativa aos modelos de Markov não observáveis e ao uso de gramáticas estocásticas.

Uma das vantagens que este modelo fornece é permitir adicionar características para classificar a sequência, facilitando a tarefa de reconhecimento de entidades mencionadas.

Na competição SemEval2014 foi apresentada pela equipa JU_CSE uma abordagem baseada no modelo Campo aleatório conditional para a análise de sentimento baseado em aspetos, que na sua avaliação obteve um medida F de 59,37 % no conjunto de dados sobre Computadores e 72,34 % no conjunto de dados sobre Restaurantes. O modelo Campo aleatório conditional foi usado por esta equipa nas três fases: Identificação de entidades e aspetos, identificação de categorias e identificação dos sentimentos expressos. Esta equipa recorreu à implementação presente na ferramenta CRF++ 0.58 para efetuar a experiência.

Na primeira fase da abordagem, identificação de entidades e aspetos, a equipa recorreu ao uso das seguintes características:

- Classes gramaticais;
- Frequência das classes gramaticais;
- Se a palavra se encontra antes de um verbo;
- Palavras inanimadas;
- Informação ontológica (Liu, 2012).

Após concluída a primeira fase foi usado novamente o CRF++ para identificar aspetos que só são constituídos por uma palavra.

Na segunda fase, onde foi realizada a identificação de categorias, recorreu-se a características como:

- Classes gramaticais;
- Relações de dependência (*Stanford dependency relations*, obtido através da ferramenta Stanford NLP Parser);
- Léxico de sinónimos (WordNet).

Para concluir a experiência, na tarefa de identificação dos sentimentos expressos foram usadas as características:

- Classes gramaticais;
- Léxico de sentimentos (SentiWordNet);
- Palavras que expressam sentimentos (e.g., Palavras positivas, negativas ou neutras);
- Número de frases que possuem a palavra (contando apenas uma vez a mesma revisão);

A característica classe gramatical, também conhecida como parte do discurso, foi usada em todas as fases da experiência por possuir um papel importante na identificação de entidades, aspetos e sentimentos.

Para obter as classes gramaticais a equipa recorreu ao Stanford CoreNLP.

2.3.2 Modelos não Probabilísticos

A equipa UNITOR (Castellucci et al., 2014), já referida na secção anterior, apresentou na tarefa de extração de aspetos um processo de classificação sequencial que começava por classificar as palavras em três classes, *beginning* (B), *inside* (I) e *outside* (O). Através desta classificação inicial (e.g., $[\text{The}]_O[\text{fried}]_B[\text{rice}]_I[\text{is}]_O[\text{amazing}]_O[\text{here}]_O$) é possível identificar aspetos (e.g., The $[\text{fried rice}]_{AspectTerm}$ is amazing here). Esta classificação foi realizada recorrendo a um classificador Support Vector Machines (SVM).

O SVM é um modelo não probabilístico supervisionado usado para classificação e regressão, que ao analisar o conjunto de dados tenta reconhecer padrões. O reconhecimento de padrões é realizado através da descoberta de fronteiras lineares, que separam dados pertencentes a cada uma das classes.

Existem vários implementações do SVM que podem ser usadas:

- SVMLight;
- LIBSVM;
- Scikit-learning (biblioteca implementada na linguagem Python, que para além do classificador SVM possui um diverso universo de classificadores).

3 *Reconhecimento de Entidades Mencionadas*

Neste capítulo são apresentadas as abordagens realizadas para efetuar a identificação de entidades. Este capítulo começa por apresentar o conjunto de dados fornecidos na conferência CoNLL e o processamento que este sofreu no âmbito das experiências realizadas. De seguida são apresentados os modelos de classificação usados nas experiências, os modelos de Markov não observáveis (HMM) e o modelo Campo aleatório conditional (CRF). Após a apresentação dos modelos são apresentadas as experiências realizadas no âmbito do reconhecimento de entidades mencionadas. Por fim é apresentada uma discussão sobre os resultados obtidos, sendo estes comparados com outras abordagens, e identificados os trabalhos a realizar para melhorar o sistema implementado.

3.1 Conjunto de Dados

Como referido anteriormente, em 2002, no contexto da conferência CoNLL (Sang, 2002), decorreu uma competição que propôs aos participantes desenvolverem sistemas de reconhecimento de entidades mencionadas independentes da língua. Foram disponibilizados dois conjuntos de dados, um conjunto de dados na língua Espanhola com 264 mil palavras para treino e 51 mil palavras para testes e um conjunto de dados na língua Holandesa com 202 mil palavras para treino e 69 mil palavras para testes. Ambos os conjuntos de dados possuem por cada palavra a respetiva etiqueta NER. Na Tabela 3.1 encontra-se um excerto do conjunto de dados fornecido em Espanhol.

Palavra	**NER**
La	B-LOC
Coruña	I-LOC
,	O

Tabela 3.1: Excerto do conjunto de dados em Espanhol fornecidos no CoNLL-2002.

Nas experiências realizadas neste documento foram usadas as etiquetas sugeridas na confe-

rência CoNLL onde as entidades são divididas em 4 tipos:

- PER, que corresponde a uma entidade do tipo Pessoa;
- ORG, que corresponde a uma entidade do tipo Organização;
- LOC, que corresponde a uma entidade do tipo Localização;
- MISC, que corresponde a uma entidade do tipo Evento.

Todos estes tipos de entidades dão origem a duas etiquetas, que correspondem ao início (B) e ao interior (I) de uma entidade (i.e., para o tipo PER vão existir as etiquetas B-PER e I-PER que corresponde ao início de uma entidade do tipo Pessoa e ao interior de uma entidade do tipo Pessoa). Para além das etiquetas que correspondem a entidades, é necessária a existência de uma etiqueta que corresponde a palavras que não são entidades, que será denominada como etiqueta O. No total serão usadas neste documento 9 etiquetas para classificação de entidades. Na Figura 3.1 é visível o universo de etiquetas usadas e a cor que será usada para representar a etiqueta em gráficos.

Figura 3.1: Conjunto de etiquetas usadas para classificação nas experiências de reconhecimento de entidades mencionadas.

3.2 Modelos de Markov não Observáveis

Os modelos de Markov não observáveis, também conhecidos como HMMs, são modelos generativos usados para efetuar a predição de sequência (Rabiner, 1989), ou seja, pretende-se associar a cada símbolo observado um estado, obtendo assim uma sequência de estados (y = y_1,..., y_n) que correspondem à sequência de símbolos observados (x = x_1,..., x_n).

Como estamos a aplicar estes modelos a reconhecimento de entidades mencionadas, os símbolos correspondem a palavras e os estados correspondem a etiquetas NER.

Na Tabela 3.2 são apresentadas algumas notações usadas neste documento para descrever os modelos de Markov.

Símbolo	Descrição
x	Sequência de palavras observadas
y	Sequência de estados
X	Universo de palavras observadas
Y	Universo de estados
T	Número total de palavras de uma sequência
N	Número total de estados
$\hat{y}$	Melhor sequência de estados não observáveis

Tabela 3.2: Notações usadas para descrever os modelos HMM.

Os modelos de Markov assumem que a probabilidade de um determinado estado depende apenas de um número limitado de estado anterior, e podemos formalizar esta assunção com a seguinte fórmula:

$$P(y_i|y_1...,y_{i-1}) = P(y_i|y_{i-1}) \tag{3.1}$$

Também é assumido por estes modelos que um determinado estado depende apenas do estado anterior, e que a probabilidade de uma palavra observada x_i depende do estado y_i que produziu a observação. Esta assunção pode ser traduzida para a seguinte fórmula:

$$P(x_i|y_1...,y_{i-1},...,y_n,x_1...,x_{i-1},...,x_n) = P(x_i|y_i) \tag{3.2}$$

Com estas duas assunções podemos verificar que como os estados não são observados mas as palavras são observadas e dependem de um estado é possível obter o estado correspondente através do cálculo do estado com maior probabilidade associada. Para realizar a predição de sequências com este modelo é necessário obter as probabilidades presentes na Tabela 3.3.

Probabilidade distribuída	Observação
$P_{\text{início}}(y_1\|start) = P_{\text{início}}(y_1)$	Probabilidade de uma sequência começar num determinado estado
$P_{\text{transição}}(y_i\|y_{i-1})$	Probabilidade de transitar de uma estado para outro
$P_{\text{emissão}}(x_i\|y_i)$	Probabilidade de observar uma determinada palavra estando num determinado estado
$P_{\text{final}}(stop\|y_i) = P_{\text{final}}(y_i)$	Probabilidade de uma sequência terminar num determinado estado

Tabela 3.3: Probabilidades usadas nos modelos HMM.

Estas probabilidades são obtidas através da contagem de ocorrência de estados e palavras num conjunto de dados.

$$P_{\text{início}}(y_1) = \frac{\text{Número de sequências que começam no estado } y_1}{\text{Número total de sequências}} \tag{3.3}$$

$$P_{\text{transição}}(y_i|y_{i-1}) = \frac{\text{Número de vezes que é observado o estado } y_i \text{ depois do estado } y_{i-1}}{\text{Número de vezes que é observado o estado } y_{i-1}} \tag{3.4}$$

$$P_{\text{emissão}}(x_i|y_i) = \frac{\text{Número de vezes que o estado } y_i \text{ emite a palavra } x_i}{\text{Número de vezes que é observado o estado } y_i} \tag{3.5}$$

$$P_{\text{final}}(y_i) = \frac{\text{Número de sequências que terminam no estado } y_i}{\text{Número total de sequências}} \tag{3.6}$$

Dando uso às probabilidades apresentadas, é possível agregar e construir uma fórmula que permite obter a probabilidade de uma sequência de palavras x corresponder a uma sequência de estados y.

$$P(x,y) = P_{\text{início}}(y_1) * \left(\prod_{i=1}^{n-1} P_{\text{transição}}(y_{i+1}|y_i) \right) * P_{\text{final}}(y_i) * \left(\prod_{i=1}^{n} P_{\text{emissão}}(x_i|y_i) \right) \tag{3.7}$$

Na aplicação prática desta fórmula probabilística podem ocorrer dois tipos de problemas computacionais. O primeiro problema pode ocorrer quando apenas observarmos uma vez uma determinada palavra num conjunto de dados de treino de grande dimensão. A probabilidade associadas dessa palavra pode ser de tal forma tão pequena que pode provocar erros na capacidade de armazenamento da unidade de memória usada (*underflow*). Para resolver este problema podemos aplicar a função logarítmica nos cálculos de forma a aumentar o valor da probabilidade evitando assim valores muito pequenos. O uso da função logarítmica nos cálculos não só resolve o problema apresentado, como também permite efetuar cálculos computacionais com menor custo. A função logarítmica pode ser aplicada sempre que sejam efetuadas multiplicações da seguinte forma:

$$\log(\exp(a) * \exp(b)) = a + b \tag{3.8}$$

e nas adições, assumindo que a é superior a b, aplicamos a função logarítmica da seguinte forma:

$$\log(\exp(a) + \exp(b)) = a + \log(1 + \exp(b - a)) \tag{3.9}$$

O segundo problema pode ocorrer durante a contagem de palavras. Quando não é visualizada uma determinada palavra no conjunto de dados de treino, não é correto atribuir uma

probabilidade de zero pois pode ocorrer no conjunto de dados de teste e corresponder a uma entidade. Desta forma, deve ser adicionada uma probabilidade mínima aos casos não observados durante o treino. Um exemplo de um mecanismo que resolve este problema é a estimativa de Laplace que incrementa uma ocorrência a todas as palavras e no cálculo das probabilidades é adicionado o tamanho do dicionário ao denominador, ou seja:

$$P_{Adiciona+1}(x_i|x_{i-1}) = \frac{Contagem(x_{i-1}|x_i) + 1}{Contagem(x_{i-1}) + \text{tamanho do dicionário}} \tag{3.10}$$

3.2.1 Determinar a Melhor Sequência de Estados

Após o cálculo das respetivas probabilidades, existe a necessidade de obter a melhor sequência de estados não observados para as palavras observadas. Para um modelo HMM com Z estados e uma sequência de W palavras, existem Z^W sequências de estados não observáveis possíveis. É de notar que o número de sequências de estados cresce exponencialmente, tornando a tarefa de obtenção do melhor caminho possível para uma sequência de palavras de grande dimensão bastante custosa em termos de cálculos computacionais. Para resolver este problema é possível dar uso ao algoritmo de programação dinâmica Viterbi.

