FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

FEUP

Extraction of opinionated profiles from comments on web news

Bruno Miguel Costa Duarte

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Eugénio da Costa Oliveira (Professor)

Co-Supervisor: Gustavo Alexandre Teixeira Laboreiro (PhD Student)

Co-Supervisor: Jorge Filipe Pinheiro Guerra de Ribeiro Teixeira (PhD Student)

July 25, 2012

Extraction of opinionated profiles from comments on web news

Bruno Miguel Costa Duarte

Mestrado Integrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

Chair: Doctor Maria Eduarda Silva Mendes Rodrigues
External Examiner: Doctor Daniel Castro Silva
Supervisor: Doctor Eugénio da Costa Oliveira

July 25, 2012

Abstract

One of the most important approaches to the extraction of information is text mining, a type of data mining that tries to derive high-quality information from text. The information is obtained through the identification of patterns and trends on the text. Opinion Mining is a specific case of text mining that deals with the extraction of opinions.

Massive spread of personal computers and Internet allowed online user-generated content (UGC) to grow exponentially in the past years. The huge amount of UGC available makes it a good source for opinion mining. The source of information for this work was extracted from Portuguese written comments on a website with news.

The first goal of this work is selecting the information to be used from the initial dataset. Then, finding entities on the obtained dataset. Entities are a list of mainly Portuguese predetermined figureheads.

The next goal is finding feelings about those entities, which includes part of speech and polarity tagging for determining what are the words expressing feelings. Following, the tagged sentences are classified using sentiment analysis methods by two different approaches: a rule based and a machine learning one, namely Naive Bayes. The last goal is to group opinions on entities by their feeling, showing entity profiles for each entity.

Different setups are used for sentiment analysis. In the rule based approach 2 different sentence structures are used, and for each one 500 random sentences containing that structure are analyzed. In the machine learning approach, other 500 sentences that respect the intended structure are randomly extracted.

As expected, results show that even though rule-based has good precision, the recall values are very limited. The machine learning approach handles better with recall, but precision suffers with it. Both approaches would also benefit from an increase in the number of tagged words with polarity.

In short, some notion of opinion is obtained by the classifier, but there are many different ways to improve final results. The most important include using different machine learning techniques to see what setup recognizes more opinions correctly, and expanding the resources used maintaining their overall quality.

Resumo

Uma da mais importantes abordagens à extração de informação é baseada em text mining, um tipo de data mining em que se tenta extrair informação de alta qualidade a partir de texto. A informação é obtida através da identificação de padrões e tendências no texto. Opinion mining é um sub-tipo de text mining lida com a extração de opiniões.

Devido à massificação dos computadores pessoais e da Internet, o conteúdo gerado por utilizadores tem vindo a crescer exponencialmente nos últimos anos de forma desorganizada. A grande quantidade de conteúdos deste género disponíveis torna-os uma boa fonte de informação para o Data Mining. O conjunto de dados utilizado é constituído por comentários em notícias web, extraídos a partir de um portal português.

O primeiro objetivo deste trabalho é seleccionar a informação a ser usada a partir do dataset inicial. De seguida, pretende-se encontrar as entidades que sejam referidas nos comentários, sendo as mesmas um conjunto pré determinado de figuras públicas maioritariamente portuguesas.

O objetivo seguinte é encontrar sentimentos acerca dessas entidades. Para tal, é efetuada a marcação da classe gramatical e polaridade das palavras para determinar quais podem expressar sentimentos. Depois, as frases marcadas são classificadas utilizando técnicas de análise de sentimentos de duas formas diferentes: uma baseada em regras e outra utilizando aprendizagem automática, nomeadamente um classificador Naive Bayes. O último objetivo é agrupar as opiniões acerca das entidades pelo sentimento expressado, resultando na criação de perfis opinativos para cada entidade.

Foram usadas várias configurações para a análise de sentimentos. Na abordagem baseada em regras são utilizadas 2 estruturas de frases, e para cada uma 500 frases que respeitem as condições pretendidas foram selecionadas aleatoriamente. Na abordagem baseada em aprendizagem automática, outras 500 frases que respeitam a estrutura delineada são extraídas aleatoriamente.

Tal como esperado, os resultados demonstram que a abordagem baseada em regras obtém precisão alta mas os valores de recuperação são muito baixos. Ao utilizar aprendizagem automática os valores de recuperação são melhores, mas a precisão baixa. Ambos os métodos beneficiariam da expansão do número de palavras com polaridade no léxico utilizado.

Em suma, o classificador consegue obter alguma noção de opinião, mas existem várias formas de melhorar os resultados finais. As mais relevantes incluem a utilização de diferentes técnicas de aprendizagem automática para determinar qual a que melhor reconhece opiniões, e expandir os recursos utilizados mantendo a sua qualidade global.

Acknowledgements

The work done and reported in this thesis would not be possible without the directly and indirectly help of some people.

I would like to thank my co-supervisors Gustavo Laboreiro and Jorge Teixeira for introducing me to the area, clarifying numerous initial doubts and also for the reviews on my work.

I would also like to thank my supervisor professor Eugénio Oliveira for giving me a general guidance through the process of writing this dissertation.

Last but not least, the cooperation with Sapo Labs was also very helpful, as it allowed me to obtain key resources for the accomplishment of this work.

Bruno Duarte

"Não confirmo nem desminto.
Antes pelo contrário, o que é importante é que interessa."

Anónimo

Contents

List of Figures

LIST OF FIGURES

List of Tables

Abbreviations

IE	Information Extraction
IR	Information Retrieval
KDD	Knowledge Discovery in Databases
ML	Machine Learning
NER	Named Entity Recognition
NLP	Natural Language Processing
OM	Opinion Mining
PMI	Point-wise Mutual Information
UGC	User Generated Content
XML	Extensible Markup Language

Chapter 1

Introduction

1.1 Context

In the scope of this work, opinionated profiles are the set of all opinions from different users, regarding a specific entity. The objective of creating such profiles is trying to understand what are the most expressed feelings and also sentiments on a given entity. There are 3 critical subjects with a great contribution for the creation of those profiles: opinion mining, sentiment analysis and the growth of user generated content.

Opinion mining is the area of research that attempts to make automatic systems to determine human opinion from text written in natural language [BX09]. Opinion mining is globally inserted in the Data Mining area, more specifically in Text Mining. As a type of data mining, opinion mining is a discipline that crosses knowledge discovery on databases with information and knowledge extraction. Opinion mining also inherits a computational linguistics component from text mining.

According to Bhuiyan et al. (2009) [BX09], computational linguistics is technically challenging because it requires natural language processing. But this is exactly what distinguishes Text Mining from other Data Mining areas. In Text Mining, the objective is to collect and process semantical information, and then retrieve knowledge from the earlier steps.

Sometimes also called Sentiment Analysis, the task of mining opinions is formally defined by Liu et al. [LH05] as follows: Given a set of evaluative text documents D that contain opinions (or sentiments) about an object, opinion mining aims to extract attributes and components of the object that have been commented on in each document D and to determine whether the comments are positive, negative or neutral.

Before Internet emerged there were very little written discussions on text opinion articles. Only a few number of people were selected by mass media to give their opinions and participate in discussions, with few different points of view. If people wanted to share opinions, they were typically restricted to her family and friends. But things were near a radical change.

The number of people using the web greatly increased: according to a study, the number of Internet users in Europe have grown more than 300% from 2000 to 2011, and more than 500% globally. This results in more than 2 thousand million people with access to the web, according to the information consulted on 9th of June, 2012 [Gro08].

Then the appearance of the "Web 2.0" allowed an easy platform for sharing opinions, which lead to an incredible growth in this type of content. One could post reviews of products, express views on almost anything in Internet forums and blogs, comment on websites and more recently do social networking. All this different types of expressing opinions can be grouped into what is called user generated content. Increasing blogs and social networking usage such as facebook [Smi08] (page accessed on 9th of June 2012), lead to an exponential growth in non-structured user generated content (UGC) in the web.

The trends introduced by this opinion oriented web also affected news related websites along last decade: they started as a simple repository and copy of the news published in the physical newspaper; the current paradigm implies a participation of the readers in those websites using tools to express their opinion. An example of these tools is the comments functionality that can be seen on many International (Figure 1.1) and National (Figure 1.2) websites.

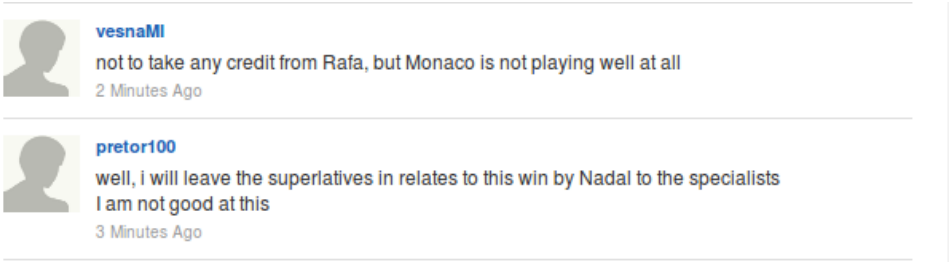


Figure 1.1: Screenshot of comments on sports website ESPN

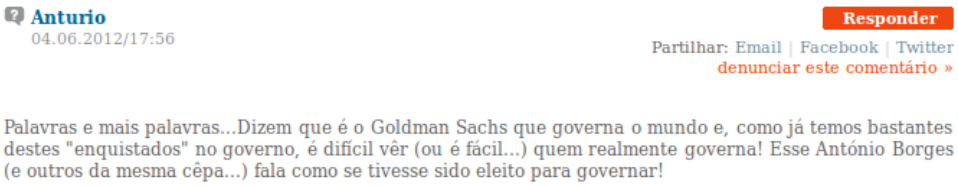


Figure 1.2: Screenshot of comments on Portuguese news website JN

Today, it is very common to see people sharing thoughts with others on web platforms such as social networks. The generated information has little external influences, since behind the computer people tend to express themselves without the social pressures of real life. A system that can extract knowledge from all these content has benefits comparing to traditional surveys or external consultants: consistency in results obtained throughout time and larger samples; there are more chances that the opinion reveals what people really think; the information and knowledge

extracted is up-to-date with world trends; easier applicability of extraction methods in different languages.

Realizing the potential value of non-structured information, more and more efforts are being made in order to extract valuable information from it. Many diverse areas are trying to obtain information from UGC: banking information for cross selling (eg., know what other offers from the bank might interest the client), acquire information about customer loyalty, health care, forecasts (e.g., determining the popularity of a political candidate), understanding the profile of the visitors which leave comments in a given website or even selling the acquired knowledge to external enterprises.

1.2 Motivation and goals

"Text is the most significant repository of human knowledge" [LP01]. With the amount of increasing unprocessed text resulting of users' activity on online platforms, such as news comments and social networking, the amount of user-generated text around the web is today significantly bigger.

The motivation for this work is essentially to create opinionated profiles with feelings and opinions of users regarding an entity - generally a famous person such as politics or athletes. The creation of profiles allows the automatic analysis of the more common opinions of users about an entity.

During the realization of this work, there are a set of goals to be achieved:

- selection of information to be used from the initial dataset. In this phase occurs selection of the useful data, encoding problems are dealt with as well as spam behaviors;
- find entities referred in comments;
- find feelings about those entities, by part of speech tagging and extraction of polarity relevant sentences;
- sentiment analysis using features previously extracted, by a rule based and a machine learning approach (Naive Bayes);
- discover the most common opinions linked to an entity, by grouping opinions on entities through the expressed feeling. In the end, entity profiles for each entity are created.

1.3 Document Structure

Following the introduction section, Chapter 2 presents a literature review about Opinion Mining and the most important areas intersecting it, focusing on what was helpful for the accomplish of this work. Chapter 3 presents resources used through the semester. On Chapter 4 the implementation steps are explained. Then, Chapter 5 contains results and their interpretation. Finally, the conclusions and goals of this dissertation are presented, and future directions to extend the project are mentioned.

Chapter 2

Literature Review

An opinion is formally defined as a belief held with confidence but not substantiated by positive knowledge or proof. Other definitions suggest a judgment or estimation of the merit of a person or thing. An opinion is a subjective belief that results from feelings or from the interpretation of facts. An opinion may be supported by an argument, and rarely changes without new arguments being presented [Dam08].

In the scope of this dissertation, opinion is a narrowed concept of the broader definition. The term opinion is the result of people perspectives, understandings, particular feelings, beliefs, and desires. It usually refers to unsubstantiated information, in contrast to knowledge and fact-based beliefs.

2.1 Text Mining overview

Text Mining has been gaining more value since companies started to understand that the majority of their information is contained in text documents. In fact, according to a study close to 80% of a company information is contained in text documents [Tan99]. As mentioned in the previous chapter, Text Mining is a sub-genre of Data Mining that aims to discover patterns from unstructured or semi-structured text, extracting useful knowledge from it. Text Mining relates Knowledge Discovery on Databases (KDD) and Information Extraction (IE). KDD consists on the application of statistical and machine learning techniques to discover relationships between data and IE is the process of locating pieces of data in natural-language documents, extracting structured information from unstructured or semi-structured text.

Data Mining assumes that information is already in a desirable form for extraction (e.g., relational database), but in Text Mining the information is in the form of natural-language freely written. This characteristic gives Text Mining an uncertain nature.

In order to overtake this issue, a previous knowledge about the dataset is recommend. This knowledge is normally acquired by analyzing small samples from the whole dataset. The referred analysis gives some insight on the general problems that will need to be faced when working with that dataset.

Examples of common problems occurring in user-generated content are grammar errors and unintentional or deliberate typos. In both cases they require a preprocessing step where text is normalized; this way, incorrect words are substituted by the correct form.

With some insight in the nature of the dataset, it is possible to select suitable approaches and methods for accomplishing objectives. This is a very important procedure, because there is no best approach for all Text Mining cases. Therefore, the best approach differs for each case, depending on numerous factors including the nature of the dataset and the goals intended for the project.

For instance, a project with the objective of maximizing precision (percentage of good instances retrieved in total retrieved instances) is very different from a project where recall (percentage of good instances retrieved in the set of all good instances) is the most important. Thus, the methods used in each case are different concerning the objectives and also the nature of text.

2.2 Opinion Mining background

Text information can be separated in two main categories, facts and opinions. Facts are objective statements about entities or events, and are the main purpose of information retrieval and web search engines. They were the main focus of text mining for several years, shown by the competition between Google, Yahoo and Microsoft search engines, among others. On the other side, the subject of this work are opinions; they are subjective to each one's feelings and perceptions about a subject. Global interest in opinions only appeared years later comparing to interest in facts, but the research in this area is now growing very fast [BX09].

The ever growing quantity of data collected for Opinion Mining is powered by people needs in expressing personal opinions. Also, when they need to make a decision or simply understanding what other people think about a subject, they like to listen to other opinions. Joining these two factors is the key to understand such up-growth in user generated content, and consequently the great development verified in sentiment analysis in last years, as stated by PBT, a Business Intelligence and Health care solutions provider [Sch12].