O algoritmo de Viterbi tem como objetivo obter a melhor sequência de estados não observáveis para uma sequência de palavra. Esta sequência conhecida por caminho de Viterbi, é obtida através do pseudo-código apresentado no Algoritmo 3.1. Em resumo, o algoritmo começa por criar uma matriz onde as linhas correspondem a estados (incluíndo os estados de início e fim) e as colunas correspondem às palavras da sequência observada. Após criada esta estrutura são preenchidos todas as posições da matriz com a maior probabilidade de alcançar o estado em análise:

$$\hat{y} = \max_{y_0,\cdots,y_{i-1}} P(y_0,...,y_{i-1},x_0,...,x_{i-1}) \tag{3.11}$$

Depois de preenchida a matriz, é obtido o caminho de Viterbi seguindo o caminho com maior probabilidade (que foi guardado durante o preenchimento da matriz).

Algoritmo 3.1 Pseudo-código do algoritmo de Viterbi.

function VITERBI(observations *x*of len T, state-graph of len N) **returns** melhor sequência de estados

criação de matriz com as probabilidades das sequências viterbi[N+2,T]
for cada estado s **from** 1 **to** N **do**
viterbi[s,1] $\longleftarrow P_{\text{início}}(s|start) * P_{\text{emissão}}(x_1|s)$
backpointer[s,1] $\longleftarrow 0$
for cada palavra observada t **from** 2 **to** T **do**
for cada estado s **from** 1 **to** N **do**
viterbi[s,t] $\longleftarrow \left(\max_{s'=1}^{N} viterbi[s',t-1] * P_{\text{transição}}(s|s') \right) * P_{\text{emissão}}(x_i|s)$
backpointer[s,t] $\longleftarrow \left(\operatorname{argmax}_{s'=1}^{N} viterbi[s',t-1] * P_{\text{transição}}(s|s') \right)$
viterbi[q_F,T] $\longleftarrow \max_{s=1}^{N}$ viterbi[s,T] * $P_{\text{final}}(stop|q_F)$
backpointer[q_F,T] $\longleftarrow \operatorname{argmax}_{s=1}^{N}$viterbi[s,T] * $P_{\text{final}}(stop|q_F)$
return o caminho com maior probabilidade seguindo os estados guardados no backpointer a partir do backpointer[q_F , T]

3.3 Campo Aleatório Condicional

Os modelos de Markov não observáveis, apresentados na secção anterior, não são práticos para representar características ou dependências entre observações de longo alcance. Em alternativa, é possível dar uso ao modelo Campo aleatório condicional, também conhecido como CRF. Enquanto a predição nos modelos HMMs têm por base a sequência, a predição nos modelos CRFs têm por base as características da sequência.

Nas tarefas da área do processamento de língua natural, mais concretamente reconhecimento de entidades mencionadas, existem características presente em palavras ou sequências que podem ajudam o reconhecimento de entidades, como por exemplo verificar se uma palavra possui dígitos ou se uma palavra possui letras maiúsculas. Uma característica muito importante nesta área é a classe gramatical, também conhecida como parte do discurso ou POS (*Part of Speech*). Esta característica permite não só obter informação sobre a palavra que estamos a observar, como também obter informação da sequência a que pertence a palavra observada. Por exemplo, uma palavra etiquetada com a classe gramatical nome tem maior probabilidade de ser uma entidade do que uma palavra etiquetada com a classe verbo.

Tal como o modelo HMM, o modelo CRF, apresentado por Lafferty, McCallum e Pereira (2001), possui um X que corresponde a uma sequência de palavras e um Yque corresponde a

uma sequência de etiquetas NER. Lafferty, McCallum e Pereira definem o modelo como sendo um modelo unidirecional que pode ser dividido em dois conjuntos (X e Y). Esta definição é apresentada pelos autores da seguinte forma:

$G = (V,E)$ é um grafo, onde Yé indexado pelos vértices de G. Então (X,Y) é um campo aleatório condicional nos casos, quando condicionados por X, a variável aleatória Y_v obedece à propriedade de independência de Markov: $p(Y_v|X,Y_w,w\neq v) = p(Y_v|X,Y_w,w \sim v)$, onde $w \sim v$ significa que w e v são vizinhos em G.

Os autores apresentam a probabilidade conjunta de uma sequência de etiquetas Yobservando uma dada sequência de palavras X através da seguinte fórmula:

$$p(y|x) = \exp\left(\sum_{e\in E,k} \lambda_k f_k(y_{i-1},y_i,x,i) + \sum_{v\in V,k} \mu_k g_k(y_i,x,i)\right) \tag{3.12}$$

onde:

- $f_k(y_{i-1},y_i,x,i)$ corresponde à função característica da transição da sequência observada e das etiquetas na posição i e $i-1$ da sequência de etiquetas;
- $g_k(y_i,x,i)$ corresponde à função característica de emissão de uma etiqueta na posição i e da sequência observada;
- λ_ke μ_k corresponde a parâmetros fixo que são estimados a partir do conjunto de dados de treino.

A esta fórmula ainda é aplicado um fator normalizador $Z(x)$, de forma a distribuir as probabilidades pelos vários estados (que correspondem a etiquetas):

$$p(y|x,\lambda) = \frac{1}{Z(x)} \exp \sum_{i=1}^{n} \sum_{j} \lambda_j f_j(y_{i-1},y_i,x,i) \tag{3.13}$$

$$Z(x) = \sum_{y\varepsilon Y} \sum_{i=1}^{n} \sum_{j} \lambda_j f_j(y_{i-1},y_i,x,i) \tag{3.14}$$

As funções características são funções que avaliam uma determinada característica, como por exemplo a existência de hífen, e retornam valores reais, normalmente valores binários. A seguinte função exemplifica uma função característica que avalia a existência de hífen:

$$f_j(y_{i-1}, y_i, x, i) = \begin{cases} 1, & \text{se a observação na posição i possui hífen} \\ 0, & \text{caso contrário} \end{cases}$$

A tarefa de identificação de características é uma das tarefas mais importantes para uma boa aplicação do modelo CRF. As características usadas devem fornecer informação relevante para o tipo de classificação pretendida, caso contrário, ao adicionarmos características que não possuem informação estamos a prejudicar as características que possuem informação relevante.

É de salientar que é possível replicar um modelo HMM através de um modelo CRF, apenas definindo uma característica por cada par de etiquetas (y',y) e uma característica por cada par de observação e etiqueta (y,x). Estas duas características correspondem às probabilidades de transição e emissão apresentadas na secção anterior.

3.4 Experiências

Esta secção começa por apresentar o processamento do conjunto de dados que foi necessário realizar antes de dar início às experiências. A seguir à apresentação do processamento dos dados foram apresentadas as experiências realizadas no âmbito da tarefa de reconhecimento de entidades mencionadas.

3.4.1 Processamento do Conjunto de Dados

Para realizar as experiências presentes nesta secção foi usado o conjunto de dados da língua Espanhola disponibilizado para a competição apresentada na conferência CoNLL-2002.

Na experiência com o modelo HMM foi usado o conjunto de dados fornecido sem qualquer adição de informação, ou seja, o conjunto de dados apenas era constituído por palavras e as respetivas etiquetas NER.

Para as duas primeiras experiências com o modelo CRF, modelo que permite adicionar características, foi adicionado ao conjunto de dados a classe gramatical de cada palavra, obtida a partir do sistema Stanford Part-of-Speech Tagger. Na Figura 3.2 é possível visualizar um excerto do conjunto de dados após este processo.

Palavra	POS	NER
La	DA	B-LOC
Coruña	NC	I-LOC
,	Fc	O

Figura 3.2: Excerto do conjunto de dados da língua Espanhola da competição CoNLL-2002 com a classe gramatical (POS).

No âmbito da experiência de 3 camadas, que iremos abordar mais a frente, foram adicionadas, para além da classe gramatical, as seguintes características:

- Se a palavra é uma entidade ou não é entidade (EO);
- Se a palavra corresponde ao início de uma entidade, se está no interior da entidade ou não pertence a uma entidade (BIO).

Estas duas características são extraídas a partir da etiqueta NER e, como iremos ver mais a frente, são usadas para efetuar a avaliação de cada uma das camadas do sistema. Na Figura 3.3 encontra-se um excerto do conjunto de dados usado na experiência de três camadas.

Palavra	POS	EO	BIO	NER
La	DA	E	B	B-LOC
Coruña	NC	E	I	I-LOC
,	Fc	O	O	O

Figura 3.3: Excerto do conjunto de dados da língua Espanhola da competição CoNLL-2002 com a classe gramatical (POS), etiqueta EO e etiqueta BIO.

3.4.2 Modelo de Markov não Observáveis

Para realizar a experiência com o modelo HMM foi usada a ferramenta lxmls-toolkit, disponibilizada na escola de verão LXMLS (*Lisbon Machine Learning School*, http://lxmls.it.pt/2015/), que possui uma implementação do modelo HMM.

Antes de iniciar a tarefa de reconhecimento, houve a necessidade de adaptar o módulo de leitura dos dados para suportar a estrutura do conjunto de dados da conferência CoNLL-2002. De seguida, foi extraído a partir do conjunto de dados original um pequeno conjunto de dados, com cerca de 1000 palavras, que foi usado para definir, através de várias classificações, os valores dos parâmetros de configuração que melhor se ajustavam ao tipo de dados. Por fim, foi

efetuada a classificação com o conjunto de dados completo obtendo os resultados apresentados na Tabela 3.4 e a visualização da matriz de confusão das predições na Figura 3.4.

Conjunto de dados Espanhol	Precisão	Cobertura	Medida F
Treino	98,70 %	99,04 %	98,87 %
Teste	95,09 %	11,91 %	21,18 %

Tabela 3.4: Resultados obtidos para o conjunto de dados Espanhol da competição CoNLL-2002, recorrendo ao modelo HMM.

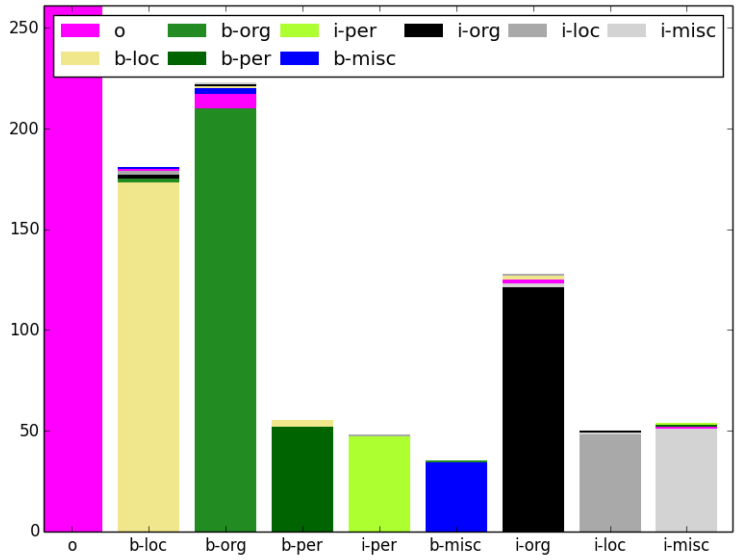


Figura 3.4: Visualização da matriz de confusão obtida para o conjunto de dados Espanhol da competição CoNLL-2002, recorrendo ao modelo HMM.

3.4.3 Campo Aleatório Condicional

Ao todo foram realizadas três experiências com o modelo CRF. Nas duas primeiras experiência, denominadas Experiência Completa e Experiência com 3 Camadas, foi usada a ferramenta lxmls-toolkit, que para além de disponibilizar uma implementação do modelo HMM também disponibiliza uma implementação do modelo CRF. Na terceira experiência foi usada a ferramenta Stanford NER disponibilizada pelo grupo de processamento da linguagem natural da Universidade de Stanford (https://nlp.stanford.edu/software/CRF-NER.shtml).