Nowadays it is usual to find some space in websites where users can leave their opinion, normally through a comments functionality. Topics as Sports or Politics generally are the most participative, but not always the most informative. This happens due to many factors, including off-topic conversation that makes comments becoming bigger in size but less focused in the original topic, and direct attacks between supporters of different ideologies, among others.

These factors makes it difficult for a new user to enter the conversation and understand the problematic discussed. Finding relevant sources and extracting appropriate sentences is therefore a clear need for these users. An opinion mining system that can identify and summarize the most relevant ideas in a text, or a succession of texts, is essential for people to gather information and acquire a point of view as little biased as possible.

Sentiment analysis can be effective for both enterprises and single individuals. Imagine a person that needs to buy a product for a defined goal and needs to know what is the best suitable option for her; or another person that needs to decide in what politician to vote for the elections;

or even an enterprise that needs to understand the real needs of their clients in order to relocate their efforts. All of these are cases that can benefit from having a brief summary of what is being said about what they need, instead of reading all the existent information. In real life situations, there is no time for that.

Automatically mining for the general sentiment expressed on a text is then a viable solution. The initial idea behind Opinion Mining is finding the general sentiment present in the text: the simpler approach is based on grouping texts in a positive or negative sentiment.

A better approach includes another class to qualify a text as neutral. Using neutral sentiment can improve accuracy of results and allows extraction of other relevant information; when a sentence is classified as positive, that is not necessarily true, but it is more probable that the sentence is not negative. The same happens for examples classified as having negative sentiment [Kop05]. Most of the approaches to sentiment mining uses a list of bearing words (also called opinion lexicon) for the purpose. These words express desirable (e.g., great, amazing, etc.) or undesirable (e.g., bad, poor, etc) states [DL08] [LZ08].

2.3 Opinion Mining application domains

This section shows some common uses for Opinion Mining. It is based in a study made by Binali et al. [BPW09] and presents some specific cases that provide a good overview about the wide range of problems and possible solutions using Opinion Mining.

2.3.1 Shopping

Perhaps the most popular use of opinion mining is decision support for consumers. Consumers are actively involved in comparison shopping over the Internet. Popular websites like amazon [1] allow customers to express their opinions on their websites.

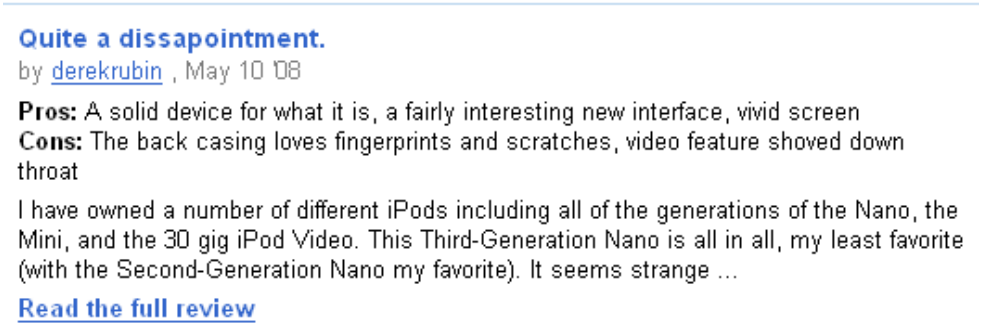


Figure 2.1: Example of online shopping opinion

Customers can easily view the opinions for products and identify how features from different products compare with each other. In some cases, after an opinion has been mined and processed,

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

knowledge is presented to the user graphically for easy comparison of product features. Consider the comments on an electronic product from an online shop.

> "I needed a high-powered laptop for my business needs. Dell offered a variety of products that met my requirements. Their product information was concisely and completely explained, which made my selection process very easy. Their website was easy to use and attractively presented. The merchandise was delivered on time, as per their promise. Shipping charges were very reasonably priced. I was able to take care of all my needs online, thereby not having to use their telephone support service. Their online help and support was excellent. I would recommend Dell Small Business to anyone[2]."

It is easy for humans to notice that the opinion is about Dell Small Business and the various features being talked about are delivery time, shipping charges, support and web site navigation. The objective of applying Opinion Mining techniques is to extract that type of information automatically from numerous similar texts.

2.3.2 Entertainment

Movie goers and home TV viewers can quickly access the opinion on recent releases and popular movies and programs. Currently, there is the internet movie database (IMDB) which provides online reviews for movies as well as TV programs. This acts as a guide for people who are unsure about which movies to watch. Below there is an extract sample from the IMDB.

> "Christopher Nolan's second bundle of joy "The Dark Knight" EXCEEDED all of my expectations!!! I can HONESTLY tell you that: as good as Jack Nicholson was in Batman'89 he is CHILD'S PLAY compared to this Joker. He is sadistic, psychotic, and downright SCARIER and PSYCHOLOGICALLY disturbing than the previous incarnation of The Clown Prince of Crime and Ledger gives it his all to do him justice. The action is great, and the plot is deeper and engrossing[3]."

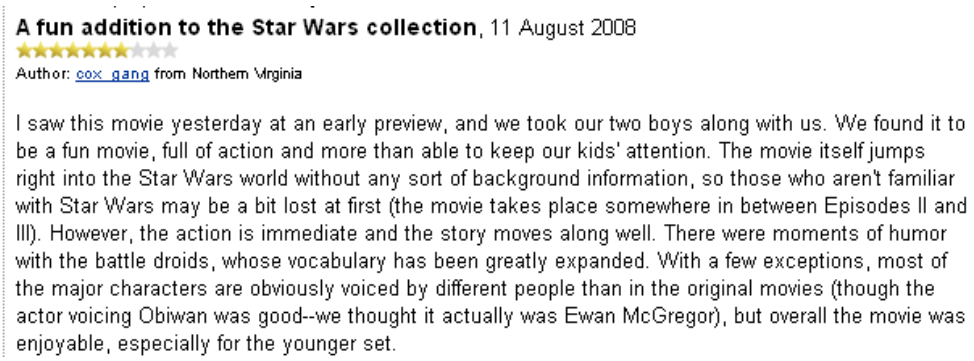
From the previous opinion, capital letters and exclamation signs are being used to emphasize emotions. Furthermore, the first comment is positive and refers directly to the movie, "The Dark Knight". However, subsequent statements refer to the actors and it is their attributes that are being mentioned. The last statement mentions the film once again. This kind of opinion, which revolves between the actors and the movie, is relatively simple for a human reader to understand but not so for a machine.

Therefore, this presents some complexity to machine learning. It is evident that two objects are being described, the movie and actors. Although words with a negative connotation (sadistic, psychotic, and disturbing) are being used to express positive aspects of the movie, it does not mean

[2] retrieved in 2009 from http://www.amazon.com/

[3] http://www.imdb.com/title/tt0468569/, viewed at 10th June 2012

that the film is not highly recommended but rather, just an illustration of the complexity that exists for machine learning. Moreover, most opinions about movies are expressed in this way.



A fun addition to the Star Wars collection, 11 August 2008

Author: cox_gang from Northern Virginia

I saw this movie yesterday at an early preview, and we took our two boys along with us. We found it to be a fun movie, full of action and more than able to keep our kids' attention. The movie itself jumps right into the Star Wars world without any sort of background information, so those who aren't familiar with Star Wars may be a bit lost at first (the movie takes place somewhere in between Episodes II and III). However, the action is immediate and the story moves along well. There were moments of humor with the battle droids, whose vocabulary has been greatly expanded. With a few exceptions, most of the major characters are obviously voiced by different people than in the original movies (though the actor voicing Obiwan was good--we thought it actually was Ewan McGregor), but overall the movie was enjoyable, especially for the younger set.

Figure 2.2: Example of online movie review

2.3.3 Government

Governments can mine the prevailing opinions on public policy. Election candidates can become more knowledgeable about specifics of the opinion poll. This knowledge can assist politicians to identify where their strengths and weaknesses lie according to their electorate. Consider the following political opinions that have been expressed. “Expect more inflation. More unemployment. Really, we need some better selection process. Who chooses these people? They make history by raising rates for the first time in the lead up to an election. I have no confidence in the published figures.”[4] A quick glance at these terms indicates a sense of dissatisfaction among the electorate.

Furthermore, key areas of concern are addressed in terms of what is lacking and what the expectations are. Issues that deal with public policy normally categorize voters into one of three groups, for, against or neutral. A good example is the statement, “I think this all seems extremely harsh. Boredom, if anything, is a sign of intelligence.”[4] A statement of this kind makes it clear that the opinion is for the motion. The advantage of opinion mining over traditional opinion polls like telephone polls is that it can be determined why electorates are for or against a proposal. Most web sites, particularly those whose fundamental objective is to provide news, have a facility for web users to express their opinions on their websites.

2.3.4 Research and Development

Product reviews can be used by manufacturing companies to improve features and provide a platform for innovation. Web based applications could offer platforms for customers to design products and submit the designs to the manufacturing companies. An approach of this nature could significantly assist in establishing features that are liked by customers. Consider the following

[4] retrieved in 2008 from http://www.theaustralian.com.au/news/opinion

review for an electronic product, "The click wheel is HORRIBLE and completely lacks response and sensitivity."[5]

This is a negative opinion being expressed about the click wheel. The use of upper caps signifies to the reader the extent of disappoint. If opinion mining is able to detect emotions of this kind being expressed in evaluative text, it will prove to be very beneficial. This will act as an indicator on how the product has been received by a consumer. However, after expressing negative opinions on the product features, a statement such as "although I am really disappointed this is probably still the best high capacity music player on the market"[5]. This positive statement indicates to the R & D department and marketing departments that the music player is still the best in the market and it is the high capacity which is favored by customers.

Figure 2.3: Example of online movie review

2.3.5 Marketing

Positive opinions about a subject can really improve products rating. An example of that is the following recommendation for a tourism resort, "It is a land of contrasts and majesty, Africa at its most wild and unexplored" [6].

Companies can now make savings on marketing expenses by requesting for reviews on their websites and specialized review websites. This eliminates the need for business consultants to conduct surveys as companies can now have all the data they need online. The advent of the Internet has brought along with it new ways of marketing. One great example of it is Viral Marketing.

Viral marketing is the use of social networks to spread product and service information. With the advent of the Internet, social networks such as MySpace [7] or Facebook [8] are offering a new platform for information exchange. Family and friends can now recommend products/services to each other or seek more knowledge about a product or service before committing themselves. It is analogous to the traditional word of mouth marketing of products and services. To encourage

[5] retrieved in 2008 from http://tanzaniatouristboard.com/
[6] retrieved in 2008 from http://www.theaustralian.com.au/news/opinion
[7] http://www.myspace.com/
[8] www.facebook.com

postings and recommendations among peers, marketers normally offer incentives like discounts for recommendations that turn into purchases [LA07].

2.3.6 Education

In e-learning systems, users opinions can be used to evaluate academic institutions and academics. Academics can know the sentiment on courses based on sentiment analysis of opinions expressed by students. This can help to improve service delivery and bolster marketing campaigns. Unit coordinators can know what students think about their team members and tutors by requesting them to provide online reviews as a part of course requirement.

For instance, Faculdade de Engenharia da Universidade do Porto (FEUP) has a system that sends surveys to each student, asking for a classification on professors, what was done right and what was done wrong in each course, throughout the semester. But the surveys are very generic and most of the times they do not have options to express the real feel about the course.

This is an excellent case of possible application for Opinion Mining. A good complementary functionality might be using Opinion Mining to discover the most common ideas associated with each course or professor, and use that information to improve the course.

There are many different possibilities for opinions with academic purpose. An example of it is presented below:

> "My research is improving my analytical, problem solving skills and ability to plan my own work. The feedback from the supervisor is valuable. The computing facilities are excellent. However, the monthly down load quota is too low for conducting research without being exceeded." [9]

It is possible to understand that overall opinion is for research (object) and it is positive. The features to extract an opinion on would be the supervisor, computing facilities and download quota. Record data as the previous example and make it readily available enables a more humane comparison of courses performance.

2.4 Sentiment Analysis

In a 2006 article called "Blog Mining through Opinionated Words" Attardi and Simi [AS06] stated that Intent Mining goal is "to assess the attitude of the document author with respect to a given subject" and "Opinion mining is a kind of intent mining where the attitude is a positive or negative opinion". By being a type of Intent Mining, Opinion Mining is a way of establishing connections between a subject and an opinion about it. Sentiment Analysis tries do classify two different problem structures: direct opinion mining problems, as in "This chair is great", and comparative opinion mining problems, as in "This chair is better than the old one".

In opinion mining, the typical approach comprises the use of three machine learning techniques for training and testing the dataset:

[9] retrieved in 2008 from http://planning.curtin.edu.au/mir/cass.cfm

- Naive Bayes [FB06]
- Maximum Entropy [Soo00]
- Support Vector Machines [BGV92]

Naive Bayes based classifiers consist in a simple application of Bayesian probabilities. Doing so, these classifiers assume independence between all features existing in the dataset. Despite starting from a wrong assumption, this type of classifier has proven working well on real data. This apparently strange fact was explained by Zang [Zha04], where he held that the different dependencies tend to cancel each other, making the dependence violation lose influence in the classification.

Maximum Entropy (ME) classifiers consist in the application of greedy algorithms that provide the least biased estimation. They are used alternatively to Naive Bayes for not assuming conditional independence between features, even though the learning process is much more time consuming.

Support Vector Machine (SVM) is a very different method as it is not based on statistic methods. The idea behind SVM is to define a hyperplan with the biggest possible margin separating data in two classes. This margin is extended until it reaches the closest points, which are the support vectors. So, support vectors implicitly define hyperplan margins and help the classification of new data. When the training data is not linearly separable, that is, it is not possible to define a hyperplan to separate data, there is a technique known as the "kernel trick" that maps the information in a high-dimension feature space. With the new dimension, the hyperplane can then be defined and classify the information.

There have been a great number of researchers working in the area of Opinion Mining, approaching and solving problems in many different ways. A widely accepted way of separating studies in the area was suggested by Bhuiyan et al. [BX09]. A detailed view on how different works in the area relate can be seen in Figure 2.4. In their research, two main research directions are identified: Sentiment Classification and feature-based Opinion Mining.

2.4.1 Sentiment Classification phases

Sentiment classification is concerned with the overall sentiment of a sentence or a document towards a subject. To do so, three stages need to be accomplished. First, an entity is extracted from the document. Entity extraction is very important to understand for whom is the opinion targeted, and some literature on the unsupervised extraction of web entities is available [ECD+05].

The next phase is entity sentiment collection, where the overall sentiment being expressed on the entity is acquired. Examples of that are good/bad, excellent/boring or smart/dumb as the prevailing sentiment of a document about an entity. For instance, Turney [Tur02] presented a paradigm which provides a basis for extracting the opinion about an item.