3.4.3.1 Experiência Completa

Antes de realizar as experiências com a ferramenta lxmls-toolkit foi necessário identificar as características que consideramos importantes para distinguir entidades e tipos de entidades. As características identificadas foram:

- Palavra observada;
- Número de vezes que a sequência começa numa determinada palavra (características de início);
- Número de vezes que se transita de um estado para outro (características de transição);
- Número de vezes que se emite a palavra observada estando um determinado estado (características de emissão);
- Número de vezes que a sequência termina com a palavra observada (características de fim);
- Classe gramatical da palavra observada (POS)*;*
- Palavra que se encontra antes da palavra observada (PW);
- Classe gramatical da palavra que se encontra antes da palavra observada (PPOS)
- Palavra que se encontra depois da palavra observada (AW);
- Classe gramatical da palavra que se encontra depois da palavra observada (APOS);
- Se a palavra observada contém letras maiúsculas;
- Se a palavra observada possui dígitos;
- Se a palavra observada possui hífen;
- Prefixo da palavra observada (3 letras iniciais);
- Sufixo da palavra observada (3 letras finais).

As características que têm por base as palavras que rodeiam a palavra observada, foram adicionadas por possuírem informação sobre a estrutura da frase e por poderem possuir informação sobre a presença de uma entidade na sua vizinhança. Todas as outras características identificam entidades a excessão do prefixo e sufixo que permitem excluir possíveis entidades.

Após definidas as características, foi necessário alterar a ferramenta lxmls-toolkit de forma a ser possível extrair do conjunto de dados a classe gramatical de cada palavra e adicionar essa mesma classe às funções características.

Com as características configuradas, foi realizada a primeira experiência com modelo CRF que obteve os resultados presentes na Tabela 3.5 e a visualização da matriz de confusão das predições na Figura 3.5.

Conjunto de dados Espanhol	Precisão	Cobertura	Medida F
Treino	78,32 %	79,70 %	79,00 %
Teste	71,38 %	71,19 %	71,29 %

Tabela 3.5: Resultados obtidos para o conjunto de dados Espanhol da competição CoNLL-2002, recorrendo ao modelo CRF.

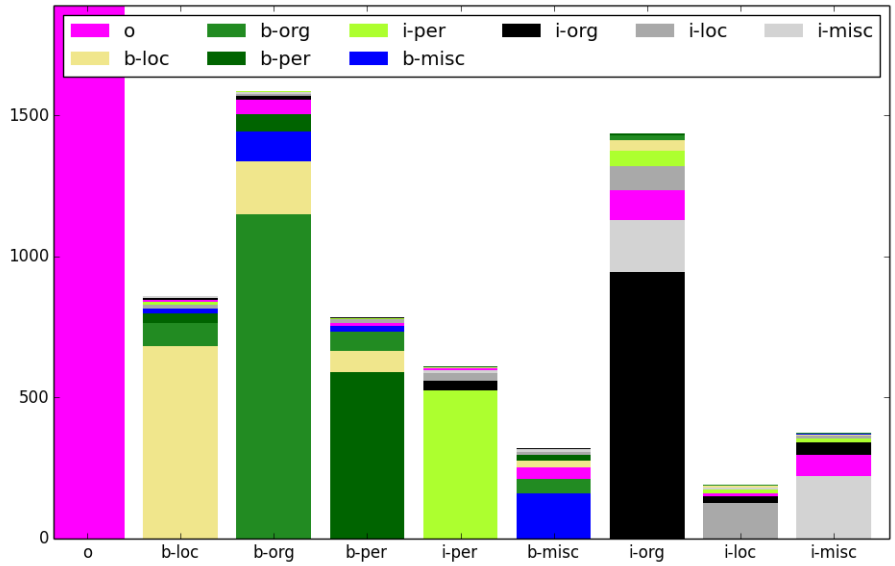


Figura 3.5: Visualização da matriz de confusão obtida para o conjunto de dados Espanhol da competição CoNLL-2002, recorrendo ao modelo CRF.

3.4.3.2 Experiência com 3 Camadas

Com o bom desempenho prestado pelo modelo CRF foi efetuada uma experiência com fases de análise, semelhante ao que foi feito na abordagem realizada pela equipa UNITOR (Castellucci et al., 2014) na competição SemEval-2014 e pela abordagem Adaboost (Carreras et al., 2002) na competição CoNLL.

No âmbito desta experiência, foram realizadas três pequenas experiências com o intuito de esclarecer se era vantajoso dividir o reconhecimento de entidades em fases sequenciais, e usar a predição obtida como características na fase seguinte.

Começou-se por fazer a experiência EO que classificava se uma palavra pertencia a uma entidade. De seguida, a experiência BIO que classificava se uma palavra pertencia ao início de uma entidade (B, *Beginning*), interior de uma entidade (I, *Inside*) ou não pertencia a uma entidade (O, *Outside*). Por último, com o mesmo fim que a experiência BIO, foi realizada a experiência BIO2 que era composta por duas fases. A primeira fase classificava se uma palavra pertencia a uma entidade, tal como a experiência EO, e a segunda fase classificava da mesma forma que a experiência BIO mas com mais uma característica que dava uso à predição obtida na primeira fase.

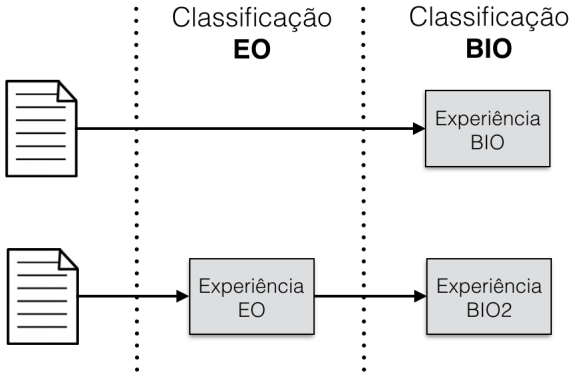


Figura 3.6: Diagrama das camadas usadas para classificação da etiquetas BIO.

Na Tabela 3.6 são apresentados os resultados obtidos nestas três pequenas experiências. É de salientar que a performance na experiência BIO2 comparando com a experiência BIO, embora pequena, foi ligeiramente superior nas métricas Cobertura e Medida F.

Conjunto de dados Espanhol	Precisão	Cobertura	Medida F
EO	93,95 %	94,45%	94,19 %
BIO	90,68 %	90,40 %	90,54 %
BIO2	90,38 %	90,80 %	90,59 %

Tabela 3.6: Resultados obtidos nas experiências EO, BIO e BIO2, para o conjunto de dados Espanhol da competição CoNLL-2002, recorrendo ao modelo CRF.

Após estas três pequenas experiências, foi construído um sistema com três camadas com o objetivo de dividir a tarefa de reconhecimento de entidades em três sub-tarefas e verificar se era

vantajoso classificar entidades em pequenos grupos e reutilizar essa predição para classificações mais complexas, como a atribuição de uma das nove etiquetas NER. Estas três sub-tarefas são executadas sequencialmente devido à dependência das predições das camadas anteriores.

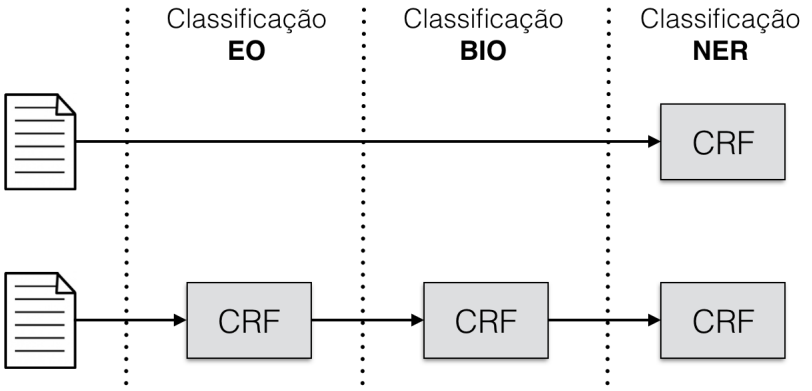


Figura 3.7: Diagrama das camadas usadas para classificação da etiqueta NER.

Primeira Camada: Entidade ou Não Entidade (EO)

A primeira camada tem como objetivo identificar se uma palavra corresponde a uma entidade (E) ou não corresponde a uma entidade (O). Nesta camada foram usadas as mesmas características da primeira experiência com o modelo CRF. O resultado obtido nesta camada é apresentado na Tabela 3.7.

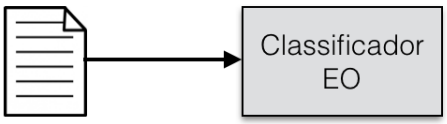


Figura 3.8: Desenho da estrutura e percurso de classificação, na experiência da primeira camada.

Conjunto de dados Espanhol	Precisão	Cobertura	Medida F
Treino	93,73 %	95,25 %	94,48 %
Teste	93,94 %	94,45 %	94,20 %

Tabela 3.7: Resultados obtidos na primeira camada (EO), para o conjunto de dados Espanhol da competição CoNLL-2002, recorrendo ao modelo CRF.

Segunda Camada: Início, Interior ou Fora da Entidade (BIO)

A segunda camada, que depende das predições realizadas pela primeira camada, tem como objetivo identificar se uma palavra corresponde ao início de uma entidade (B), interior de uma entidade (I) ou não pertence a uma entidade (O).

Esta segunda camada usa todas as características da primeira camada e ainda foram adicionadas três novas características a partir das predições da primeira camada. Estas três novas características são:

- Se a palavra observada é uma entidade ou não (EO);
- Se a palavra anterior à palavra observada é uma entidade ou não (PEO);
- Se a palavra a seguir à palavra observada é uma entidade ou não (AEO).

O resultado obtido nesta segunda camada é apresentado na Tabela 3.10.

Figura 3.9: Desenho da estrutura e percurso de classificação, na experiência da segunda camada.

Conjunto de dados Espanhol	Precisão	Cobertura	Medida F
Treino	99,09 %	99,08 %	99,09 %
Teste	90,38 %	90,81 %	90,59 %

Figura 3.10: Resultados obtidos na segunda camada (BIO), para o conjunto de dados Espanhol da competição CoNLL-2002, recorrendo ao modelo CRF.

Terceira Camada: Etiqueta NER

A terceira e última camada desta experiência tem como objetivo atribuir a cada palavra observada uma etiqueta NER. Esta camada realiza a mesma classificação que a primeira experiência realizada com o modelo CRF mas com mais características.

As características usadas nesta camada são as mesmas da segunda camada e ainda foram adicionadas as seguintes três características criadas a partir das predições obtidas na camada anterior:

- Se a palavra observada corresponde ao início, interior ou não pertence a uma entidade (BIO);

- Se a palavra anterior à palavra observada corresponde ao início, interior ou não pertence a uma entidade (PBIO);

- Se a palavra a seguir à palavra observada corresponde ao início, interior ou não pertence a uma entidade (ABIO).

Ao todo, são usadas mais 6 características que a Experiência Completa, que não recorre a uma estrutura por camadas.

O resultado final desta experiência de 3 camadas é apresentado na Tabela 3.8.

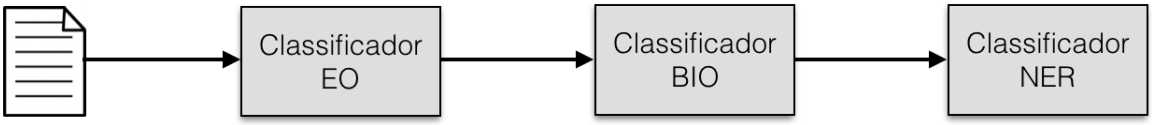


Figura 3.11: Desenho da estrutura e percurso de classificação, na experiência da terceira camada.

Conjunto de dados Espanhol	Precisão	Cobertura	Medida F
Treino	82,03 %	82,03 %	82,03 %
Teste	67,61 %	67,91 %	67,76 %

Tabela 3.8: Resultados obtidos na terceira camada (NER), para o conjunto de dados Espanhol da competição CoNLL-2002, recorrendo ao modelo CRF.

Na Figura 3.12 é apresentado um diagrama das camadas usadas nas duas experiências que têm como fim a atribuição de uma etiqueta NER por palavra observada. Neste mesmo diagrama são apresentados os resultados obtidos em cada uma das camadas e a dependência entre elas. A primeira linha do gráfico corresponde à Experiência Completa realizada com o modelo CRF que deu uso a 15 características. A segunda linha corresponde à Experiência com 3 camadas, onde são usadas como características as predições das camadas anteriores. Nesta última experiência são usadas num total 21 características, sendo que neste conjunto de características estão presentes as 15 características usadas na primeira experiência.