The terms used in the review are assumed to be subjective and can be divided into one of three groups, positive, negative or neutral. As mentioned in the previous section [Kop05], most

Figure 2.4: Classification of Opinion Mining Research

researchers started focusing on weighing the subjectivity of positive and negative terms to extract the review sentiments while ignoring neutral ones. However, after studies showing that neutral terms can improve the accuracy of results were made, neutral subjectivity started to be more used.

The last phase is called Entity Comparison. This phase relies on comparing opinions of an Entity A and other entities. It is specially used for reviews on items: if a person needs to decide what item to buy in little time, a good decision factor might be comparing the opinions of two items and choose the one with best reviews.

But this methodology can also be applied to persons instead of items. For example, to know who is the favorite politic for an election, comparing opinions about them is a good solution. Wang [WA08] have developed a novel way of graphically depicting entity comparison.

2.4.2 Sentiment Classification research

Research on opinion mining basically started with identifying opinions on sentiment bearing words, (e.g., great, amazing, wonderful, bad, poor and so forth). Many researchers have worked on mining such words and identifying their semantic orientations or polarity such as positive, negative or neutral [GZ06] [LWWH06].

Hatzivassiloglou and McKeown [HM97] identified several linguistic rules that can be exploited to identify opinion words and their orientations from a large corpus. This method has been applied, extended and improved by other researchers [DL08] [KN06] [PE05].

Table 2.1: Sentiment Classification results by Pang et al.

Features	**number of features**	**frequency or presence**	**NB**	**ME**	**SVM**
unigrams	16165	freq.	78.7	N/A	72.8
unigrams	N/A	pres.	81.0	80.4	82.9
unigrams+bigrams	32330	pres.	80.6	80.8	82.7
bigrams	16165	pres.	77.3	77.4	77.1
unigrams+POS	16695	pres.	81.5	80.4	81.9
adjectives	2633	pres.	77.0	77.7	75.1
top 2633 unigrams	2633	pres.	80.3	81.0	81.4
unigrams+position	22430	pres.	81.0	80.1	81.6

The next major development is sentiment classification of product reviews at the document level [DLP03] [PLV02] [Tur02]. In Dave et al. (2003) [DLP03] sentiment classifiers are built from some training corpus. The objective of this task is to classify each review document as expressing a positive or a negative sentiment about an object (e.g., a movie, a camera, a book, a laptop computer or even a car).

In 2002, Pang et al. [PLV02] compared the three earlier referred methods (Naive Bayes, Maximum Entropy and SVM) between them and against human produced baselines. The objective was the classification of sentiment in a movies dataset. The measure used for the three-fold cross-validation was the precision of the classification and many features were tested, as it can be viewed in the table 2.1.

There were different training scenarios in this work. In the case of features the author used unigrams, bigrams, position and adjectives. Unigrams and bigrams are the definition of one or two words as the unit of text to be parsed, respectively. Position is the part of the phrase where the word is (i.e. first, middle or last quarter) and adjectives consists of only look for adjectives in the text as they have a great deal of information regarding a document's sentiment. Choosing between presence count or frequency count of words was easier because the last one has shown better results in similar setups.

In a quick analysis of the obtained results, it is possible to conclude that the best setup was achieved using both unigrams and bigrams simultaneously, followed by unigrams plus position. However, back in the days this work was done there were several limitations in processing power. For this reason, the number of features tested was harshly decreased comparing with the original data.

With the intent of overtake these limitations, Vachaspati and Wu [VW12] replicated Pang's work as closely as they could, but this time extending the work by exploring an additional dataset, additional preprocessing techniques, and combining classifiers. They also tested how well classifiers trained on Pang's dataset extended to reviews in another domain. Although Pang limited many of his tests to use only the 16165 most common n-grams, advanced processors have lifted this computational constraint, and so they additionally tested on all n-grams. Results of Vachaspati and Wu's work can be seen in table 2.2.

Table 2.2: Sentiment Classification results by Vachaspati et al.

Test configurations				Naive Bayes			MaxEnt			SVM		
Domain	Features	# of features	Frequency	+	-	±	+	-	±	+	-	±
No-negation	Unigrams	16165	Frequency	0.94	0.62	0.78	-	-	-	0.82	0.82	0.82
No-negation	Unigrams	16165	Presence	0.87	0.72	0.82	0.85	0.87	0.86	0.85	0.84	0.84
No-negation	Bigrams	16165	Frequency	0.92	0.64	0.78	-	-	-	0.77	0.81	0.79
No-negation	Bigrams	16165	Presence	0.89	0.73	0.81	0.79	0.82	0.81	0.8	0.81	0.8
adjectives	Unigrams	16165	Frequency	0.95	0.52	0.73	-	-	-	0.75	0.77	0.76
default	Bigrams	2633	Frequency	0.91	0.46	0.69	-	-	-	0.74	0.75	0.75
default	Bigrams	16165	Frequency	0.92	0.64	0.78	-	-	-	0.78	0.79	0.78
default	Unigrams	2633	Frequency	0.96	0.5	0.74	-	-	-	0.81	0.79	0.8
default	Unigrams	16165	Frequency	0.93	0.59	0.76	-	-	-	0.82	0.81	0.82
default	Unigrams	maximum	Frequency	0.95	0.49	0.72	-	-	-	0.82	0.81	0.82
partofspeech	Bigrams	16165	Frequency	0.96	0.47	0.71	-	-	-	0.82	0.82	0.82
partofspeech	Unigrams	16165	Frequency	0.96	0.54	0.75	-	-	-	0.82	0.81	0.81
position	Bigrams	16165	Frequency	0.96	0.49	0.73	-	-	-	0.77	0.78	0.78
position	Unigrams	16165	Frequency	0.93	0.58	0.76	-	-	-	0.81	0.82	0.82
verbs	Unigrams	maximum	Frequency	0.8	0.55	0.67	-	-	-	0.61	0.65	0.63
adjectives	Unigrams	16165	Presence	0.93	0.59	0.76	0.79	0.77	0.78	0.75	0.73	0.74
default	Bigrams	2633	Presence	0.86	0.64	0.75	0.75	0.75	0.75	0.73	0.75	0.74
default	Bigrams	16165	Presence	0.89	0.74	0.81	0.81	0.82	0.81	0.78	0.79	0.78
default	Unigrams	2633	Presence	0.84	0.8	0.82	0.84	0.82	0.83	0.78	0.82	0.8
default	Unigrams	16165	Presence	0.87	0.77	0.82	0.84	0.85	0.85	0.83	0.82	0.83
default	Unigrams	maximum	Presence	0.91	0.7	0.81	0.84	0.86	0.85	0.83	0.85	0.84
partofspeech	Bigrams	16165	Presence	0.89	0.73	0.81	0.84	0.84	0.84	0.79	0.82	0.8
partofspeech	Unigrams	16165	Presence	0.86	0.76	0.81	0.85	0.85	0.85	0.84	0.83	0.84
position	Bigrams	16165	Presence	0.87	0.66	0.76	0.82	0.83	0.82	0.73	0.76	0.74
position	Unigrams	16165	Presence	0.86	0.78	0.82	0.84	0.85	0.85	0.8	0.8	0.8
verbs	Unigrams	maximum	Presence	0.8	0.54	0.67	0.65	0.65	0.65	0.64	0.63	0.635
adjectives	Unigrams	16165	TF-IDF	0.82	0.6	0.71	-	-	-	0.79	0.76	0.77
default	Bigrams	2633	TF-IDF	0.92	0.46	0.69	-	-	-	0.76	0.71	0.74
default	Bigrams	16165	TF-IDF	0.9	0.68	0.79	-	-	-	0.83	0.74	0.79
default	Unigrams	2633	TF-IDF	0.85	0.52	0.74	-	-	-	0.81	0.79	0.8
default	Unigrams	16165	TF-IDF	0.88	0.68	0.78	-	-	-	0.83	0.77	0.8
default	Unigrams	maximum	TF-IDF	0.86	0.65	0.76	-	-	-	0.83	0.78	0.81
partofspeech	Bigrams	16165	TF-IDF	0.89	0.67	0.78	-	-	-	0.79	0.74	0.76
partofspeech	Unigrams	16165	TF-IDF	0.89	0.63	0.76	-	-	-	0.81	0.78	0.79
position	Bigrams	16165	TF-IDF	0.89	0.59	0.74	-	-	-	0.79	0.69	0.74
position	Unigrams	16165	TF-IDF	0.91	0.61	0.76	-	-	-	0.81	0.71	0.76
verbs	Unigrams	maximum	TF-IDF	0.64	0.57	0.6	-	-	-	0.62	0.66	0.64

After evaluating the results presented in table 2.2, it is possible to observe that the utilization of the maximum number features did help performance. In fact, when using the exact same setup and the maximum number of features, the performance is slightly better than with the regular setup.

One of the differences between this work and the previous one is that TF-IDF (term frequency–inverse document frequency) measure was added. This numeric value reflects how important a word is to each text in the dataset. The term frequency represents the number of times a given word appears in a given text, and inverse document frequency states for a logarithmic ratio between the total number of texts by the number of texts that contains the word.

The most significant variation introduced in this work was the usage of majority voting. In

this scheme, the results obtained by each classifier are combined giving the same weight for all of them. These can eliminate weaknesses existing in a single classifier but also eliminate strengths that only one classifier have. Results have shown that the combination of the three classifiers (Bayes, ME and SVM) provided a three to four percent boost over results of the best classifier alone.

The final results obtained by this work are fairly positive, as it has achieved, in some specific setups, values around 95% and 96% of precision. Also, when using an harder alternative dataset to train data (Yelp [Yel]) the results were positive, despite being lower than with the default movies dataset.

Some researchers analyze sentiment at document level, while others do it at the sentence level. Sentence level is done by classifying each sentence as a subjective or objective sentence and/or as expressing a positive or negative opinion [KH04] [WR05] [WWH04].

Sentence level subjectivity classification is studied in Hatzivassiloglou and Wiebe (2000) [HW00], which determines whether a sentence is a subjective sentence but may not express a positive or negative opinion or a factual one. Wiebe and Riloff (2005) [WR05] distinguish subjective sentences from objective ones.

Kim and Hovy (2004) [KH04] propose a sentiment classifier for English words and sentences, which utilizes a thesauri. However, template-based approach needs a professionally annotated corpus for learning; words in thesauri are not always consistent. Like the document-level classification, the sentence-level sentiment classification does not consider object features that have been commented on in a sentence.

Abbasi et al. (2008) [ACS08] proposed the use of sentiment analysis methodologies for classification of the Web forum opinions in multiple languages (Fig. 2.5). The design has two major steps: extract an initial set of features, and then perform feature selection. These steps are used to carry out sentiment classification of forum messages. The experiment produces a fantastic result on the benchmark movie review dataset. Their method focuses on document level classification of sentiment only.

Gamon at al. (2005) [GACOR05] presented a prototype system named Pulse, for mining topics and sentiment orientation jointly from customer feedback. However, this technique is limited to domain of products and highly dependent on the training dataset, so is not generally applicable to summarize opinions about an arbitrary topic. Most sentence level and even document level classification methods are based on identification of opinion words or phrases. There are basically two types of approaches:

- Corpus-based approaches
- Dictionary-based approaches

Corpus-based approaches find co-occurrence patterns of words to determine the sentiments of words or phrases [HW00] [Tur02]. Dictionary-based approaches use synonyms and antonyms in WordNet to determine word sentiments based on a set of seed opinion words [Fel98].

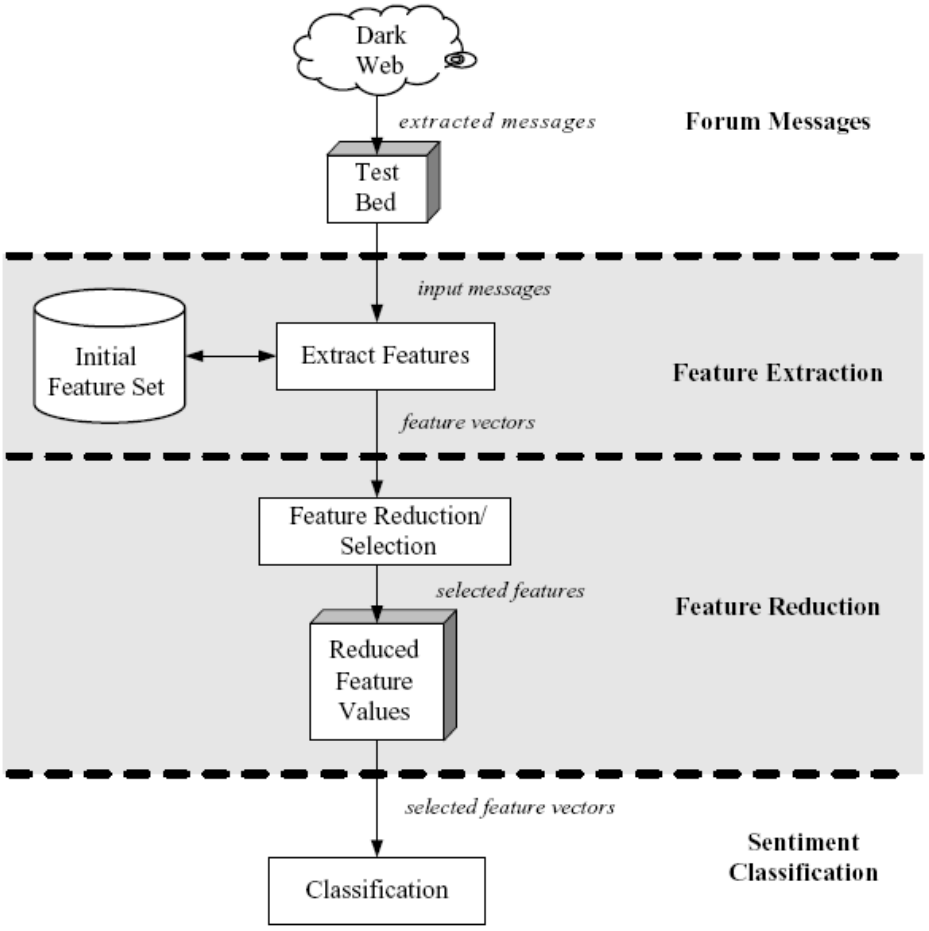


Figure 2.5: Sentiment classification system design

In Hu et al.(2004) [HL04] and Kim and Hovy (2004) [KH04], a bootstrapping approach is proposed, which uses a small set of given seed opinion words to find their synonyms and antonyms in WordNet [10] to predict the semantic orientation of adjectives. In WordNet, adjectives are organized into bipolar clusters and share the same orientation of their synonyms and opposite orientation of their antonyms.

To assign orientation of an adjective, the synonym set of the given adjective and the antonym set are searched. If a synonym/antonym has known orientation, then the orientation of the given adjective could be set correspondingly. As the synonym set of an adjective always contains a sense that links it to the head set, the search range is rather large. Given enough seed adjectives with known orientations, the orientations of all the adjective words can be predicted [11].

Yu et al. (2008) [YMTR08]proposed a method for combining How-Net and sentiment classifier. They divide the sentiment text features into characteristic words and phrases extracted from the training data. Then they compute semantic similarity of characteristic words, phrases with

[10] wordnet.princeton.edu
[11] Lee2008

tagged words in How-Net, and adopt the positive or negative terms as features of sentiment classifier. Negative rules for negation sentences are also added to sentiment classifier. If a word is matched, the whole meaning of the sentence is changed contrarily. However, the performance of their proposed method is not that satisfactory according to their experiment result.

Dey and Haque (2009) [DH09] proposed a hybrid approach while focusing on opinion extraction from noisy text data. They have argued that most of the existing Natural Language Processing (NLP) techniques assume that the data is clean and correct. But generally opinions expressed in the online environment as blog comments or written reviews are full of spelling mistakes and grammatical errors due to "noisy text".

Their proposed system uses a plugged in domain ontology to extract opinions from pre-defined websites which allows opinions to view at multiple levels of granularity based on the requirements. They proposed a text pre-processing mechanism which exploits domain knowledge to clean the text. Those clean texts are then processed by NLP tools. But the process is iterative and difficult to implement.

2.4.3 Feature based Opinion Mining

The model of Feature-based opinion mining and summarization is proposed by many researchers [HL04] [LH05] [PE05] while others [BM08] propose feature driven opinion summarization method. They emphasize on the term d̈riven̈to describe the concept-to-detail approach. In Feature-based opinion mining, features broadly mean product features or attributes and functions. The main tasks in this technique are:

- Identifying product features that have been commented on
- Decide whether the comments are positive or negative
- Summarizing the discovered information

Feature-based Opinion Mining identifies different features of the subject, and then tries to obtain the sentiment regarding each one of those features. The basic motivation of feature-based approach is that a negative customer opinion on a product does not necessarily mean this customer dislikes every aspect of the product, and vice versa.

This approach models a product consisting on a number of sub-components. Each product is associated with a set of attributes that can be evaluated through opinion expressions. Note that feature represents both components and attributes. Consider a digital camera, which has several features: picture quality, battery life, zoom, size, weight and so on. For instance, a camera with poorer picture quality may have a very long battery life and light weight [BPW09]. A situation like this would result in different sentiment for each feature.

Classifying evaluative texts at the document level or the sentence level does not tell what the opinion holder likes and dislikes. A positive document on an object does not mean that the opinion holder has positive opinions on all aspects or features of the object. Likewise, a negative document does not mean that the opinion holder dislikes everything about the object.

In an evaluative document such as a customer review of a product, the opinion holder typically writes both positive and negative aspects of the object, although the general sentiment on the object may be positive or negative. To obtain such detailed aspects, Feature-based opinion mining has been proposed [HL04] [PE05] to summarize the overall opinion.

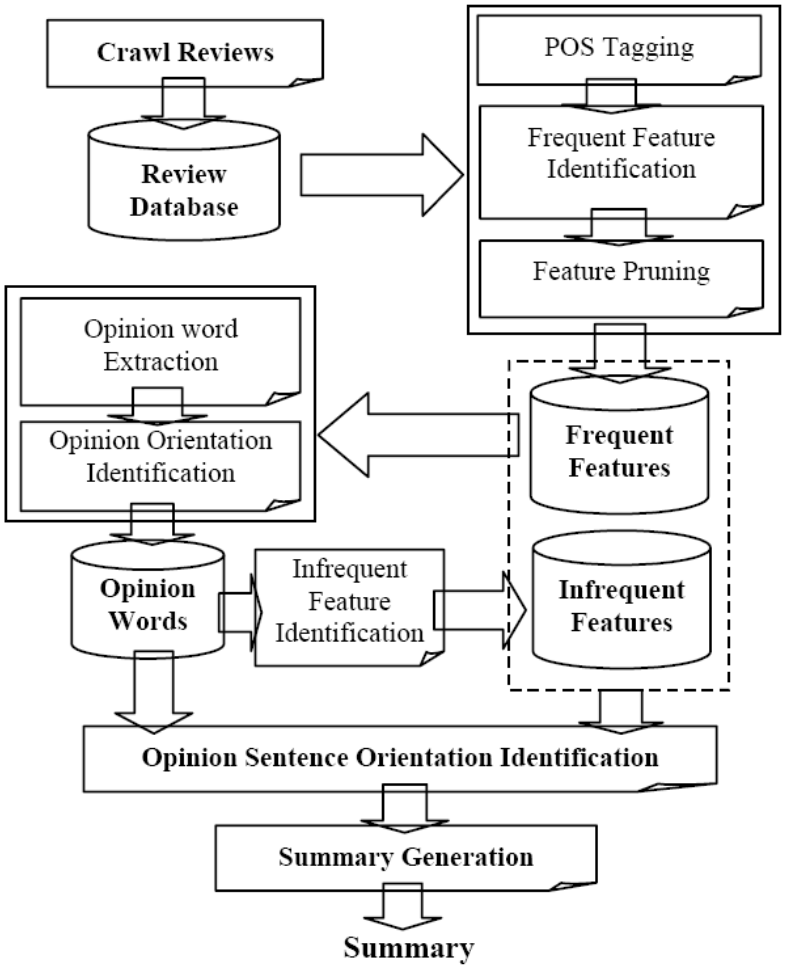


Figure 2.6: Feature-based Opinion Summarization

The first stage is called Feature Extraction. Instead of just extracting semantic orientated opinion, feature extraction looks for the features whose opinion can be used to extract the entity sentiment. An example of definition and modeling of this kind of work was made by Hu et al. [HL04].

The following phase is named Feature Sentiment. Feature Sentiment is the opinion that is expressed for an item based on its features. After the features have been identified, a feature sentiment can be expressed for each feature, which tells us something about the weaker and finer points of item's features. For instance: long battery life, portable size, beautiful colors, excellent cast, poor actors.

A point of interest here is the case of the movie review. Researches show that it is hard for current algorithms to distinguish between opinions of the movie in general and those regarding the actors. In some cases, a movie may get a positive review because a popular actor is involved but the movie itself may be uninteresting. With scenarios of this nature, it becomes necessary to differentiate between reviews of the movie and those of the actors [BPW09]. An algorithm for feature sentiment extraction was used by Wang [WA08], but with an handicap of using reviews in Japanese as the source for the work.

The third phase is Feature Comparison. It is very similar to entity comparison, but with comparisons made in a feature level. Comparing the result of different features individually allows more accurate results, as the sentiment expressed for a specific feature is unique. This way it is possible to analyze and compare all the pros and cons of an entity individually.

In the feature driven opinion summarization method, for each product class, at first it automatically extracts general features. Then specific features and attributes and then assigns polarity to each of the feature attributes using a corpus and Support Vector Machines Sequential Minimal Optimization machine learning with the Normalized Google distance [CV05] [Pla98].

Hu and Liu (2004) [HL04] proposed an opinion summarization of products, categorized by opinion polarity. Their work is the most representative one in this area of study. Initially, they proposed the association rule mining to extract feature words. Then extract the opinion words in the sentences that contain at least one feature word. Finally, the feature-opinion pairs are generated and summarized according to the extracted features (Fig.2.6). They identified the sentiment orientation by the adjective synonym set in WordNet [Fel98] [MYTF02].

Liu et al (2005) [LH05] then illustrated an opinion summarization of bar graph style, categorized by product features. This model gives a more complete formulation of the opinion mining problem. It identifies the key pieces of information that should be mined and describes how a structured opinion summary can be produced from unstructured texts. Though, both of them are domain-specific. Popescu and Etzioni (2005) [PE05] proposed a domain-independent information extraction system. They identified four tasks in the review analysis:

- Product feature identification
- Identification of opinions regarding product features
- Determination of the opinion polarity
- Opinion ranking based on their strength

Feature driven methods extracts explicit product features using Pointwise Mutual Information (PMI). It uses explicit features to identify potential opinion phrases based on the intuition that an opinion phrase associated with a product feature will occur in its vicinity on syntactic parse tree. After the extraction of the opinion expression, relaxation labeling [HZ83], which is an unsupervised classification technique, is used to disambiguate the semantic orientation of opinion words. As a result, set of (feature, ranked opinion list) tuples are extracted [LJL08].

More sophisticated methods such as described by Koboyashi et al. (2007) [KIM07] states a quadruple as the opinion unit. In addition to the subject being evaluated and the opinion expressing positive or negative value, the opinion holder and the part (or the attribute) of the subject being evaluated are also included. A machine learning method which does not use domain-specific features is then applied with good results, resulting in greater precision and recall comparing to baseline models who do not use the same approach.

Another method for Opinion Mining was presented by Nakagawa et al. (2010) [NIK10] and uses subjective sentences using conditional random fields with hidden variables. When a sentence contains a word that reverse the sentiment polarity, all the sub trees depending on that variable are affected and their own variable is changed accordingly. The test was performed both in Japanese and English datasets, and the results were better than traditionally bag-of-features approaches.

2.5 Common techniques in Sentiment Analysis

This section presents some of the techniques commonly used in Sentiment Analysis in order do get more accurate results. These techniques are mostly used in pre-processing phases of Sentiment Analysis, with a filtering function to manipulate data, making it easier for applying machine learning algorithms.

2.5.1 Text Normalization and de-obfuscation

Text normalization (or Text Refining) is a widely used approach by which text is cleaned and transformed to a more consistent version. There is a set of text normalization techniques used for many purposes, including text pre-processing. As comments on websites are very likely to be full of errors and obfuscated in an inadvertently or even purposely way, normalizing text in a previous phase leads to obtaining much better results in the following mining phase. This happens in part because methods can then more easily identify relations and patterns on the text.

There are numerous examples of this kind of normalization from Unicode normalization [LS07], removal of punctuation [FSB06] to the removal of stopwords [XTW10, SR03]. In this case it is an extremely important feature, regarding the unstructured and grammarless nature of the text used for this case study (extract profiles about entities from news comments).

2.5.1.1 Techniques for text de-obfuscation

Concerning the normalization phase on text, there are some major techniques used: Orthographic text de-obfuscation, mainly used in Spam Filtering [FSB06] consists in substitutions of different characters with others, (or even removal of irrelevant ones, like hyphens in the middle of some words) allowing the discovery of a real word (a word present in the dictionary).

Compared with the normal version of Spam Assassin filter, who doesn't use de-obfuscation techniques , this technique achieved good results in identifying obfuscated words: using the subject field of 2377 received emails, this technique has achieved the identification of only 11 False

Negatives and no False Positives (False Positive are sentences incorrectly recognized as spam and False Negative are sentences incorrectly recognized as not spam), comparing to the regular Spam Assassin which identified 3 FP and 117 FN. In the case of news comments, it is frequent to find misspelled words with changed characters (be it a grammatical error or a way to let the word by-pass filters). So, finding the correspondent correct way of writing a word, gives us the opportunity to identify more opinionated expressions.

Another useful approach using word de-obfuscation techniques is the Unicode de-obfuscation variant [LS07].Sometimes users use characters from other alphabets, which hampers the identification of a word. Using 1000 spam emails for test effects, Spam Assassin was capable of identifying much more spam-related words applying de-obfuscation techniques, and therefore classified more emails as spam. With this method of de-obfuscation it is easier to identify words written in such way, because it allows the identification of external characters and their substitution by, in this case, their Portuguese counterparts.

2.5.2 Context

Personal opinionated texts frequently have the same words meaning different things and different words meaning the same, depending on the context. Understanding the context of a sentence is a good indicator for the real meaning of some words and expressions in a text. This third pre-processing method for Text Mining involves the context in which a given expression is inserted, as in a Sports context, an Economic context and so forth. That way, it can aid in the identification of the original meaning of the expression written by the author.

For instance, if a person uses a figure of speech, and by writing "massa" (Portuguese word for dough) or "papel" (Portuguese word for paper) maybe what he really wanted to mean was "dinheiro" (Portuguese word for money). This measure of context is based on testing a word “A” in contexts of word “B”. If the word "A" has similarities with "B", sharing some contexts means that the word "A" will substitute "B" without loosing the original idea of the sentence. If “A” does not have a similar meaning to "B", it will not fit well in the contexts of “B” [JAG08].

Tests were conducted using a collection of 1.4 billion words from Gigaword v.1 (collection of words distributed in DVD) and sentences extracted from British National Corpus. Conclusions suggested that some notion of similarity between words and different contexts can be acquired. The nature of opinionated text invites the usage of figures of speech such as euphemisms, and it can be helpful to identify which words are out of context and do not mean necessarily what is written.

2.5.3 String Matching

"String-matching consists in finding one, or more generally, all the occurrences of a string (more generally called a pattern) in a text" [CL]. There are two major variants of String Matching: Exact String Matching, where words are matched when one word is contained in the other in the exact same form (i.e. play in player), and Inexact String Matching that seeks for approximate patterns

between words. This last type is the one that is important for this work, as it can be used, for instance, to match a misspelled word with an existing word on the dictionary. This technique comprises finding approximate patterns on strings, obtaining a set of words, and then compares them with valid words extracted from dictionary or a lexicon.

2.5.3.1 Techniques for String Matching

In order to identify words used in comments, it is very useful to use techniques of string matching between words on the text and real words extracted from dictionary [SWB06, FSB06]. These methods are quite efficient because they support wildcards (character used to substitute any other character or characters in a string) and gaps when matching is being done. Supporting these features facilitates the discovery of words with incorrectly repeated characters or with some of those characters lacking, respectively.

Applying inexact string matching is a good way of finding the correct word correspondence in a dictionary as Sculley et al (2006) had demonstrated [SWB06]: using TREC (Text Retrieval Conference) 2005 public corpus, TREC 2006 Chinese spam dataset among other private datasets, using Perception Algorithm with Margins allowed a better efficiency in matching words when the length of messages was shorter, which is the case in web news comments. A similar technique will be used to match words on comments with words of the dictionary, extracted from SentiLex-PT02 [CSR11].

2.5.4 Stemming

"Word stemming is a technique that reduces closely related words to a basic canonical form or 'stem'. For example, the user inputs 'swims' and 'swimming' can be reduced to the basic stem 'swim' before performing an exact match against expected inputs" [Whi04]. Stemming is a widely used approach to infer the meaning of a word, by reducing it to the infinitive form. Even though it implies the creation of a lexicon to make the correspondence between a word and is stem, according to Cui et al. [CMD06] using stems improves performance of classifiers.

2.6 Related Projects

This section includes some implemented tools related with this thesis work. Those works can be directly related, like works of the same area and similarly structured. But they also can be from different areas but have some components in common with Opinion Mining, or even deal with some aspects that served as inspiration for this thesis work. The chapter is organized starting by simplest approaches to the more complex ones.

2.6.1 We feel fine

We Feel Fine [12] is an emotional search engine and web-based artwork whose mission is to collect the world's emotions to help people better understand themselves and others [KH11]. We Feel Fine is rule-based: it continuously crawls blogs, microblogs, and social networking sites, extracting sentences that include the words "I feel" or "I am feeling", as well as the gender, age, and location of the people authoring those sentences. Figure 2.7 shows an example of popular sentiments cached by the application.