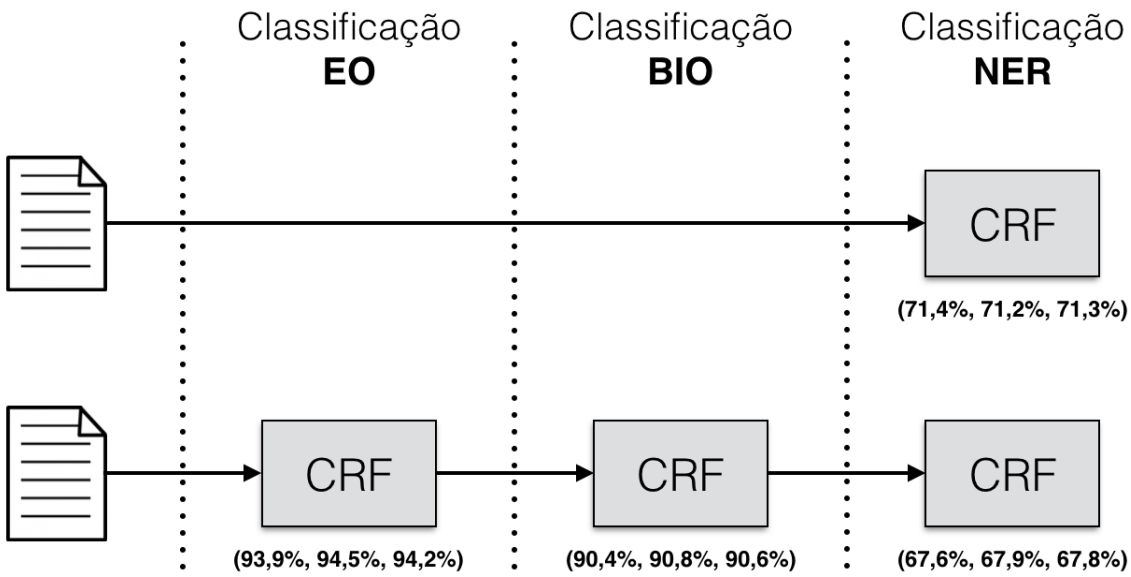


Figura 3.12: Diagrama de camadas usadas nas experiências com o modelo CRF. Os valores apresentados por baixo de cada módulo correspondem à Precisão, Cobertura e Medida F.

3.4.3.3 Stanford NER

Como apresentado anteriormente foi realizada uma experiência com a ferramenta Stanford NER. Esta ferramenta, também conhecida como CRFClassifier, disponibiliza a implementação do modelo sequencial CRF (Finkel et al., 2005).

Para esta experiência não foi usado o conjunto de dados com a classe gramatical, como nas restantes experiências com o modelo CRF. A ausência da classe gramatical no conjunto de dados deve-se ao facto das características usadas pelo Stanford NER serem muito semelhantes às características usadas pela ferramenta Stanford POS Tagger. A ferramenta Stanford POS Tagger é uma ferramenta usada para realizar a classificação da classe gramatical, o que torna o uso da classe gramatical desnecessário.

Nesta ferramenta as características já se encontram implementadas e são definidas de acordo com parâmetros de configurações. Alguns dos parâmetros usados nas configurações desta experiência foram:

- *useTitle*, que verifica se a palavra observada está presente numa lista de títulos, como por exemplo: Mr, Mrs, entre outros;
- *useWord*, que adiciona como característica a palavra observada;

- *useWordPairs*, que adiciona como características dois pares de palavras que correspondem à palavra observada e às palavras vizinhas ((w_{i-1}, w_i) e (w_i, w_{i+1}));
- *useNGrams*, que adiciona como características conjuntos de letras com tamanhos de 1 a 6 letras;
- *maxNGramLeng*, que limita o tamanho dos n-gramas usados (este foi usado para limitar a dimensão dos prefixos e sufixos até tamanho 6);
- *usePrev*, que adiciona características que têm por base a palavra anterior (w_{i-1});
- *useNext*, que adiciona características que têm por base a palavra seguinte (w_{i+1});
- *useSequences* e *usePrevSequences*, que permite características combinadas, isto é, características que combinam a palavra observada e a classe gramatical correspondente ou a palavra observada e a classe gramatical da palavra anterior;
- *useLongSequences*, que adiciona como características o restante conteúdo da sequência;
- *featureDiffThresh*, descarta todas as características que não possuem um peso superior a 0,05.

Realizadas as configurações foi efetuado o treino do modelo que extraiu 158 061 características. Depois do treino, foi realizada a predição para o conjunto de dados testes que obteve como resultados geral os valores apresentados na Tabela 3.9. Também foram extraídos os resultados para cada um dos 4 tipos de entidades, que são apresentados na Tabela 3.10.

Conjunto de dados	Precisão	Cobertura	Medida F
Treino	99,42 %	99,27 %	99,34 %
Teste	80,90 %	79,74 %	80,32 %

Tabela 3.9: Resultados obtidos na experiência Stanford NER para o conjunto de dados Espanhol da competição CoNLL-2002.

Tipo de entidade	Precisão	Cobertura	Medida F
LOC	82,42 %	76,57 %	79,39 %
MISC	67,16 %	52,94 %	59,21 %
ORG	79,55 %	83,07 %	81,27 %
PER	86,25 %	90,48 %	88,31 %

Tabela 3.10: Resultados por tipo de entidade na experiência Stanford NER para o conjunto de dados Espanhol da competição CoNLL-2002.

3.5 Discussão

Nesta secção são discutidos os resultados obtidos nas várias experiências de reconhecimento de entidades mencionadas e comparados os resultados com outras abordagens. Na Tabela 3.11 são apresentados os resultados finais de cada uma das experiências e, para fins de comparação, são apresentados os resultados obtidos pelas equipas que participaram na competição CoNLL-2002 na Tabela 3.12.

Experiência	**Precisão**	**Cobertura**	**Medida F**
HMM	95,09 %	11,91 %	21,18 %
CRF - Experiência Completa	71,38 %	71,19 %	71,29 %
CRF - Experiência com 3 Camadas	67,61 %	67,91 %	67,76 %
CRF - Stanford NER	80,90 %	79,74 %	80,32 %

Tabela 3.11: Resultados obtidos nas experiências de reconhecimento de entidades realizadas neste documento para o conjunto de dados da língua Espanhola da competição CoNLL-2002.

Sistema	**Precisão**	**Cobertura**	**Medida F**
CMP02	81,38%	81,40%	81,39%
Flo02	78,70%	79,40%	79,05%
CY02	78,19%	76,14%	77,15%
WNC02	75,85%	77,38%	76,61%
BHM02	74,19%	77,44%	75,78%
Tjo02	76,00%	75,55%	75,78%
PWM02	74,32%	73,52%	73,92%
Jan02	74,03%	73,76%	73,89%
Mal02	73,93%	73,39%	73,66%
Tsu02	69,04%	74,12%	71,49%
BV02	60,53%	67,29%	63,73%
MM02	56,28%	66,51%	60,97%
Baseline	**26,27%**	**56,48%**	**35,86%**

Tabela 3.12: Resultados obtidos pelos participantes da competição CoNLL-2002 para o conjunto de dados da língua Espanhola.

Na competição CoNLL-2002 participou um sistema apelidado de Mal02 que inicialmente deu uso a modelos de Markov não observáveis. Este sistema obteve uma fraca prestação com o modelo HMM e a equipa que o desenvolveu acabou por substitui-lo pelo modelo de máxima entropia. Apenas este segundo modelo apresentou resultados que permitiram entrar na tabela de classificações.

Tal como a primeira versão do sistema Mal02, os resultados obtidos na experiência apresentada neste documento com os modelos HMM também apresentaram uma fraca prestação. Esta fraca prestação deve-se ao facto dos modelos HMM não ter em conta características presentes na palavra e na sequência. Como vimos anteriormente, estas características contribuem imenso para a identificação de entidades ou exclusão de entidades na tarefa de reconhecimento de entidades.

Outra desvantagem encontra-se presente na assunção realizada pelos HMM. Esta assunção, que assume independência entre a posição na sequência e a probabilidade de transitar de um estado para outro, penaliza um sistema que faz análise de sequências porque as sequências de língua natural seguem normas de sintaxe. Estas normas estão presentes ao longo da sequência e definem a relação entre as palavras.

Estas desvantagens não se verificam no modelo CRF que permite adicionar informação relevante para predição, como a classe gramatical, a existência de hífen, entre outras características. O uso deste modelo permitiu obter resultados muito superior à base de referência imposta na competição, permitindo alcançar um lugar no top 10 da tabela de participantes.

Como apresentado na secção anterior, foram realizadas três experiências com o modelo CRF. A primeira experiência focou-se na identificação de características úteis para o reconhecimento de entidades. A segunda experiência, Experiência com 3 Camadas, que surgiu com a ideia de dividir a tarefa de reconhecimento em fases e reutilizar a predição de cada uma dessas fases para realizar classificações mais complexas. Os resultados nesta segunda experiências foram inferiores nas três métricas usadas. Um problema que podemos verificar nesta experiência, e que acaba por prejudicam a classificação, é a propagação de erros, isto é, as predições realizadas pelas camadas iniciais podem ser erradas e acabam por transmitir essa informação às camadas seguintes.

A terceira experiência com o modelo CRF, que deu uso à ferramenta Stanford NER, alcançou resultados bastante bons que permitiram alcançar o segundo lugar da tabela de participantes. O sucesso desta experiência deve-se ao facto de ter sido usado um conjunto alargado de características relevantes para o reconhecimento de entidades.

Na terceira experiência foram ainda recolhidos dados bastantes interessantes no que toca à identificação do tipo de entidades. Na Tabela 3.10 podemos visualizar os resultados obtidas para cada um dos tipos de entidades e verificar que existe um tipo de entidade que é mais propício a erro, que é o tipo MISC, que corresponde a Eventos.

Análise de Sentimento Baseada em Aspetos

4

Neste capítulo é apresentada a abordagem realizada para efetuar a análise de sentimento baseada em aspetos. Este capítulo começa por apresentar o conjunto de dados fornecidos na competição SemEval-2014 e o processamento que este sofreu no âmbito das experiências realizadas. De seguida é apresentada a base de resultados e a forma como foram obtidos. Esta base de resultados é usada para comparar as abordagens de análise de sentimento e análise de sentimento baseada em aspetos realizadas ao longo deste capítulo. Por fim e após a apresentação das experiências realizadas, é apresentada uma discussão sobre as abordagens realizadas e os resultados obtidos.

4.1 Conjunto de Dados

Para realizar a tarefa de análise de sentimento baseada em aspetos foi usado o conjunto de dados fornecidos para 4º tarefa da competição SemEval-2014 (*International Workshop on Semantic Evaluation*). Nesta competição, que ocorreu em 2014, foram apresentados vários desafios em dez tarefas da área de análise semântica computacional (Pontiki et al., 2014). Algumas das tarefas foram: Análise de Sentimento Baseada em Aspetos (Tarefa 4), Análise do Texto Clínico (Tarefa 7), Análise de Dependência Semântica de Ampla Cobertura (Tarefa 8) e Análise de Sentimentos no Twitter (Tarefa 9).

Para realizar a tarefa de Análise de Sentimento Baseada em Aspetos (Tarefa 4), a competição propôs a realização de quatro sub-tarefas:

1. Identificação de aspetos;
2. Identificação da polaridade dos aspetos;
3. Identificação da categoria dos aspetos;
4. Identificação da polaridade da categoria dos aspetos.

No capítulo Reconhecimento de Entidades Mencionadas deste documento, foram realizadas abordagens que permitem resolver a 1º sub-tarefa proposta. Neste capítulo, o foco é realizar abordagens que permitem realizar a 2º sub-tarefa proposta, que corresponde a identificação da polaridade dos aspetos presentes.

O conjunto de dados fornecidos nesta competição encontra-se dividido em dois domínios: Computadores e Restaurantes, e está anotado com 4 tipos de polaridades que são: Positivo, Negativo, Conflito ou Neutro. O tipo de polaridade Conflito encontra-se presente em situações onde é expresso pelo menos um sentimento positivo e um sentimento negativo em relação a um aspeto.