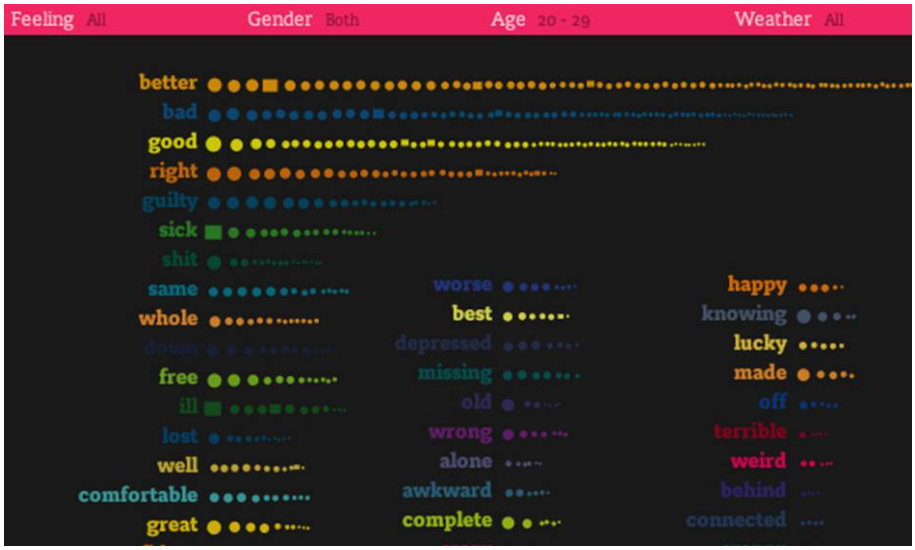


Figure 2.7: Example of interface from We Feel Fine Project

2.6.2 Tweetfeel

Tweetfeel [13] monitors positive and negative feelings in twitter conversations about many different topics including movies, musicians, TV shows and popular brands and displays these feelings in a clear and simple. It asks for the insertion of an entity by the user (a person, a brand name, a movie, and so on) and then looks for sentences in twitter talking about it. After that, it evaluates the sentiment of those sentences as positive and negative and presents the global sentiment based in that number.

One of best contributions of this work is in the white paper available at the enterprise website[14] called "There's Nothing Neutral about Neutral". It explains in a clear way how important neutral examples can be, and why they should be used for better results.

The positive side of this attempt on sentiment mining is the great accuracy achieved in the results; as it can be viewed in Figure 2.8 all the examples shown are correctly categorized as positive or negative, even the one using a negation form.

[12] http://www.wefeelfine.org/
[13] http://www.tweetfeel.com
[14] http://www.conversition.com

The downside of Tweetfeel is that the user have no control on the searched tweets and they don't seem to be obtained in real time; only a small number of sentences are evaluated, it seems like they are "hand-picked" and they represent a very small sample of all the possible tweets about a subject.

Figure 2.8: Tweetfeel screenshot searching for "Apple"

2.6.3 Twitrratr

Twittratr [15] is a StartupWeekend [16] project that started simply with the question of whether tweets about Barack Obama were generally positive or negative, and evolved to classify any topic. It works with a list of positive keywords and a list of negative keywords. Then, the application searches Twitter for a keyword and the results are cross-referenced against the keywords lists and then results are displayed.

When compared with Tweetfeel, this implementation (Figure 2.9 crawls more tweets and shows more detailed information like the neutral sentences. But the results seems less accurate, because of the rule-based approach instead of a machine learning one.

2.6.4 Emotext

Emotext [17] is the result of a dissertation that considers linguistic and psychological aspects to perform computer-aided categorization of opinions and emotions in texts. It discusses various

[15] http://www.twitrratr.com/
[16] http://startupweekend.org/about/
[17] http://socioware.de/EmoTextDemo/servlet/InputForm

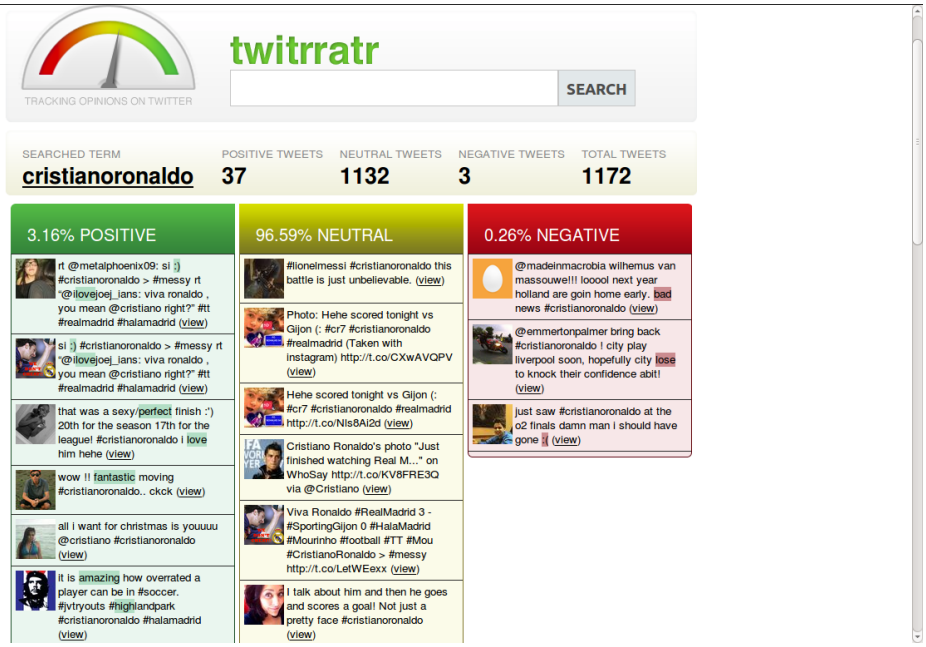


Figure 2.9: Twittratr screenshot searching for "cristianoronaldo"

emotional corpora (movie reviews, weblogs, product reviews, and natural-language dialogues) and describes different approaches to affect classification of their texts: a statistical approach that utilizes lexical, deictic, stylometric, and grammatical information; a semantic approach that relies on emotional dictionaries and on deep grammatical analysis; a hybrid approach that combines the statistical approach and the semantic approach.

The theoretical basis for this work is opinion mining and lexical affect sensing. As a result, Emotext is a more scientific and complete solution than the others presented, with a more functional and not so appealing interface [OA09]. It lets the user choose the classifier used (SVM or NB) and shows results using many different resources to train the classifier. A medium and majority of the results are shown to, but the user have indivual access to each result of using different resources to create his own method of classification.

2.6.5 Twitteuro

Twitteuro [18] is a barometer, developed by SapoLabs [19], that tracks the popularity and trends of the Euro 2012 teams and players in the Twittersphere. Twitteuro processes in real-time all the tweets that contain the #Euro2012 hashtag and identifies mentions to team and individual players. Thus, the more tweets containing the team or the player name, the higher their popularity.

It also presents the latest tweets collected by Twitteuro in real-time, for a given visualization context. In Figure 2.11 we can see the popularity of the German team, with special emphasis on Mario Gomez inside the bigger "bubble"; he is the most popular player for the time, because he

[18] http://twitteuro.sapo.pt/
[19] http://labs.sapo.pt/

Figure 2.10: Emotext screenshot for a long text classification using SVM

gave the victory to Germany, scoring two goals against Netherlands little time before the screenshot was taken.

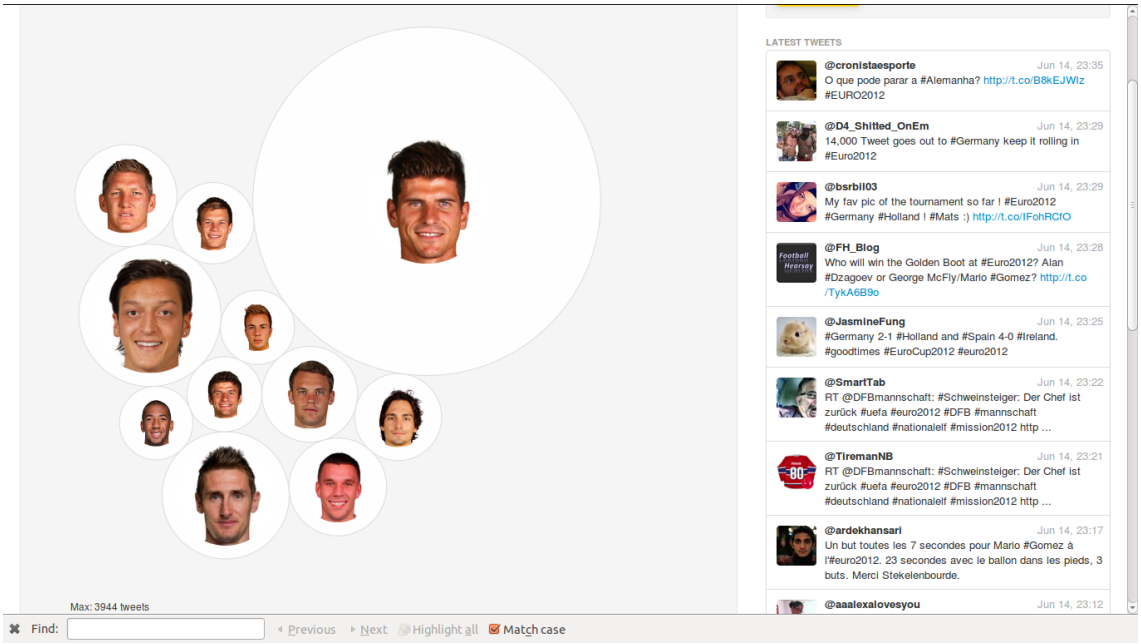


Figure 2.11: Twitteuro screenshot for German team

2.6.6 Socialmention

SocialMention [20] is a social media search and analysis platform that aggregates user generated content from across the universe into a single stream of information. It allows to easily track and measure what people are saying about companies, products, or any topic across the web's social media landscape in real-time. Social Mention monitors 100+ social media properties directly including: Twitter, Facebook, FriendFeed, YouTube, Digg, Google, and so forth.

Social Mention provides a point-in-time social media search and analysis service, daily social media alerts, and a third-party API for the user to manipulate the information given. But what really distinguishes this application is the knowledge build with the information collected (see Figure 2.12).

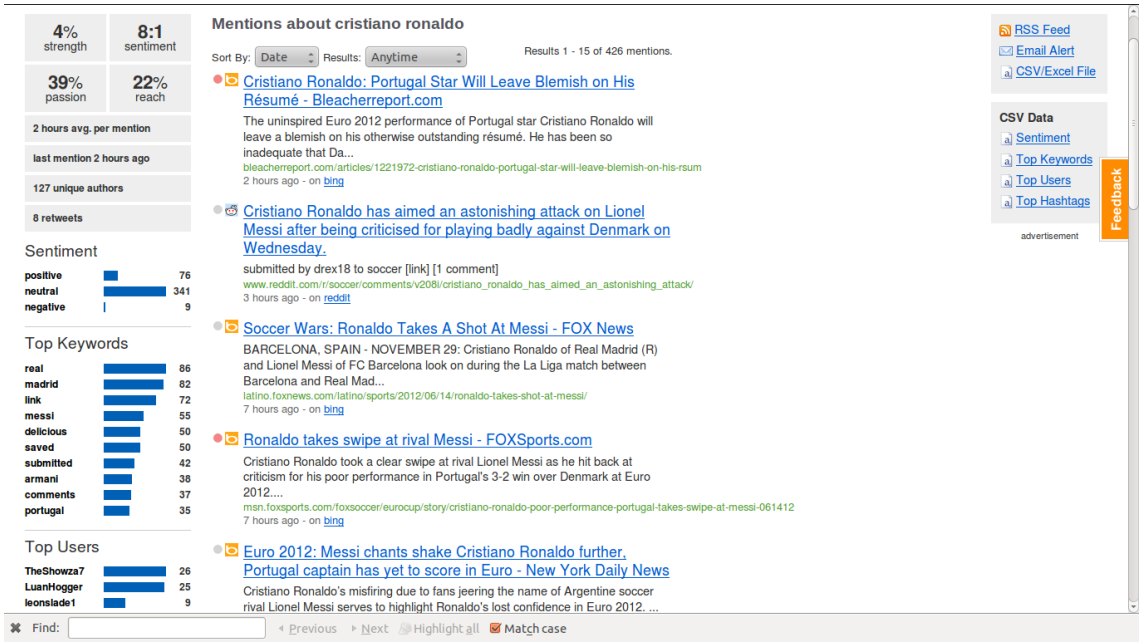


Figure 2.12: SocialMention screenshot searching for "Cristiano Ronaldo"

In the left side of the page it is possible to see the general sentiment, top keywords, top users, top twitter hashtags and the main sources where the information was collected. But beyond that, there are also four interesting measures created in an attempt to generate some knowledge from the collected information:

Strength is the likelihood that a brand is being discussed in social media. A very simple calculation is used: phrase mentions within the last 24 hours divided by total possible mentions.

Sentiment is the ratio of mentions that are generally positive to those that are generally negative.

Passion is a measure of the likelihood that individuals talking about a brand will do so repeatedly. For example, if a small group of very passionate advocates talk about products from a brand

[20] http://www.socialmention.com

all the time the brand will have a higher Passion score. Conversely if every mention is written by a different author the brand will have a lower score.

Reach is a measure of the range of influence. It is the number of unique authors referencing a brand divided by the total number of mentions.

2.6.7 Twendz

The twendz [21] Twitter-mining Web application uses the power of Twitter Search, highlighting conversation themes and sentiment of the tweets that talk about topics people are interested in. The crawling mechanism is collecting tweets in real-time.

This application provides some interesting features for the user. It lets the user choose the speed new tweets are added and sentimentally analyzed, always showing the last tweets on screen. The percentage of positive, negative and neutral comments are show in a bar below the searched term. Other interesting information display some subtopics related with the search, and also a word cloud with the most important tags around the topic. As Figure 2.13 shows searching for footballer David Silva, after he played from Spain against Ireland, the related information includes both countries.