As frases presentes em ambos os domínios encontram-se na língua Inglesa e foram extraídas a partir de avaliações de clientes (Ganu et al., 2009). Cada domínio possui 3 mil frases anotadas com os aspetos e respetiva polaridade. Na Figura 4.1 é apresenta uma das frases anotadas do conjunto de dados do domínio Computadores em formato XML.

```
<sentence id="175">
  <text>The decor is vibrant and eye-pleasing with several semi-
      private boths on the right side of the dining hall, which are
      great for a date.</text>
  <aspectTerms>
    <aspectTerm term="decor" polarity="positive" from="4" to="9"/>
    <aspectTerm term="dining hall" polarity="positive" from="95" to="
        106"/>
    <aspectTerm term="semi-private boths" polarity="positive" from="
        51" to="69"/>
  </aspectTerms>
  <aspectCategories>
    <aspectCategory category="ambience" polarity="positive"/>
  </aspectCategories>
</sentence>
```

Figura 4.1: Exemplo de uma frase anotada do conjunto de dados do domínio Computadores em formato XML.

Para realizar as experiências presentes neste capítulo, e com o fim de identificar a polaridade dos aspetos, foi necessário efetuar um processamento prévio. Para realizar este processamento foi criado um processo que começa por converter o formato de XML para JSON e remover as anotações: ”aspectCategories” e ”aspectCategory” que apenas são usadas nas sub-tarefas que analisam categorias. Em seguida é criada por cada aspeto identificado no conjunto de dados uma estrutura que possui: o aspeto, a frase onde se encontra e a polaridade expressa sobre o aspeto. Na Figura 4.2 é apresenta o resultado deste processamento.

```
{
"frase": "The decor is vibrant and eye-pleasing with several semi-
   private boths on the right side of the dining hall, which are
   great for a date.",
"aspeto": "decor",
"polaridade": "positive"
},
{
"frase": "The decor is vibrant and eye-pleasing with several semi-
   private boths on the right side of the dining hall, which are
   great for a date.",
"aspeto": "dining hall",
"polaridade": "positive"
},
{
"frase": "The decor is vibrant and eye-pleasing with several semi-
   private boths on the right side of the dining hall, which are
   great for a date.",
"aspeto": "semi-private boths",
"polaridade": "positive"
}
```


Figura 4.2: Exemplo de uma frase processada do conjunto de dados do domínio Computadores em formato JSON.

Ao executar este processo foram extraídos do conjunto de dados do domínio Computadores 2.358 ocorrências de aspetos, sendo 1.042 aspetos distintos, e do conjunto de dados do domínio Restaurantes 3.693 ocorrências de aspetos, sendo 1.288 aspetos distintos. A partir destas ocorrências extraídas e respeitando a ordem sequencial das frases no conjunto de dados, foram criados para cada um dos domínios dois conjuntos de dados, um conjunto para treino (80%) e outro para teste (20%) aplicando um rácio de 80:20.

4.2 Base de Resultados

Para podermos comparar os resultados das abordagens apresentadas neste capítulo, foram realizadas duas experiências que permitem obter resultados de sistemas de baixa complexidade. Ao todo foram obtidos dois resultados, um obtido através de um processo fornecido pela equipa que organizou a competição SemEval-2014, e o outro através de um simples processo que atribui a polaridade com maior frequência no conjunto de dados de treino.

O processo fornecido pela equipa do SemEval-2014 começa por dividir o conjunto de dados

em duas partes, uma para treino e outra para teste, usando um rácio de 80:20. De seguida, verifica por cada aspeto se este existe no conjunto de treino, se existir atribui a polaridade com maior frequência associada a esse aspeto, se não existir atribui a polaridade como maior frequência entre os vários aspetos do conjunto de treino. Na Tabela 4.1 são apresentados os resultados obtidos para cada um dos domínios do conjunto de dados com este processo.

Conjunto de dados	Taxa de Acerto
Computadores	47,06 %
Restaurantes	57,80 %

Tabela 4.1: Resultado obtido através do processo fornecido no SemEval-2014.

Para além do processo fornecido pela equipa do SemEval-2014, e durante a análise dos conjuntos de dados de treino, foi criado um processo que realiza a contagem de ocorrências dos tipos de polaridade. Com os resultados das contagens, apresentados na Tabela 4.2, podemos constatar que se um sistema que apenas atribuísse o tipo de polaridade com maior frequência, que neste caso é o tipo positivo, este conseguia obter os resultados apresentados na Tabela 4.3. Estas contagens são apresentadas nesta secção por possuirem informação relevantes para decisões tomadas durante a realização das experiências apresentadas neste capítulo.

Domínio	Conjunto de dados	Nº de Aspetos	Tipos de Polaridade			
			Positivo	Negativo	Neutro	Conflito
Computadores	Treino	1886	793	677	372	44
	Teste	472	194	189	88	1
Restaurantes	Treino	2954	1736	662	487	69
	Teste	739	428	143	146	22

Tabela 4.2: Número de ocorrência dos tipos de polaridade nos conjuntos de dados SemEval-2014.

Conjunto de dados	Taxa de Acerto
Computadores	41,10 %
Restaurantes	57,92 %

Tabela 4.3: Resultado obtido atribuindo apenas a polaridade com maior frequência no conjunto de dados de treino.

4.3 Métodos de Classificação

Nesta secção são apresentados os métodos que são usados nas experiências para classificar a polaridade do sentimento.

4.3.1 Textblob

A primeira experiência realizada deu uso ao sistema TextBlob. Este sistema, que se encontra desenvolvida na linguagem de programação Python, fornece uma interface que permite realizar tarefas como: obtenção de classes gramaticais de uma sequência, extração de nomes, classificação, tradução, análise de sentimentos, entre outras. Para além destas funcionalidades, este sistema disponibiliza dois modelos de classificação treinados, com conjuntos de dados na língua Inglesa fornecidos pela biblioteca NLTK. O classificador padrão disponibilizado por este sistema pertence à biblioteca Pattern (Smedt and Daelemans, 2012), que recorre a um léxico de adjetivos frequentes em avaliações de produtos. Para além do classificador padrão, PatternAnalyzer, é disponibilizado o classificador NaiveBayesAnalyzer, que corresponde a um classificador Naive Bayes. Estes dois modelos apenas realizam predições para 3 tipos de polaridade: Positivo, Negativo ou Neutro.

4.3.2 Classificador Naive Bayes

O classificador Naive Bayes surge dando uso ao teorema de Bayes que define a probabilidade de x ocorrer sabendo que y ocorre através da seguinte fórmula:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} \tag{4.1}$$

Através desta regra de Bayes e assumindo a independência entre as várias características (x) dada a classe (y), é possível obter a probabilidade de um conjunto de características ($x = x_1, x_2, ..., x_n$) corresponderem a uma determinada classe (y) através da seguinte fórmula:

$$p(x|y) = p(x_1, x_2, ..., x_n|y) = \prod_{i=1}^{n} p(x_i|y) \tag{4.2}$$

Na análise de sentimentos a classe corresponde ao tipo de polaridade e as características correspondem a informação extraída da sequência, como por exemplo: palavras que constituem a sequência, classes gramaticais, n-gramas, entre outras.

Através desta última fórmula, é possível calcula a probabilidade para cada tipo de polaridade e atribuir ao conjunto de características observadas a polaridade com maior probabilidade.

$$y* = \arg\max_{y} p(x|y)p(y) \tag{4.3}$$

4.3.3 Modelo de Regressão Logística

Um classificador de Regressão Logística, também conhecido como Máxima Entropia, baseia-se num modelo discriminativo que calcula a probabilidade de y ocorrer sabendo que x ocorre.

$$y* = \arg\max_{y} p(y|x) \tag{4.4}$$

Esta probabilidade é estimada através de uma combinação linear entre as características (x) e o seu peso (p), ou seja,

$$\sum_{i=1}^{n} p_i x_i \tag{4.5}$$

Como esta combinação é calculada através de um somatório e pode resultar em valores negativos, por causa do peso, é necessário aplicar a função exponencial, para tornar o valor positivo. Para além da função exponencial, também é necessário aplicar um fator normalizador, para tornar o resultado num valor probabilístico (entre 0 e 1).

$$P(y|x) = \frac{\exp\left(\sum_{i=1}^{n} p_i x_i\right)}{\sum_{y'\varepsilon Y} \exp\left(\sum_{i=1}^{n} p_i x_i\right)} \tag{4.6}$$

Para além destas condições, um classificador de Regressão Logística, em vez de dar uso apenas à característica (x_i), associa uma característica e uma classe (ou tipo de polaridade), através de uma função $f_i(y,x)$, como por exemplo:

$$f_1(y,x) = \begin{cases} 1, & \text{se } \text{"}bom\text{"}\varepsilon x \text{ e } y = Positivo \\ 0, & \text{caso contrário} \end{cases}$$

A variável peso é normalmente determinada automaticamente recorrendo a uma função gradiente descendente para obter um máximo global.

A equação final usada para realizar a classificação do tipo de polaridade, recorrendo a um classificador de Regressão Logística, é:

$$P(y|x) = \frac{\exp\left(\sum_{i=1}^{n} p_i f_i(y,x)\right)}{\sum_{y' \varepsilon Y} \exp\left(\sum_{i=1}^{n} p_i f_i(y',x)\right)} \tag{4.7}$$

4.4 Características

Nesta secção são apresentadas as características usadas nas experiências deste capítulo para detetar a polaridade do sentimento.

4.4.1 Análise Sintática

Por vezes numa frase encontram-se presentes vários aspetos e polaridades, como por exemplo a frase ”That Screen is awesome but the battery is really bad.” que possui o aspeto ”Screen” com um sentimento positivo associado e o aspeto ”battery” com um sentimento negativo. Para resolver este problema, e para poder ser possível efetuar a análise de sentimento apenas para o aspeto em questão, é necessário identificar as palavras que apresentam uma relação com o aspeto e dão contexto a esse aspeto.

Para identificar as palavras que apresentam uma relação direta ou indireta com o aspeto foi usada a ferramenta Stanford Parser (Manning et al., 2014). Implementada na linguagem Java, esta ferramenta disponibiliza funcionalidades para realizar uma análise sintática como: análise por constituintes e identificação de relações de dependência.

Uma análise por constituintes permite obter conjuntos de palavras presentes na frase que atuam como uma unidade. Em oposição, as relações de dependência apresentam ligações unidirecionais entre as palavras e a relação entre essas palavras.

Nas experiências realizadas neste documento, e na maioria das abordagens realizadas pelos participantes do SemEval-2014, foi usada a identificação de relações de dependência para obter as palavras que possuem uma relação com o aspeto.

Vejamos a representação das dependências obtida para as palavras da frase ”That Screen is awesome but the battery is really bad.” (na Figura 4.3 são ilustrada as dependências em forma de grafo):

det(Screen-2, That-1)
nsubj(awesome-4, Screen-2)
cop(awesome-4, is-3)
root(ROOT-0, awesome-4)
cc(awesome-4, but-5)
det(battery-7, the-6)
nsubj(bad-10, battery-7)
cop(bad-10, is-8)
advmod(bad-10, really-9)
conj(awesome-4, bad-10)

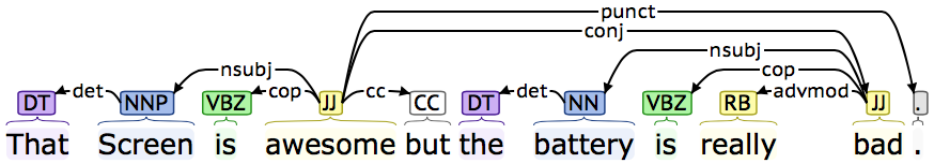


Figura 4.3: Representação das dependências para as palavras da frase ”That Screen is awesome but the battery is really bad.”

Através desta estrutura de dependências e analisando os aspetos presentes, ”Screen” e ”battery”, é possível verificar que apenas tendo em conta a relação de 1º nível, ”(awesome-4, Screen-2)” e ”(bad-10, battery-7)”, é possível obter as palavras que expressam sentimentos.

Embora neste exemplo o sentimento é facilmente detetado a partir da relação de 1º nível, existem casos onde o sentimento é expresso num conjunto de palavras ou encontra-se num nível muito distante. Um exemplo deste tipo de casos é a frase ”the microphone doesn’t work” que possui um sentimento que não é detetado numa relação de 1º nível (na Figura 4.4 são ilustradas as dependências em forma de grafo):

det(microphone-2, the-1)
nsubj(work-5, microphone-2)
aux(work-5, does-3)
neg(work-5, n’t-4)
root(ROOT-0, work-5)

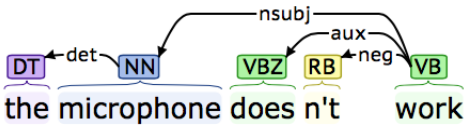


Figura 4.4: Representação das dependências para as palavras da frase ”the microphone doesn’t work.”

Tendo em conta estes casos, foi criado um processo recursivo que através da identificação de relações de dependências, extrai a sequência de palavras que apresentam relação com o aspeto, com excessão dos seguintes tipos de relação:

- Coordenações (cc), que identificam palavras como: ”and”, ”or” e ”but”;
- Conjugações (conj), que apresentam dependências com base em coordenações;
- Pontuações (punct).

Ao todo, o Stanford Parser classifica as relações em 50 tipos (M. Marneffe and B. Maccartney and C. Manning, 2006), sendo os 3 tipos de relação acima mencionados considerados como identificadores de dependências que não fornecem contexto para a análise de sentimentos do aspeto.

Na Tabela 4.4 são apresentadas algumas dependências extraídas para frases presentes no conjunto de dados SemEval-2014.

4.4.2 Léxicos de Sentimentos - SentiWordNet & Vader

Tal como na maioria dos participantes da competição SemEval-2014, foram usados, nas experiências deste capítulo, léxicos de sentimentos para extrair características das palavras que expressam sentimentos. Este tipo de léxicos permite obter a polaridade de uma palavra (positivo ou negativo) e, em alguns casos, obter a força do sentimento (normalmente entre -1 e 1 ou até entre $-\infty$ e $+\infty$).

Um dos léxicos mais usados nas abordagens dos participantes do SemEval-2014 foi o léxico SentiWordNet (Baccianella et al., 2010). Este léxico, famoso na área de análise de sentimentos, foi desenvolvido a partir do léxico de sinónimos WordNet, que possui palavras do tipo substantivos, verbos, adjetivos e advérbios associados a um conjunto de sinónimos. Dando uso a este conjunto de sinónimos e realizando uma anotação automática, o léxico SentiWordNet atribui

Frase	Aspeto	Dependências extraídas
It even has a great **webcam**, and Skype works very well.	webcam	It even has a great webcam
It even has a great webcam, and **Skype** works very well.	Skype	Skype works very well
It is easy to use, its keyboard easily accommodates large hands, and its **weight** is fantasic.	weight	its weight is fantasic
Seems to slow down occassionally but can run many applications (ie **Internet tabs**, programs, etc) simultaneously.	Internet tabs	can run many applications ie Internet tabs simultaneously

Tabela 4.4: Sequência de palavras extraídas dando uso à identificação de relações de dependências.

a cada palavra três valores que correspondem ao sentimento positivo, sentimento negativo e objetividade.

É de salientar que estes três valores variam entre 0 e 1, e apresentam uma dependência entre eles (o somatório destes três valores é sempre 1).

Em alternativa, é possível recorrer ao léxico de sentimentos Vader (C.J. Hutto and Eric Gilbert, 2014). Disponibilizado na biblioteca NLTK, o léxico Vader, para além de atribuir a uma palavra os mesmos três valores que o léxico SentiWordNet, também atribui um valor para a ausência de sentimento (i.e., polaridade neutro). Este léxico, em vez de atribuir valores de sentimentos aos sinónimos de uma palavras, atribui valores de sentimentos diretamente à palavra através de uma abordagem baseada em regras.

Tendo, este léxico, apresentado uma óptima prestação para conjuntos de dados formados a partir de avaliações de clientes, foi usado nas experiências juntamente com o léxico SentiWordNet 3.0.

As características extraídas recorrendo ao léxico de sentimentos SentiWordNet foram:

- Média do sentimento positivo dos sinónimos da palavra;
- Média do sentimento negativo dos sinónimos da palavra;
- Número de palavras da frase que possuem um valor média do sentimento positivo dos sinónimos superior a 0,25;

- Número de palavras da frase que possuem um valor média do sentimento positivo dos sinónimos superior a 0,25.

A constante 0,25 usada nas duas últimas características foi adicionado para excluir palavras que possuem um valor muito baixo de sentimento. Esta constante foi obtida através de um processo experimental.

As características extraídas recorrendo ao léxico de sentimentos Vader foram:

- Número de palavras da frase que possuem valor de sentimento positivo superior ao valor de sentimento negativo;
- Número de palavras da frase que possuem valor de sentimento negativo superior ao valor sentimento positivo;
- Somatório do valor de sentimento positivo das palavras;
- Somatório do valor de sentimento negativo das palavras.

4.4.3 Outras Características

Para além das características obtidas a partir da análise sintáticas e do uso de léxico de sentimentos, foram extraídas as seguintes características:

- Aspeto;
- Se as palavras presentes na frase foram observadas na fase de treino (*Bag of Words*);
- Se as palavras extraídas a partir da relação de dependência foram observada na fase de treino (*Bag of Words*).

4.5 Experiências

Nesta secção são apresentadas as experiências realizadas no âmbito da tarefa de análise de sentimento baseada em aspetos. As experiências encontram-se divididas em dois grupos, o primeiro grupo corresponde a experiências que atribuem ao aspeto a polaridade da frase e o segundo grupo corresponde a experiências que usam características baseadas no aspeto.

Nestas experiências foi usada a implementação do classificador Naive Bayes disponibilizada na biblioteca NLTK e a implementação do classificador Regressão Logística da biblioteca Sklearn.

4.5.1 Atribuição da Polaridade da Frase

Nesta sub-secção são apresentadas as experiências que tem como objetivo identificar o sentimento global da frase e atribuir aos aspetos presentes na frase essa mesma polaridade.

4.5.1.1 TextBlob

A primeira experiência realizada deu uso ao sistema TextBlob, que tal como apresentado anteriormente, fornece um classificador treinado para classificar frases da língua Inglesa em 3 tipos de polaridade: Positivo, Negativo ou Neutro.

Dando uso a este classificador, e assumindo o erro para sentimentos com tipo de polaridade Conflito, foi obtida a polaridade da frase e atribuída essa mesma polaridade ao aspeto.

Os resultados para esta experiência encontram-se presentes na Tabela 4.5.

Sistema	Conjunto de dados	Taxa de Acerto
TextBlob	Computadores	55,72 %
	Restaurantes	61,30 %

Tabela 4.5: Resultado obtido com a sistema TextBlob atribuindo a polaridade da frase ao aspeto.

4.5.1.2 Naive Bayes e Regressão Logística

Com o mesmo objetivo da experiência anterior, foram treinados dois classificadores, Naive Bayes e Regressão Logística, com o conjunto de dados de treino e realizada a classificação da polaridade para os aspetos do conjunto de dados de teste.

Características usadas nesta experiência foram as seguintes:

- Se as palavras presentes na frase foram observadas na fase de treino (*Bag of Words*);
- As 8 características extraídas a partir dos léxicos de sentimentos SentiWordNet e Vader.

Na Tabela 4.6 são apresentados os resultados obtidos recorrendo aos classificadores Naive Bayes e Regressão Logística.

Classificador	Conjunto de dados	Taxa de Acerto
Naive Bayes	Computadores	64,19 %
	Restaurantes	64,14 %
Regressão Logística	Computadores	67,80 %
	Restaurantes	62,25 %

Tabela 4.6: Resultado obtido usando os classificadores Naive Bayes e Regressão Logística ao atribuir a polaridade da frase ao aspeto.

4.5.2 Atribuição da Polaridade com Base no Aspeto

Nesta sub-secção são apresentadas as experiências que tem como objetivo realizar a análise de sentimento baseada no aspeto. Para alcançar este objetivo foi realizada uma análise sintática da frase, que permitiu extrair as dependências de cada aspeto. Estas dependências, tal como apresentado anteriormente, são extraídas a partir da ferramenta Stanford Parser e fornecem informação sobre a relação do aspeto com as restantes palavras da frase.

4.5.2.1 TextBlob

Depois de extraídas as dependências dos aspetos, foi usado o sistema TextBlob para classificar a polaridade da sequência de palavras que apresentam relação com o aspeto. Esta classificação foi realizada em duas experiências. A primeira experiência apenas forneceu ao sistema as palavras que apresentavam dependências. A segunda experiência para além de fornecer as palavras que apresentavam dependências também forneceu todas as palavras da frase. Os resultados obtidos nestas experiências são apresentados na Tabela 4.7.

Sistema	Palavras	Conjunto de dados	Taxa de Acerto
TextBlob	Dependências	Computadores	54,87 %
		Restaurantes	59,81 %
	Frase e Dependências	Computadores	56,57 %
		Restaurantes	61,02 %

Tabela 4.7: Resultado obtido recorrendo ao sistema TextBlob para classificar a polaridade das dependências sintáticas do aspeto.

4.5.2.2 Naive Bayes e Regressão Logística

Enquanto a experiência com o sistema TextBlob apenas teve em conta as dependências extraídas, foi realizada uma experiência com os classificadores Naive Bayes e Regressão Logística que mantiveram todas as características que possuem como base a frase:

- Se as palavras presentes na frase foram observadas na fase de treino (*Bag of Words*);
- As 8 características extraídas a partir dos léxicos de sentimentos SentiWordNet e Vader.

Além destas características, foram adicionadas as seguintes características que fornecem informação baseada no aspeto:

- Aspeto;
- Se as palavras extraídas a partir da relação de dependência foram observada na fase de treino (*Bag of Words*).

Depois de programadas estas características, foram treinados os classificadores com o conjunto de dados de treino, e de seguida efetuada a classificação da polaridade para o conjunto de dados de teste.

Os resultados obtidos nesta experiência são apresentados na Tabela 4.8.

Classificador	Conjunto de dados	Taxa de Acerto
Naive Bayes	Computadores	65,47 %
	Restaurantes	65,22 %
Regressão Logística	Computadores	68,64 %
	Restaurantes	65,22 %

Tabela 4.8: Resultado obtido dando uso aos classificadores Naive Bayes e Regressão Logística para classificar a polaridade do aspeto.

4.6 Discussão

Nesta secção são discutidos os resultados obtidos nas várias experiências de análise de sentimento baseada em aspetos e apresentado o sistema desenvolvido.

Nas Tabelas 4.9 e 4.10 são apresentados os resultados finais de cada uma das experiências para o conjunto de dados do SemEval-2014, e na Figura 4.5 são ilustrados os componentes que formam o sistema construído ao longo destas experiências.

Tal como podemos observar na tabela de resultados, as experiências realizadas obtiveram resultados muito superiores à base de resultados.

Comecemos por analisar as experiências com o sistema TextBlob. Este sistema permitiu obter resultados superiores à base de resultados. No domínio Computadores chegou a superar a base de resultados em 8,66% e no domínio Restaurantes em 3,38%.

Classificador	Nº de Aspetos Corretos	Nº de Aspetos	Taxa de Acerto
Regressão Logística	324	472	68,64 %
Naive Bayes	309	472	65,47 %
TextBlob	263	472	55,72 %
Base de Resultados (SemEval-2014)	222	472	47,06 %
Base de Resultados (todos positivo)	194	472	41,10 %

Tabela 4.9: Resultados final das experiências para o conjunto de dados do domínio Computadores do SemEval-2014.

Classificador	Nº de Aspetos Corretos	Nº de Aspetos	Taxa de Acerto
Regressão Logística	482	739	65,22 %
Naive Bayes	482	739	65,22 %
TextBlob	453	739	61,30 %
Base de Resultados (todos positivo)	428	739	57,92 %
Base de Resultados (SemEval-2014)	427	739	57,80 %

Tabela 4.10: Resultados final das experiências para o conjunto de dados do domínio Restaurantes do SemEval-2014.