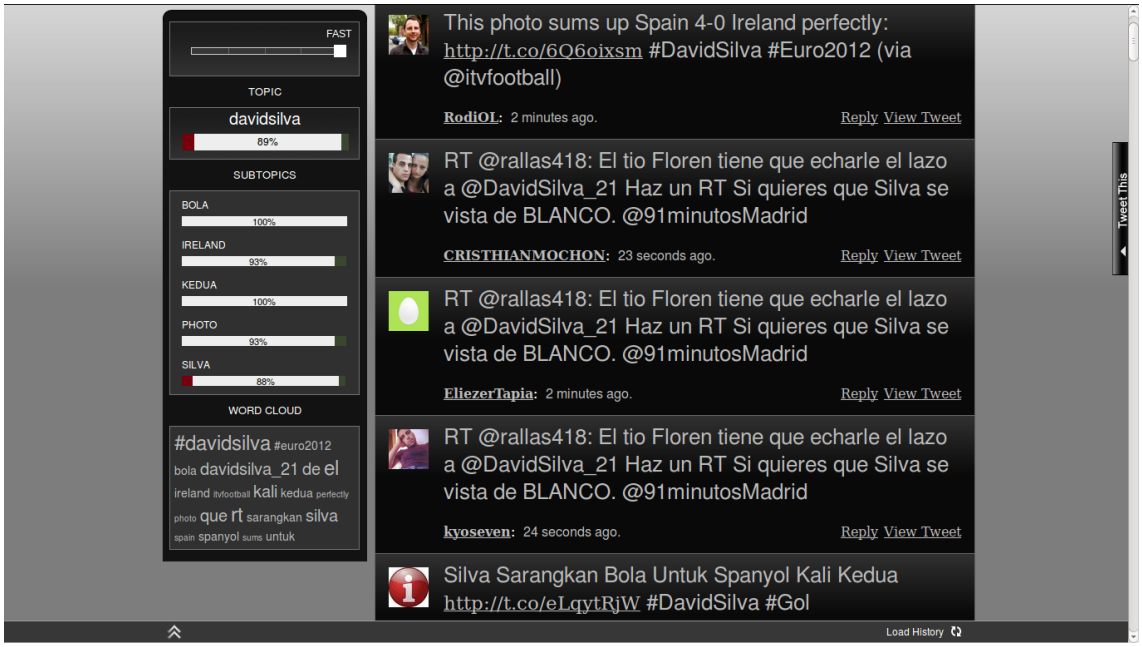


Figure 2.13: Twendz screenshot searching for "David Silva"

[21] http://twendz.waggeneredstrom.com/

Chapter 3

Resources

This chapter shows the resources used throughout this thesis. The starting point for this work are the comments made upon news articles, collected from a Portuguese website. All the other tools were used over those comments, with the objective of extracting information from them. They act as supporting tools to identify entities and then words or expressions that include opinions about those entities.

3.1 Comments

The dataset source for the comments was the Portuguese generalist website called "Sapo.pt", which includes a news section [1]. This choice is the result of the initial restriction of focusing on Portuguese written text. As the website news are written in Portuguese, most of the comments are expected to also be written in Portuguese.

This option has disadvantages related with the lack of similar work in the area, as well as poor resources to use, when compared with the available resources to a widespread used language as English. On the other side, applying techniques used in other languages to Portuguese, given the different nature and structure of languages, can be quite challenging.

The dataset collection includes comments between 27th April 2011 until 8th February 2012, performing a total number of more than 800000 comments.

3.2 Entity list

In the scope of this dissertation, an entity is defined only as the name of a person. Identification of entities in the dataset makes use of an entities list. This list is obtained using an on-line service called Verbetes [2], provided by Sapo. All the information from Verbetes is collected automatically from news sources, and its information is updated on a hour basis, as new news are collected. The service provides is accessible by a public API [3].

[1] http://www.sapo.pt/

[2] https://store.services.sapo.pt/en/Catalog/other/free-api-information-retrieval-verbetes

[3] https://store.services.sapo.pt/en/Catalog/other/free-api-information-retrieval-verbetes/technical-description

The list includes both National and International entities, with special incidence to national entities or directly related with Portugal (e.g, a foreign football player playing in Portugal); this happens due to the source website [4] being focused on generic Portuguese matters and only the most important international matters.

Selection of entities was based on choosing the most famous ones. This is made using the Verbetes functionality of selecting entities by the frequency they appear in the news, which means the most common are more discussed and will appear frequently in comments. The number of entities collected is superior to 6500. Other list with nicknames for entities was merged with the original entity list.

3.3 SentiLex-PT02

A lexicon is a set of words and expressions that represent the vocabulary, or part of it, contained in a language. Lexicons are useful for word recognition in computational linguistics. One example of a lexicon is Sentilex-PT02 which was used in this work. The lexicon contains annotations on the sentiment of the vocabulary. Annotations are used to find the sentiment of a sentence by matching an expression in the text with another one present in the lexicon.

SentiLex-PT02 is a sentiment lexicon for Portuguese, made up of 7,014 lemmas, and 82,347 inflected forms [CSR11]. The lexicon can be acquired by request following the instructions described in the official page [CSR11]. In detail, the lexicon describes:

- 4,779 (16,863) adjectives,
- 1,081 (1,280) nouns,
- 489 (29,504) verbs, and
- 666 (34,700) idiomatic expressions.

The sentiment entries correspond to human predicates, i.e. predicates modifying human nouns, compiled from different publicly available resources (corpora and dictionaries). An example of an entry is "aberração.PoS=N;TG=HUM:N0;POL:N0=-1;ANOT=MAN". The sentiment attributes for each entry are:

- the target of sentiment (eg: TG=HUM),
- the predicate polarity (eg: N0), and
- the polarity assignment (POL:N0=-1).

[4] www.sapo.pt

3.4 JSPELL

JSpell is a morphological analyzer and spell checking software. It is based on ISpell spell checker and it is directed to analysis on words and texts written in Portuguese [5]. The Portuguese dictionary is used along other available open source applications, such as Firefox, Thunderbird, and OpenOffice. Along with diverse usage for different kinds of research projects.

For each of the given words, the program obtains a morphological and semantic classification from the respective dictionary. The spell checker functionality allows the user to discover and fix misspelled words, with resource to word suggestion.

In this work, the Perl package Lingua::Jspell was used, which is a Perl interface to the Jspell morphological analyser [6]. The morphological information contained is crucial for extracting features used by classifiers.

The dictionaries purposefully exclude archaic and obsolete variations of words. They also exclude racial, religious and ethnic slang which may be considered offensive or illegal in some environments. Jspell dictionaries always try to be up to date, with words concerning emerging trends, current events, political figures, technology, and so forth.

[5] http://natura.di.uminho.pt/webjspell/jsolhelp.pl

[6] http://search.cpan.org/ ambs/Lingua-Jspell-1.84/lib/Lingua/Jspell.pm

Resources

Chapter 4

Implementation

This chapter is divided in four main sections. The first phase shows the overall architecture of the developed system. Following, an explanation on how the dataset was constructed with the available resources. Then, it is showed how the dataset was handled in order to identify the most common patterns regarding opinions on entities. Subsequently, the approaches to identify possible opinions in the text are explained as well as how the data was manipulated to improve final results.

4.1 Architecture

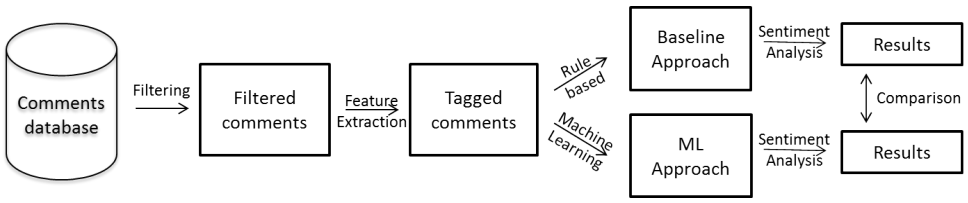


Figure 4.1: Overall system architecture

Figure 4.1 shows the architecture of the system developed regarding the achievement of proposed goals. It is just an overview about how the work was accomplished, and the rest of the chapter describes it with more detail.

The first step of the process is based on filtering the original dataset. This includes choosing the important information to keep from each comment and what information should be discarded. Encoding problems are also dealt with in this phase as well as spam behaviors: when a user repeats the same comment many times, only one is considered.

Feature extraction includes discovery of entities in sentences and possible opinions about each entity. Effective opinion identification differs depending on the respective approach, each corresponding to the achievement of a different goal. For the rule based approach only a specific

sentence structure is searched and identified using SentilexPT, but the chances of it being in fact an opinion are good. On the other hand, the machine learning approach implies the identification of many different structures, needing both SentilexPT and JSPELL to identify a wider range of words and expressions.

For both cases, the next step is the assignment of sentiments for the entities. After assigning and running both approaches, the results obtained are discussed and compared.

4.2 Obtaining the dataset

The methodological approach of this work started with the collection of comments from sapo.pt [1]. The information collected is the result of a database dump containing data in a table structure. This means that the crawling phase was already done in SapoLabs [2] and is not in the scope of this work.

Those comments are the result of the involvement and contribution of readers in discussions about the news topic. However, there are some patterns differentiating each topic: in politics comments often critically address all the entities involved. In sports related comments, comparisons between different entities are also very frequent in discussions between supporters of different teams.

A small extract of the comments and the additional information from it can be seen in Table 4.1. Beyond the comment itself, a unique identification (ID) number and the posted date for each comment were also stored. The reason why date was maintained is to detect exactly equal comments with small differences in the timestamp. This means they are duplicated comments posted by the same user with little delay. This is a typical spam behavior, and only one of those comments is considered, as an attempt to obtain a less biased dataset.

Table 4.1: Extracted comments example

Id	Timestamp	Comment
53	2011-04-27 17:09:34	estás perto de Marrocos e ainda por cima és Marroquino , é mesmo azar , coitado .
54	2011-04-27 17:09:50	Que no final do jogo , o primeiro a sair não esqueça de rebentar com os fusíveis , apagar a luz e ligar a rega ... Lol
55	2011-04-27 17:09:55	Será que a ASAE tem condições para fiscalizar a partidarice , uma vez que ninguém sabe quem paga e de onde provem o dinheiro para as monumentais jantarada ? Quase que poderia apostar que a factura das jantaradas não vão ser custeada pela partidarice .
56	2011-04-27 17:09:55	As moscas são sempre as mesma nunca mudam

[1] http://noticias.sapo.pt/
[2] http://labs.sapo.pt/

4.3 Feature Extraction

The first phase consisted in obtaining the list of entities to be identified through comments. Only entities (Table 4.2) that appeared more than 4 times to the crawler were selected. That list was joined with another one containing nicknames (Table 4.2) commonly used when people refer to public figureheads, in an attempt to identify entities when their real name was not used in the comment.

Table 4.2: Entities and nicknames example

Entities	**Nicknames**
Khaled Kaim	Putin
Daniel Bahr	Puyol
Musa Bility	Quaresma
Peter Praet	Queiroz
Gao Jianguo	Quintanilha
Jeremy Hunt	Racine
José Mujica	Radcliffe
Agnès Buzyn	Radu
Manuel Cruz	Rafael Branco
Jorge Bruno	Rafic

The following phase consisted in looking for all entities names into each comment. Whenever a sentence includes an entity, the entity is tagged in a XML (Extensible Markup Language) format, surrounding the entire entity name within <NAME> and </NAME>. If a sentence does not contain any entity, it is discarded and no longer taken into account for further analysis.

The remaining sentences were tagged again, but this time with Portuguese adjectives (e.g., "aborreça") or idiomatic expressions (e.g., "abrirá o coração") present in Sentilex-PT. As Sentilex has complementary information to each adjective and expression, the tags contain also information on the polarity of the tagged text (Table 4.1).

So, this time the tagging format contains more information; for adjective *bonita*, which means beutiful, the resulting tag is <ADJ bonito="1"> bonita </ADJ>. The word *bonito* is the radical form (or lemma) from where *bonita* derives. A radical is the origin of a word, normally in the singular and masculine form. The number "1" that appears next is the polarity indicator for the word, which means that the adjective used in this case has a positive connotation.

4.4 Precision, Recall and Accuracy

Before going into the used approaches with more detail, it is useful to be aware of some important notions involving performance measurement. This section starts with a brief explanation on the existing measures for opinion mining, and then shows how those measures were applied in this specific case. After that, the details of implementation are explained.

Precision and recall are two measures that have a significant meaning when analyzed together. For data mining purposes, precision is the fraction of correctly classified instances in the total amount of classified instances, whilst recall is the fraction of correctly classified instances in the total amount of relevant instances contained in the dataset. The formal definition of recall and precision implies the definition of four other notions:

True Positive (TP) The number of correct entries labeled;

True Negative (TN) The number of incorrect entries not labeled;

False Positive (FP) The number of incorrect entries labeled;

False Negative (FN) The number of correct entries not labeled.

In this work case, positive results are sentences that have both positive or negative polarity; negative results are sentences without an opinion expressed, as in with neutral polarity. A more detailed explanation is following shown:

True Positive Correct positive + correct negative values;

True Negative Correct neutral values;

False Positive Incorrect positive + incorrect negative values;

False Negative Incorrect neutral values.

Using these definitions allows a definition of recall and precision based on them:

$$Precision = \frac{TP}{TP+FP} \tag{4.1}$$

$$Recall = \frac{TP}{TP+FN} \tag{4.2}$$

A high precision value means a good proportion of correct categorizations made by the classifier, while a high recall value means a good proportion of categorized instances. If analyzed alone, precision and recall are not sufficient. Imagine sentiment analysis case with a dataset of 1000 instances in which 300 are sentimentally polarized:

- a precision value of 1 with only 10 instances analyzed is not a good result;
- a recall value of 0.95 with only 35% of precision is not a good result also.

This is the reason why both measures need to be used simultaneously, and to better understand the quality of results obtained there are measures that relates them both. One of the most simple is F-score, which attributes the same weight to precision and recall (considering a β value of 1):

$$Fscore = (1+\beta^2) * \frac{precision * recall}{(\beta^2 * precision) + recall} \tag{4.3}$$

Some earlier approaches only used positive and negative labeling (categorize opinion on entities), while some newer approaches use many more, classifying instances in a scale of different positive and negative intensity. As in this work, typically sentiment analysis has three different label options: positive, negative and neutral. Using neutral labeling adds the necessity of using True Neutral (correctly identified as Neutral) and False Neutral (incorrectly identified as Neutral) values. A measure for the global performance of the system is Accuracy [BOSB10], as it measures the proportion of true results obtained from all the analyzed ones.

$$Accuracy = \frac{TP + TNeutral}{TP + TNeutral + FP + FNeutral} \tag{4.4}$$

4.5 Baseline approach

After having obtained the dataset, a baseline was defined with the intent of better knowing the nature of the dataset used. The baseline expression serves as a starting point for the definition of more complex ones and also as a comparing measurement for them. With that in mind, the baseline needs to be very simple but common as well. So, the most common patterns expressing opinions were manually identified. The defined baseline is based in two structures (X represents an entity and Y represents an opinion):

- *X é um/uma Y* (X is a Y)
- *X é Y* (X is Y)

Other not so common structure consist in the direct adjectivation of an entity by finding the adjective and the entity next to each other or separated by an article. The translation does not match the same structure in the English language because of the different nature and structure of Portuguese language.

- (Article) Y X (e.g., *O ditador Passos Coelho* - The dictator Passos Coelho);
- Y (Article) X (e.g., *Aldrabão do Sócrates* - Sócrates that swindler);
- (Article) X Y (e.g., *do Sócrates pinóquio* - (that) Sócrates Pinocchio).

These are the structures initially used for finding opinions about entities. It is a special case of opinion atribution commonly used by people when they want to express feelings about an entity, recurrently using figures of speech. The example for the third structure showed above is an example of that, where the user tried to say that *Sócrates* is a liar by referring to him as Pinocchio. Finding the meaning of these types of sentences is easy for a human but very complex for a computer, and is out of the scope of this work.