Embora os valores obtidos com este classificador treinado não serem altos, comparados com as restantes abordagens, é de salientar que a prestação deste classificador foi boa, isto porque, este classificador não foi treinado com dados do mesmo domínio e apenas realiza análise de sentimento da frase.

Quanto aos outros dois classificadores usados, que deram uso ao conjunto de dados de treino do respetivo domínio, apresentaram resultados muito superiores, chegando a ultrapassar em alguns casos a base de resultados em mais de 20%.

O classificador Naive Bayes conseguiu superar a base de resultados em 18,41% no domínio Computadores, e 7,3% no domínio Restaurantes. Enquanto o classificador de Regressão Logística, que alcançou 68,64% no domínio Computadores, superou a base de resultados em 21,58% e no domínio Restaurantes apenas superou em 7,3%.

Em geral, os resultados obtidos foram bons. No entanto, os resultados obtidos para o domínio Restaurantes não escalaram tanto como os resultados para o domínio Computadores,

chegando a apresentar 14% de diferença de melhoria entre os domínios.

Neste tipo de experiências, onde os componentes mais importantes são o classificador e as características extraídas, verificámos ao longo das experiências que houve dois tipos de características que contribuíram bastante para os resultados apresentados. As mais relevantes foram as características que se baseiam na relação de dependência com o aspeto e as características baseadas no léxico de sentimentos.

Para além das características identificadas neste capítulo, foram extraídas e testadas outras características ao longo das experiências, no entanto estas não acrescentaram mais valor nenhum ou, até mesmo, prejudicavam o sistema. Estas características foram:

- Número de palavras na frase que identificam a presença de uma negação (i.e., ”not”, ”n’t”, ”none”);
- Número de palavras na frase anotadas nos léxicos de sentimentos como neutras (tanto no léxico Vader, como no léxico SentiWordNet);
- Classe gramatical;
- N-gramas (até 3 n-gramas) das palavras na frase;
- N-gramas (até 2 n-gramas) das palavras que apresentavam relação de dependência com o aspeto;
- Número de palavras da frase;
- Número de palavras que eram constituídas apenas por letras maiúsculas.

Durante as experiências foi criado um mecanismo que identificava as frases e os aspetos que eram classificados com tipos de polaridade diferentes do anotado. Com base nestas frases identificadas, verificamos que em certos casos a atribuição de polaridade é controversa até mesmo para o ser humano. Vejamos os seguintes casos presentes na Tabela 4.11.

Embora a frase nº 1 da Tabela de caso tenha sido classificado com o tipo de polaridade Negativo, não é expresso qualquer sentimento negativo nessa mesma frase.

A frase nº2, que possui um sentimento global positivo, não apresenta nenhum sentimento positivo aplicado ao aspeto OSX. A frase nº3, anotada com o tipo de polaridade Neutro, expressa um sentimento positivo que não pertence apenas ao aspeto ”desktop” mas também ao aspeto ”Core2 Quad”.

Nº	Frase (e aspeto)	Tipo de Polaridade Anotada	Tipo de Polaridade Atribuída
1	Summary: They played games with me for the **warranty period**.	Negativo	Neutro
2	I can have both **OSX** and Windows XP running at the same time!	Positivo	Neutro
3	It is much faster than my desktop which is a **Core2 Quad** running at 2.83 GHz.	Neutro	Positivo

Tabela 4.11: Algumas frases extraídas do conjunto de dados de teste que possuem uma polaridade discutível.

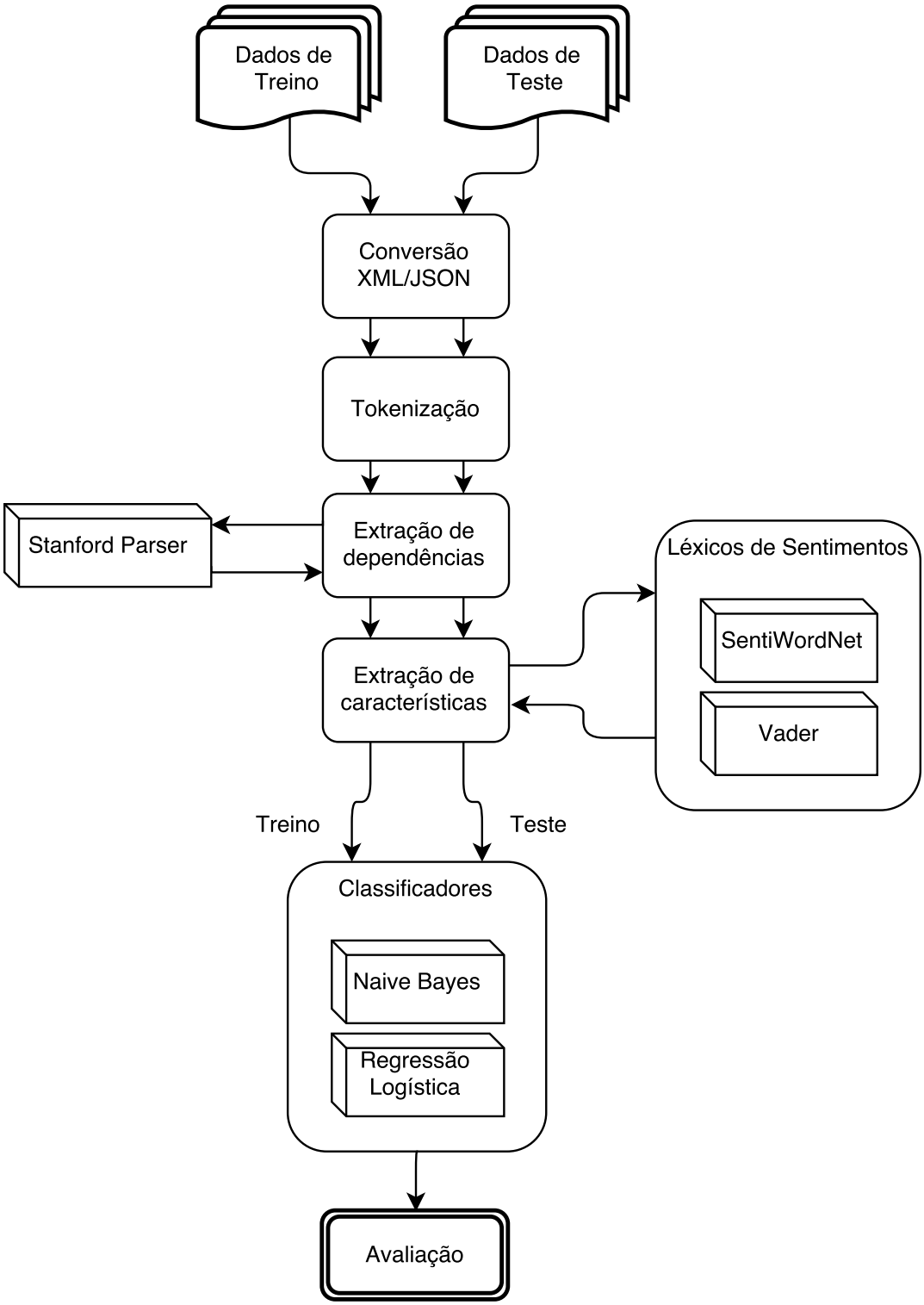


Figura 4.5: Diagrama dos componentes do sistema desenvolvido para análise de sentimento baseada em aspetos.

5 *Conclusão e Trabalho Futuro*

Neste documento foram apresentadas e comparadas inúmeras abordagens usadas para realizar a tarefa de análise de sentimento baseada em aspetos. Esta tarefa foi decomposta em duas sub-tarefas, reconhecimento de entidades mencionadas e análise de sentimentos, de forma a separar a classificação de entidades da classificação de sentimentos.

Em ambas as sub-tarefas, foram realizadas várias abordagens com modelos probabilísticos, onde foram extraídas características próprias para o objetivo de cada sub-tarefa.

No âmbito da sub-tarefa reconhecimento de entidades mencionadas, foi desenvolvido um sistema capaz de identificar e classificar entidades do tipo Pessoa, Organização, Local ou Evento. O modelo que se destacou mais nesta identificação e classificação foi o modelo Campo Aleatório Condicional (CRF), que chegou a obter valores bastante elevados nas três métricas usadas (Precisão: 80,90%, Cobertura: 79,74% e Medida F: 80,32%). Este sistema foi avaliado com um conjunto de dados na língua Espanhola de acordo com os critérios de avaliação CoNLL-2002, e os resultados foram comparados com os vários sistemas que participaram na competição.

Ainda relativamente à sub-tarefa de reconhecimento de entidades mencionadas, verificou-se que, dos 4 tipos de entidades, o tipo Evento foi o que obteve resultados mais baixos, não chegando a ultrapassar a barreira dos 60% nas métricas Cobertura e Medida F. Esta baixa prestação, para detetar uma entidade do tipo Evento, verificou-se em todos os sistemas participantes na competição.

No âmbito da sub-tarefa de análise de sentimentos, foi desenvolvido um sistema capaz de classificar o sentimento baseado no aspeto. Este sistema, que recebe um conjunto de dados onde o aspeto já se encontra identificado, realiza a classificação do sentimento com base na análise sintática, que extrai as relações de dependência associadas ao aspeto, e de léxicos de sentimentos para identificar o sentimento de cada palavra observada. Os sentimentos são classificados num dos 4 tipos de polaridade: Positivo, Negativo, Neutro ou Conflito. O modelo que se destacou mais na classificação de sentimento foi o modelo Regressão Logística, que obteve 68,64% de

taxa de acerto para o conjunto de dados do domínio Computadores e 65,22% para o domínio Restaurantes. Estes dois conjuntos de dados foram disponibilizados na competição SemEval-2014 e encontram-se na língua Inglesa.

Como não foi disponibilizado o conjunto de dados usado para testar os sistemas participantes, foram realizadas experiências para obter valores base de resultados. Ao todo obtiveram-se três valores, um obtido a partir do processo da equipa organizador do SemEval-2014, outro a partir da atribuição da polaridade mais frequente no conjunto de dados e, por fim, o último valor obtido que deu uso ao sistema TextBlob para classificar a polaridade das frase e atribuir esse mesma polaridade aos aspetos.

Contudo, os resultados obtidos nesta tarefa e respetivas sub-tarefas podem ser melhorados. Foram identificados possíveis desenvolvimentos a realizar em trabalhos futuros:

1. Testar outros classificadores.
2. Identificar e adicionar ao sistema de reconhecimento de entidades mencionadas novas características que forneçam informação relevante para detetar entidades do tipo Evento.
3. No sistema de reconhecimento de entidades mencionadas com 3 camadas, testar novos classificadores para realizar a classificação do tipo de entidade.
4. Identificar e adicionar ao sistema de análise de sentimento baseada em aspetos mais características com base em léxicos de sentimentos.
5. Identificar e adicionar ao sistema de análise de sentimento baseada em aspetos mais características com base na relação de dependência com o aspeto.
6. Identificar e adicionar características de acordo com o tipo de classificador usado. (A mesma característica pode melhorar o desempenho de um classificador e prejudicar o desempenho de outros classificadores)
7. No caso de serem avaliados conjuntos de dados provenientes de redes sociais, é necessário desenvolver um componente que identifique palavras com erros de ortografia e altere pela palavra correta, caso contrário serão processadas como palavras distintas (Exemplo de deste tipo de casos é a falta de acentuação em palavras).

Ainda no âmbito da tarefa de análise de sentimentos foi desenvolvida uma aplicação *web*, que dada uma frase é apresentado ao utilizador o sentimento dessa mesma frase através de um *emoji*. Em anexo, são apresentados exemplos desta interação. Na Figura A.1 são apresentadas frases que possuem sentimentos positivos, na Figura A.2 são apresentadas frases que possuem

sentimentos negativos, e na Figura A.3 é apresentada uma frase que possui um sentimento conflituoso ou neutro.

A frase pode ser escrita em qualquer língua, sendo usado o sistema TextBlob para detetar a língua e traduzi-la para a língua Inglesa. A análise de sentimento é realizada através do classificador Naive Bayes do sistema TextBlob por este ter sido treinado com conjuntos de dados de vários domínios.

Bibliografia

Androutsopoulos, I., Galanis, D., Manandhar, S., Papageorgiou, H., Pavlopoulos, J., and Pontiki, M. (2015). SemEval-2015 Task 12: Aspect Based Sentiment Analysis.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proc. of LREC*.

Batista, F. and Ribeiro, R. (2013). Sentiment Analysis and Topic Classification based on Binary Maximum Entropy Classifiers. *Procesamiento del Lenguaje Natural, Revista no 50 marzo de 2013*, pages 77–84.

Bollen, J., Mao, H., and Zeng, X.-J. (2010). Twitter mood predicts the stock market. *Journal of Computational Science*, pages 1–8.

Brody, S. and Elhadad, N. (2010). An Unsupervised Aspect-Sentiment Model for Online Reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, pages 804–812, Los Angeles, California.

Carreras, X., Màrquez, L., and Padró, L. (2002). Named Entity Extraction using AdaBoost. In *Proceedings of CoNLL-2002*, pages 167–170, Taipei, Taiwan.

Castellucci, G., Filice, S., Croce, D., and Basili, R. (2014). UNITOR: Aspect Based Sentiment Analysis with Structured Learning. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 761—767.

C.J. Hutto and Eric Gilbert (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text.

Finkel, J. R., Grenager, T., and Manning, C., editors (2005). *Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling.* ACL 2005, Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics.

Ganu, G., Elhadad, N., and Marian, A. (2009). Beyond the stars: Improving rating predictions using review text content. In *Proceedings of the 12th International Workshop on the Web and Databases*.

Hamdan, H., Bellot, P., and Béchet, F. (2014). Supervised methods for aspect-based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 596—600, http://alt.qcri.org/semeval2014/cdrom/pdf/SemEval104.pdf.

Hevner, A. and Chatterjee, S. (2010). Design Research in Information Systems. *Information Systems 22*, pages 9–22.

Hu, M. and Liu, B. (2004). Mining and Summarizing Customer Reviews. In *In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177.

Jansche, M. (2002). Named Entity Extraction with Conditional Markov Models and Classifiers. In *Proceedings of CoNLL-2002*, pages 179–182, Taipei, Taiwan.

Jiménez-Zafra, S. M., Martínez-Cámara, E., Martín-Valdivia, M. T., and Urenã-López, L. A. (2014). SINAI: Voting System for Aspect Based Sentiment Analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 566—571, http://alt.qcri.org/semeval2014/cdrom/pdf/SemEval099.pdf.

Jurafsky, D. and Martin., J. H. (2017). *Speech and Language Processing.*

Kiritchenko, S., Zhu, X., Cherry, C., and Mohammad, S. M. (2014a). NRC-Canada-2014: Detecting Aspects and Sentiment in Customer Reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 437–442, Dublin, Ireland.

Kiritchenko, S., Zhu, X., and Mohammad, S. M. (2014b). Sentiment Analysis of Short Informal Texts. *Journal of Artificial Intelligence Research 50*, pages 723–762.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001)*, pages 282–289.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan and Claypool.

M. Marneffe and B. Maccartney and C. Manning (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006)*, pages 449—454, Genoa, Italy. LREC, European Language Resources Association (ELRA).

Malouf, R. (2002). Markov models for language-independent named entity recognition. In *Proceedings of CoNLL-2002*, pages 187–190, Taipei, Taiwan. Alfa-Informatica Rijksuniversiteit Groningen.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

McCallum, A. and Li, W. (2003). Early results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons. In *Proceedings of CoNLL-2003*, pages 188–191, Edmonton, Canada.

OpeNER (2014). Open Polarity Enhanced Named Entity Recognition.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of EMNLP*, pages 79–86.

Patra, B. G., Mandal, S., Das, D., and Bandyopadhyay, S. (2014). JU_CSE: A Conditional Random Field (CRF) Based Approach to Aspect Based Sentiment Analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 370–374, http://alt.qcri.org/semeval2014/cdrom/pdf/SemEval063.pdf.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., and Manandhar, S. (2014). SemEval-2014 Task 4: Aspect Based Sentiment Analysis.

Popescu, A.-M. and Etzioni, O. (2005). Extracting Product Features and Opinions from Reviews. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 339—346.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77.

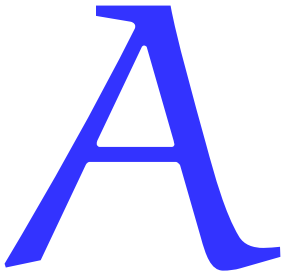
Sang, E. F. T. (2002). Introduction to the CoNLL-2002 Shared Task: Language Independent Named Entity Recognition. In *Proceedings of the 6th conference on Natural language learning*, volume 20, pages 1–4.

Smedt, T. D. and Daelemans, W. (2012). Pattern for Python. *Journal of Machine Learning Research 13*, pages 2063–2067.

Villena-Román, J., García-Morera, J., Moreno-García, C., Ferrer-Ureña, L., Lana-Serrano, S., González-Cristóbal, J. C., Westerski, A., Martínez-Cámara, E., García-Cumbreras, M. Á., Martín-Valdivia, M. T., and Ureña-López, L. A. (2012). Workshop on sentiment analysis at sepln. TASS - Taller de Análisis de Sentimientos en la SEPLN.

Wei, W. and Gulla, J. A. (2011). Enhancing the HL-SOT Approach to Sentiment Analysis via a Localized Feature Selection Framework. pages 327—335, Chiang Mai, Thailand,.

Anexos



A.1 Aplicação Web de Análise de Sentimentos

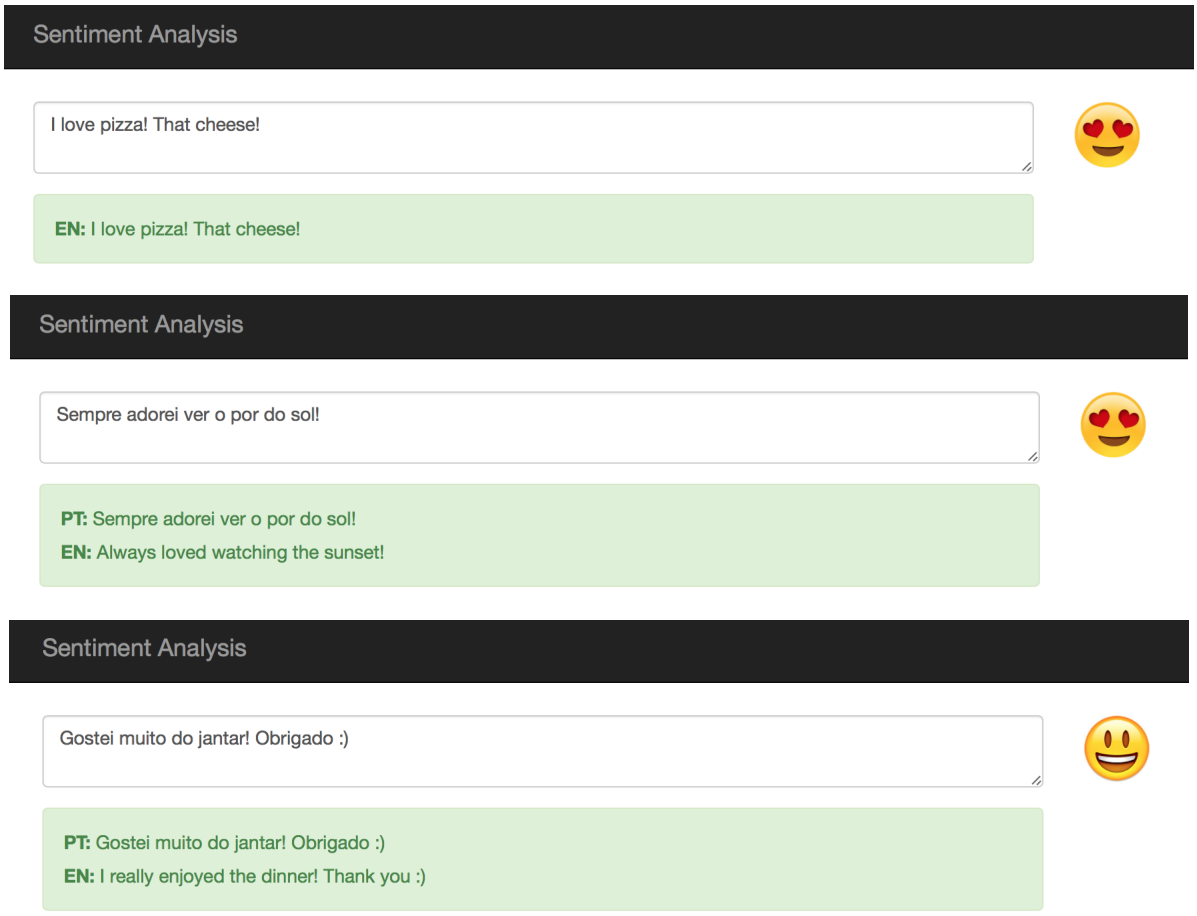


Figure A.1: Frases com sentimento fortemente positivo e positivo.

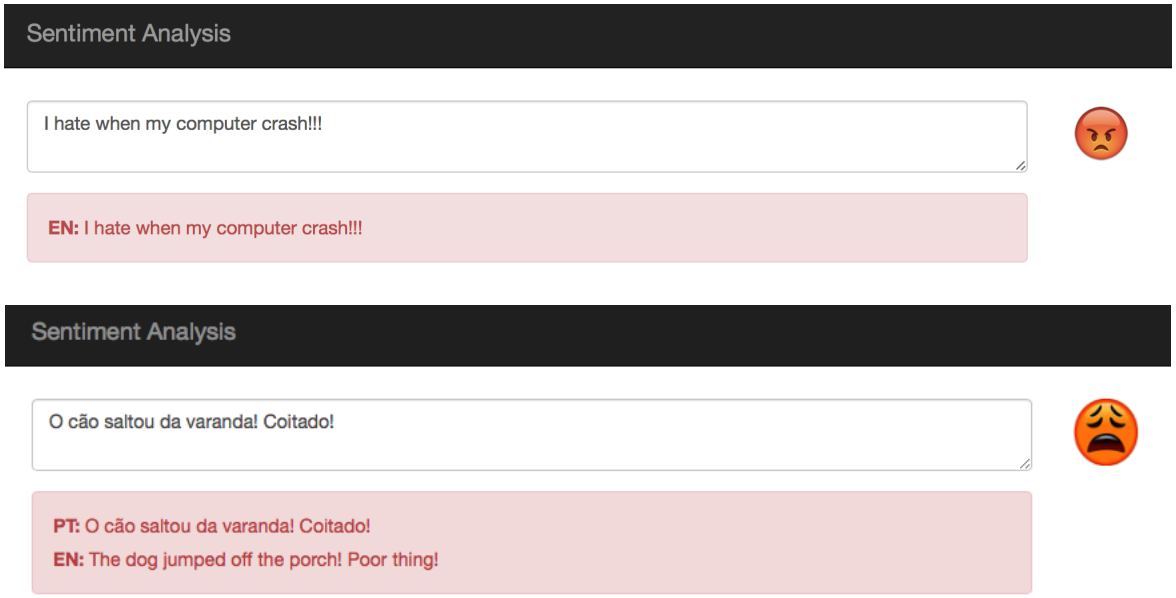


Figure A.2: Frase com sentimento fortemente negativo e negativo.

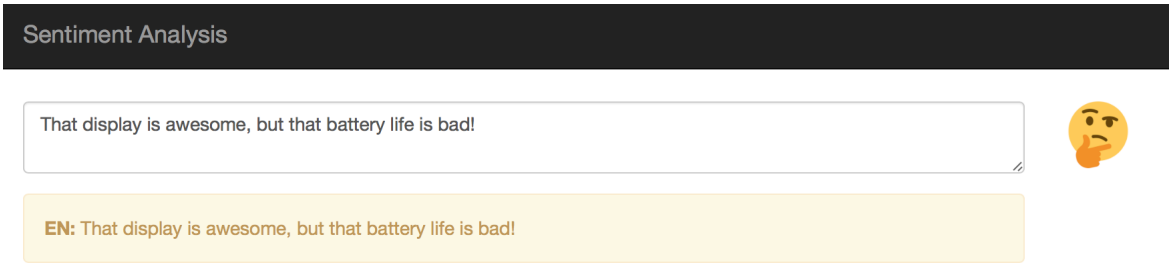


Figure A.3: Frase com sentimento de conflito ou neutro.