Using a rule based approach with this simple structure allows the classification of some of the most simple opinions present in the dataset with a great precision value. This happens due to the

big correlation between the entity and the adjective; as they are next to each other, almost all the times the adjective is referring to the entity.

So, the classification by rules is simply a direct match to the adjective polarity: if the polarity is negative the opinions is negative; if the polarity is positive the same classification is attributed to the opinion on the entity. the same happens for neutral adjectives; The expected results are higher in this first attempt. The downside of this simple approach is that most of the opinions are neglect because they are not expressed in this form, which makes the recall value very low.

4.6 Machine Learning structure approach

The following phase intends to expand the structure of sentences with opinions on entities. The accepted sentence structure is amplified when compared to the rule-based approach, to enable a number of other words between an entity and an adjective, called tokens. It is expected that this allows a great increase in the recall value when comparing with the very low value of the same measure for the baseline approach. This happens due to the fact that the baseline represents a very small percentage of the total comments with opinions. Even with the tokens approach, there are many other variables, specified in the next chapter, that contribute to a relative low number of selected sentences and consecutively a low recall value.

Important factors that might decrease the relation between Entity and Adjective are punctuation signs in the middle of the sentence. For instance, if a sentence has an "!", "?" or "." between the entity and the adjective, the adjective is not likely to be referring to the entity; they are inserted in different sentences of the same text. An example of that is the following sentence:

- "[...] Sócrates! Aldrabão é o [...]"

Clearly the entity (Sócrates) and the adjective (aldrabão) are not correlated in this case. This is the reason why sentences with one of the previous punctuation signs separating an entity and an adjective are discarded of further evaluation.

The following step consists in defining good tokens number between the entity and the adjective. The problematic involved in choosing this number is obvious: a small number does not improve the recall that much; as the tokens number increases the number of identified sentences and recall value also increases, but the connection between the entity and the adjective has a higher probability of being lost.

This means that a big number of tokens often implies that the adjective is not directed to the entity, which will lead to a wrong classification of the entity, and ultimately decrease the precision value. For a careful choice of the tokens number, a study using many different tokens number was conducted, and sentences with punctuation in the middle of an Entity and Adjective were promptly discarded. The graphic in Figure 4.2 shows the relation between the exact number of tokens used and the number of sentences present in the dataset, using an extract of the dataset of 100.000 sentences randomly chosen.

By observing Figure 4.2 it is possible to obtain some conclusions:

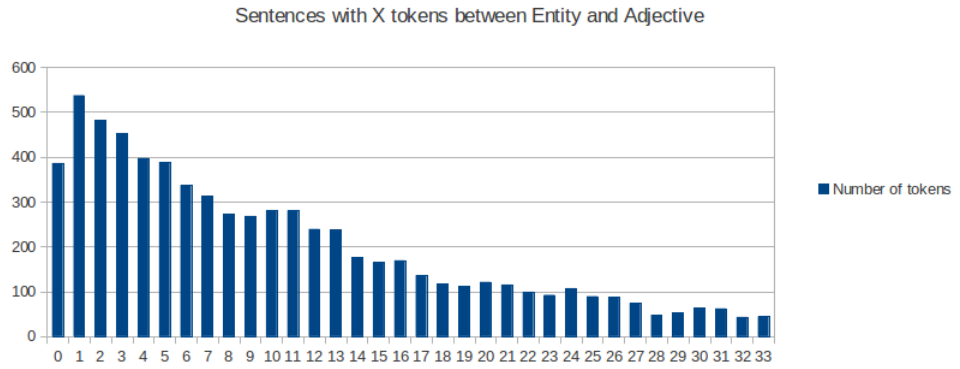


Figure 4.2: Graphic showing the number of sentences extracted using different tokens number

- it is common that 0 tokens separate the Entity and the Adjetive, which is explained by the frequent adjectivation present in this type of text (Eg: *Passos ditador*);
- the most common number of tokens is 1, consistent with the most common path being *X é Y*, as in *Louça é louco*;
- from 1 to 8 tokens the number of sentences gradually decreases;
- maintain similar results from 8 until 11 tokens: manual observation of results revealed that the number of entities and adjectives not related starts to increase;
- from 12 tokens on, there are much less paths identified; and the ones identified are commonly false positives, as entities and adjectives are not related.

After observing with more detail the results with more than 8 tokens, it is obvious that the big majority of entities and adjectives are not related. Frequently, there is an unidentified entity or adjective in the middle of both words: for the first case, it was the real entity to whom the opinion is intended; for the second case, it was the adjective intended to classify the entity.

Further analysis on 100 of the resulting sentences has been done, for each tokens number lower or equal to 12. Results revealed acceptable correlation values between entity and adjective for 8 or less tokens (68%). Knowing that, the initial number of tokens used was 8 or less. This means that only sentences with 8 or less words separating an Entity and an Adjective were considered. To obtain better results the tokens number changed later to 6 or less, expecting a minimum of 74% entity related opinions.

After finding the number of tokens to use, only the entities and adjectives were tagged in sentences. Only these two tags provide little information, so there was a need to tag all the other words in sentences. This was done using the morphological analyzer JSPELL.

JSPELL has the advantage of having a much wider range of words comparing with Sentilex, including more adjectives and nouns and also all the other morphological classes as pronouns, articles, names, verbs and so forth. But it does not have polarity values for words, which prevents using JSPELL for sentiment evaluation.

JSPELL includes also many complementary information for each word regarding gender, grammatical tense, radical form between others. With the possible utilization of such information, the JSPELL tagging function created in this work includes an option to use or omit that information when tagging a word.

An example of that information is the two resulting forms of tagging the Portuguese word *jogadores* (players in English): the word is a substantive / adjective (represented by a_nc) [3] and the tagging without additional information is <a_nc>jogadores</a_nc> , whilst tagging with that information becomes <a_nc T="inf" FSEM="dor" N="p" TR="_" G="m" rad="jogar">jogadores</a_nc>. The first one is more human readable, but the second one can give more information for processing.

JSPELL was then used to tag all the words and possible punctuation in sentences (for example "," as sentences with final punctuation were previously removed) and following is the result of a sample sentence tagged without the complementary information:

Initial Sentence <NOME>Jorge Jesus</NOME> é sem margem para dúvidas o <ADJ melhor="0"> melhor </ADJ>

Tagged Sentence

```
<html>
    <NOME>
        Jorge Jesus
    </NOME>
    <v  T="p" N="s" P="3" TR="_" rad="ser">
        é
    </v>
    <prep  rad="sem">
        sem
    </prep>
    <nc  N="s" G="f" rad="margem">
        margem
    </nc>
    <v  T="i" N="s" P="3" TR="_" rad="parir">
        <v  T="pc" N="s" P="3" TR="_" rad="parir">
            <v  T="pc" N="s" P="1" TR="_" rad="parir">
                <v  T="i" N="s" P="2" TR="_" rad="parar">
                    <v  T="p" N="s" P="3" TR="_" rad="parar">
                        <prep  rad="para">
                            para
                        </prep>
```

[3] http://natura.di.uminho.pt/webjspell/jsolhelp.pl

```
                    </v>
                </v>
            </v>
        </v>
    </v>
    <nc  N="p" G="f" rad="dúvida">
        dúvidas
    </nc>
    <ppes  C="a" N="s" P="3" G="m" rad="o">
        <art  N="s" CLA="def" G="m" rad="o">
            o
        </art>
    </ppes>
    <ADJ melhor="0">
        melhor
    </ADJ>
</html>
```

Note the amount of apparently repeated <v> (verb) tags of the word *para*. This happens because that word has many verbal forms. It can derive from radical verbal form *parir* (to give birth) or *parar* (to stop). Even for the same verbal form, the word can be in a different grammatical tense (present or past tense) or even be used in first or third person.

Besides that many verbal tags, the same word has also a different morphological value tag, <prep>, that stands for preposition. This means that all possible morphological functions of the word appear in the tag. The next challenge was how to pass all that knowledge in an understandable form for automatic processing. Passing all those tags separately would turn the analysis biased, because when processing the information each tag would be considerate a different word.

In order to deal with this issue, the solution found aggregates tags for the same word with a punctuation mark (+). This way, each set of tags for the same word is evaluated as a whole, and some more information is passed to the classifier: a word with <ppes> (personal pronoun) and <art> (article) morphological functions is different from a word that is only a <ppes> (personal pronoun), as in *o* and *meu* (the and mine). With this method, the previous shown sentence becomes the following:

Previous Sentence <NOME>Jorge Jesus</NOME> é sem margem para dúvidas o <ADJ melhor="0">melhor</ADJ>

New Tagged Sentence

```
<html>
    <NOME>
        Jorge_Jesus
```

```
        </NOME>
        <v>
            é
        </v>
        <prep>
            sem
        </prep>
        <nc>
            margem
        </nc>
        <v>+<v>+<v>+<v>+<v>+<prep>
            para
        </prep>+</v>+</v>+</v>+</v>+</v>
        <nc>
            dúvidas
        </nc>
        <ppes>+<art>
            o
        </art>+</ppes>
        <ADJ melhor="0">
            melhor
        </ADJ>
    </html>
```

Note that all the arguments regarding each morphological function are suppressed in this example for better readability, but in the actual case all the arguments are also present.

Even using all this techniques, there are many sentences with adjectives not related with the entity. One of the most important reasons for that is the nature of the comments; user-generated content is commonly disorganized and do not respect syntactic rules to form sentences. In many cases, sentences do not make any sense and are just a collection of random text with an entity and an adjective.

To make sure that each sentence analyzed respects some syntactic rules and have some notion of structure, another condition was created. For a sentence to be considered for evaluation, it is mandatory that a verb is included between the entity and adjective. This is done by verifying that there is a <v> (verb) tag in the sentence. This is a way of granting a high probability that the sentence is correct and might express an opinion.

After the tagging phase, the sentences are manually assigned to a polarity value. The classifier was used through Perl Algorithm::NaiveBayes [4], that receives a set of features to train the system with positive, negative and neutral examples. After that, each sentence is passed to the classifier

[4]http://search.cpan.org/ kwilliams/Algorithm-NaiveBayes-0.04/lib/Algorithm/NaiveBayes.pm#METHODS

as a test subject. The classifier returns the prediction that is compared with the manually labeled value, to assure if the prediction was correct or no.

Chapter 5

Results interpretation and discussion

This chapter starts with a little briefing on the used approaches detailed before, and then presents the results obtained. After that, the results meaning is discussed and also what could be done to improve them.

For the purpose of calculating recall values, the total number of sentences would be required. As it is not viable to evaluate more than 200000 sentences manually, that value needed to be estimated.

Through analysis of 1500 random sentences from the whole dataset, there were 423 sentences expressing opinions with an acceptable structure for the classifier to find it. Many others express opinions in a very subtle way, and were not considered as the classifiers would never identify it. So, it is expected that 28,2% of them have opinions. This estimated number is used in the calculation of all recall values.

5.1 Rule-based approach

The rule based approach was based in looking for two sentences with similar structure. Each structure searched returned a total number of opinionated sentences with that structure. From those, 500 randomly chosen sentences are extracted and manually marked as positive, negative or neutral. Then, sentiment on the same sentences is calculated based on the polarity given by the lexicon on the adjective feature. The first structure was *X[Entity] é Y[Adjective]*, and the results collected are visible in Table 5.1.

Table 5.1: Results of classifications on sentence structure *X[Entity] é Y[Adjective]*

Correct positive	183	**Incorrect positive**	42
Correct neutral	52	**Incorrect neutral**	61
Correct negative	154	**Incorrect negative**	8

The results shown in Table 5.1 allows the calculation of the performance measures that can be seen next. As expected, the results are generally high for precision in this specific approach to opinion mining:

Precision 0.87

Recall 0.02

F-score 0.04

Accuracy 0.78

The total number of retrieved sentences within this structure is 1442. To calculate the recall value, the estimated value of 73752 sentences with opinions is used, resulting in a recall value around 2%. Obviously, recall value suffers with this too simple approach.

The relatively low precision (the expected value was above 90%) value is due to the frequently appearance of adjectives like *bom* (good) that have a positive sentimental value, but in a sentence they usually precede another adjective not detected in the tagging phase that adds polarity to the sentence. This leads to many mistakes in automatic classification, as it is very common that a negative or neutral opinion is being expressed.

But a lowest score was given by Accuracy, with 78%. By looking at the results it is possible to see that neutral opinions are the reason for this lower score, because there are more false than positive neutrals.

Looking at the dataset enabled the identification of the motive for that to happen: words used to compare entities like *melhor* (better) are very common in the text and are marked as neutral by Sentilex. This is probably because a comparison has at least two entities, normally one with positive sentiment and another with negative sentiment.

But as the structure searched implies that the targeted entity is at the left of the adjective, it is possible to assume that the word *melhor* has a positive connotation for the entity (the typical sentence is *X é melhor que X2* which stands for X is better than X2). With this change, the number of false neutrals decrease from 61 to 14; on the other hand, the number of true positive increases from 183 to 230.

These improvements demands recalculation of both precision, F-score and accuracy, with values of 0.89, 0.04 and 0.87 respectively.

The other sentence structure searched in a rule-based approach was *X é um Y* (X is an Y). The setup was similar to the earlier one, but this time the number of retrieved sentences is smaller. This occurs because the sentence structure is not so common, resulting in a total number of 567 sentences. Detailed results are described in Table 5.2.

Table 5.2: Results attempt on sentence structure *X[Entity] é um Y[Adjective]*

Correct positive	118	**Incorrect positive**	37
Correct neutral	4	**Incorrect neutral**	32
Correct negative	305	**Incorrect negative**	4

The corresponding performance measures can be observed below:

Precision 0.91

Recall 0.01

F-score 0.02

Accuracy 0.85

When comparing with the results from the previous setup, this one has a trend of many more negative sentiment. On the other hand, there are much less neutral sentences, making this a case of almost exclusively polarized sentences. Another relevant aspect is that the first precision value is significantly higher, but there is no simple way of improving it as with the previous setup.

There are two types of problems visible in the different approaches to improve system general performance. Namely, improving recall and precision (accuracy depends on precision also) values. For improvements in precision and accuracy it would be useful that:

- sentences with entities that are not specific, as *Presidente* (President), were discarded;
- polarization by context: that is to say use different sentiment value for the same word in different contexts (for instance, words *ajudado* and *beneficiado* can have a positive connotation in an economics context, but a negative sentiment on a sports context.

Even though their contribution to the decrease in the overall performance is very small, some adjustments on the lexicon could also be made:

- discard words that are used most of the times with a different morphological function (e.g., bruno is almost exclusively used as a noun instead of an adjective);
- discard of words wrongly recognized as adjectives (e.g., bruna and brunas do not exist as adjectives but were automatically obtained from adjective bruno) from SentilexPt;

For improvements in recall there is a huge potential, possible by making some changes in the way features are obtained:

- adding more adjectives to the lexicon, because the spectrum of polarized adjectives is very restricted;
- adding more entities to the entity list;
- when a pronoun is encountered, search the rest of the sentence for the corresponding entity name.

But even applying these improvements there is more room for improvement in recall values. This is due to the restricted sentence structure of the rule-based approach, when considering many different possibilities of expressing an opinion. In an attempt to amplify the accepted opinions, techniques using machine learning were applied.

5.2 Machine Learning

One of the most simple and widely used machine learning algorithm is Naive Bayes, based on the Bayesian probabilistic theorem. This time, getting features for the classifier was based in tagging all words between and entity and an adjective, with polarity when possible. Then, each word tag is interpreted as a feature, and passed to the classifier for learning and then for testing.

The setup this time allows a bigger number of retrieved results, because the sentence structure is not so rigid. For each case, from the total number of retrieved results, 500 sentences were randomly selected and assigned a polar value manually. Then, the classifier was run and the results were measured against the manually labeled values.

The results obtained are shown on Table 5.3, using only words with an exact number of 6 tokens. The choice for 6 tokens, explained in the previous chapter, is mainly due to the strong correlation between entities and adjectives. Also, using an exact number of tokens is useful to compare results with the further approach of less than a given number of tokens.

Table 5.3: Results attempt using Naive Bayes classifier with exactly 6 tokens

Correct positive	59	**Incorrect positive**	47
Correct neutral	209	**Incorrect neutral**	80
Correct negative	59	**Incorrect negative**	46

This approach resulted in 1908 retrieved results. With all the information obtained, the calculation of performance measures was done with the following results:

Precision 0.56

Recall 0.03

F-score 0.06

Accuracy 0.65

Then, an approach using 6 or less token words between the entity and the adjective was done. The number of retrieved results was obviously bigger, totaling 19838. The results of the labeling process can be seen in Table 5.4

Table 5.4: Results attempt using Naive Bayes classifier with 6 or less tokens

Correct positive	169	**Incorrect positive**	41
Correct neutral	42	**Incorrect neutral**	38
Correct negative	164	**Incorrect negative**	46

The collected information allowed the calculation of performance measures:

Precision 0.79

Recall 0.27

F-score 0.40

Accuracy 0.75

As it is possible to see comparing these results with the last attempt, the results were fairly better this time. Recall value is higher due to the great number of sentences with possibility to be analyzed. As it was explored in the last chapter 4.2, the majority of opinions have a small number of tokens separating the entity and the adjective.

Precision and accuracy are also higher, which shows a bigger correlation between the entity and the adjective as the number of tokens decreases. A bigger correlation means that an adjective is more connected with the entity, giving more chances for the classifier to make a correct guess.

As a final improvement, 3 more changes were made to the system. First, the inclusion of negation in the features passed, made by tagging words like *não* or *nunca* with negative sentiment. The polarity of words like *melhor* were changed from neutral to positive sentiment. Also, some ambiguous entites like *Presidente* or homograph entities like *Longo*, *Vale* or *Dias* (as a surname). Or even adjectives mainly used with other purpose *são* and *vão* (both are commonly verbs) were removed from analysis. Results collect are available in Table 5.5.

Table 5.5: Results attempt using Naive Bayes classifier with 6 or less tokens, with improvements

Correct positive	168	**Incorrect positive**	40
Correct neutral	38	**Incorrect neutral**	30
Correct negative	161	**Incorrect negative**	44

This time the number of retrieved results was slightly smaller, 18982. Looking at the results reveals immediately a decrease in neutral assignments, mostly the wrongly assigned. The reasons are the elimination of entities with low correlation with the opinions, and also the change in polarity in the lexicon from neutral to positive. The detailed results are following shown:

Precision 0.80

Recall 0.26

F-score 0.39

Accuracy 0.76

5.3 Entity profiling

In an attempt to create an opinionated profile about each entity, the most common opinions on each entity were aggregated. The radical from which every adjective is derived was used as the prevailing sentiment.

As the detection of sentiment is restricted by the the limited amount of polarized adjectives, the number of opinions in a given entity is not very high. The most discussed entities are in a range

between 100 and 200 total opinions. A word cloud of opinions around José Mourinho is shown in Fig 5.1, with the number of each opinion in brackets.

These are only some of the most common adjectives used to classify an entity. The most common expression, *melhor* (the best or better) is not a direct opinion. But it shows a positive sentiment around him, either saying that he is the best one, or that he is better than the following entity.

The figure of José Mourinho is famous in part because it does not generate consensus. That is supported by opinions calling him arrogant (*arrogante*) and dirty (*sujo*) but also sincere (*verdadeiro*). One last note for the word *mau*, which is quite frequent but it does not necessarily mean a negative sentiment. It acts as a modifier, because it is followed by other word (adjective or not). Only combining both words would allow the discovery of the real sentiment, but often the second word is not detected and marked as a feature.

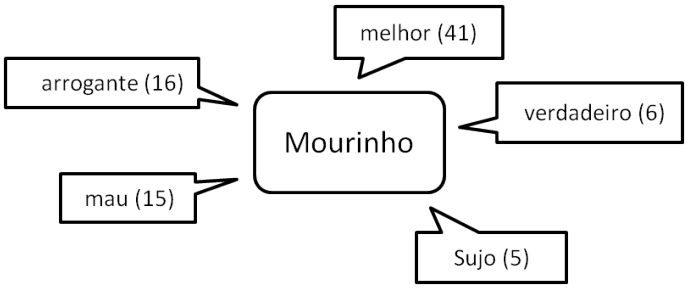


Figure 5.1: Word cloud for Mourinho

A more interesting example of the knowledge extraction is shown in Figure 5.2. These time, it is possible to see a comparison between two of the most famous Portuguese politics in recent years: the current and the last prime ministers. Once again only some of the most common sentiments are shown.

José Socrates, the ex-prime minister of Portugal, has some negatives opinions about him including scoundrel (*malandro*) and unqualified (*incompetente*). But the most interesting opinion says that he is forgiven (*perdoado*): this opinion appears massively because he is no longer the prime minister, and people make comparisons between him and the actual man on the position, Passos Coelho, forgiving José Socrates. It is expected that when Passos Coelho leaves the position, the same comparisons happen between him and the following prime minister.

There is also a significant opinion shared by both entities, people think they are both liars (*mentiroso*). This is an adjective specially connected with politics context in people opinions: just by watching a regular newscast emission it is very common to see someone calling liar to a politic.

It should be noted that Pedro Passos Coelho is here represented by two different nomenclatures: Passos Coelho and Coelho. Even knowing that Coelho is a common name in Portuguese (meaning rabbit), all the opinions are directed to the person. Both forms share the adjective *melhor* (better) and have their own opinions.

But since the opinion is about the same person, a good improvement for the profiling system will be aggregating names that represent the same person, to obtain more reliable resources. It is easy to know that Passos Coelho and Coelho represents the same person, but there are other not so direct associations, as Sócrates and Pinócrates, that would represent a more challenging task.

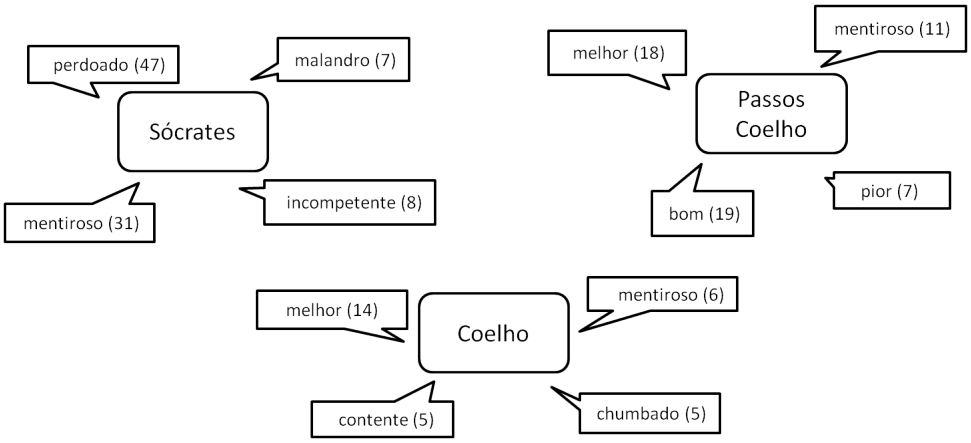


Figure 5.2: Word clouds for politics

Chapter 6

Conclusions

The problematic detailed throughout this dissertation was based on extracting opinions from user-generated content, in the form of comments on a news webpage written in Portuguese. Those opinions were extracted as features from text. Then, sentences were sentimentally evaluated and grouped to find the most common opinions on a given entity.

The big challenge presented in this dissertation is adjust knowledge and techniques used in the area to the Portuguese language. There is a great collection of studies for English, but Portuguese has a very different sentence structure as well as more complex grammar rules for verbs conjugation.

Furthermore, the available resources are limited in the amount of information provided. The annotated words are restricted in number, which complicates the discovery of sentiments in a text. Dealing with this issues was the main concern through this dissertation, with various efforts aimed to bypass this shortcoming. This efforts included using negation, change of polarity for some adjectives, refining the entity list and also the adjectives detected. But the first step when starting a similar work to this dissertation is to expand the available resources from the very beginning, without loosing their good ratio of correct entries.

Even with the efforts for improving the final results, for a more substantial improvement on results the resources available would need substantial changes. Perhaps a good solution would be using a dictionary, like JSPELL used for this work purpose, alongside with Portuguese grammar rules for formation of new words on the lexicon. This is a feasible solution for expanding polarity assignment, using some previously reliable knowledge about many words.

Another possible expansion is related with the concept of context. The same adjective or expression frequently represent different sentiment when used in different contexts. So, some entries simply do not fit in the context, and should be discarded or at least changed. Changing polarity of words by taking into account that it was used in a specific context might improve the precision of analyzing sentiments.

Other improvements can be achieved by understanding how more complex word modifiers can change the sentiment of a word. In this work a simple negation was used, but there are a variety of modifiers with subtle effects on sentiment that can be explored and inserted in a future work.

Conclusions

The main goals intended in this work were achieved. Firstly, selecting the information to work with. Then, studying correlation between entities and opinions as different sentence structures and sizes were tested. Feature extraction problems as dealing with words with many morphological functions or negation in a sentence. Also, using two different approaches to find sentiment: the simpler rule-based approach, and the more versatile machine learning approach.

However, one of the last desired goals fell short when comparing with what initially desired. Comparing multiple machine learning algorithms performance was not done (only a Naive-Bayes classifier was used), as the extraction of features had bigger priority and effect on the general system performance.

Also, the final goal of creating opinionated profiles is not very well developed: the used methods are very simple, and the results of sentiment analysis could be joined to form a more complete source of knowledge.

Future development on this work might include the creation of an interface for the system to show results in a more user-friendly way.

References

[ACS08] Ahmed Abbasi, Hsinchun Chen, and Arab Salem. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems*, 26(3):1–34, 2008.

[AS06] Giuseppe Attardi and Maria Simi. Blog mining through opinionated words. *Proceedings of TREC*, pages 2–7, 2006.

[BGV92] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory COLT 92*, 6(8):144–152, 1992.

[BM08] Alexandra Balahur and Andres Montoyo. A feature dependent method for opinion mining and classification. *2008 International Conference on Natural Language Processing and Knowledge Engineering*, pages 1–7, 2008.

[BOSB10] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. The balanced accuracy and its posterior distribution. *2010 20th International Conference on Pattern Recognition*, 0:3121–3124, 2010.

[BPW09] Haji Binali, Vidyasagar Potdar, and Chen Wu. A state of the art opinion mining and its application domains. *2009 IEEE International Conference on Industrial Technology*, pages 1–6, 2009.

[BX09] T Bhuiyan and Y Xu. State-of-the-art review on opinion mining from online customers' feedback. *Proceedings of the 9th Asia-Pacific*, (November):4–7, 2009.

[CL] Christian Charras and Thierry Lecroq. EXACT STRING MATCHING ALGORITHMS. http://www-igm.univ-mlv.fr/~lecroq/string/, year = 1997.

[CMD06] Hang Cui, Vibhu Mittal, and Mayur Datar. Comparative experiments on sentiment classification for online product reviews. *Computing*, 21(2):1265, 2006.

[CSR11] Paula Carvalho, Mário J. Silva, and João Ramalho. SentiLex-PT_02_in_English @ dmir.inesc-id.pt, 2011.

[CV05] Rudi Cilibrasi and Paul Vitányi. Automatic meaning discovery using google. *Knowledge Creation Diffusion Utilization*, (cs.CL/0412098):1–31, 2005.

[Dam08] T Edward Damer. *ATTACKING FAULTY REASONING A Practical Guide to Fallacy-Free Arguments*. Cengage Learning, 2008.

REFERENCES

[DH09] Lipika Dey and Sk Mirajul Haque. Opinion mining from noisy text data. *International Journal on Document Analysis and Recognition IJDAR*, 12(3):205–226, 2009.

[DL08] Xiaowen Ding and Bing Liu. A holistic lexicon-based approach to opinion mining. *conference on Web search and web data mining*, page 231, 2008.

[DLP03] Kushal Dave, Steve Lawrence, and David M Pennock. Mining the peanut gallery : Opinion extraction and semantic classification of product reviews. *Word Journal Of The International Linguistic Association*, 17(5):519–528, 2003.

[ECD+05] Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S Weld, and Alexander Yates. Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence*, 165(1):91–134, 2005.

[FB06] Eibe Frank and Remco R Bouckaert. Naive bayes for text classification with unbalanced classes. *Machine Learning*, 4213:503–510, 2006.

[Fel98] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*, volume 71. MIT Press, 1998.

[FSB06] Valerio Freschi, Andrea Seraghiti, and A. Bogliolo. Filtering obfuscated email spam by means of phonetic string matching. *Advances in Information Retrieval*, pages 505–509, 2006.

[GACOR05] Michael Gamon, Anthony Aue, Simon Corston-Oliver, and Eric Ringger. *Language*, 3646(3646):121–132, 2005.

[Gro08] Miniwatts Marketing Group. Internet growth statistics. http://www.internetworldstats.com/stats.htm, Jan 2008.

[GZ06] Andrew B Goldberg and Xiaojin Zhu. *Seeing stars when there arent many stars: Graph-based semi-supervised learning for sentiment categorization*, page 45. Number June. Association for Computational Linguistics, 2006.

[HL04] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining KDD 04*, 04(4):168, 2004.

[HM97] Vasileios Hatzivassiloglou and Kathleen R McKeown. Predicting the semantic orientation of adjectives. *Proceedings of the 35th annual meeting on Association for Computational Linguistics*, pages(2):174–181, 1997.

[HW00] Vasileios Hatzivassiloglou and Janyce M Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. *Proceedings of the 18th conference on Computational linguistics*, 1(3):299–305, 2000.

[HZ83] R A Hummel and S W Zucker. On the foundations of relaxation labeling processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(3):267–287, 1983.

REFERENCES

[JAG08] Sanaz Jabbari, Ben Allison, and Louise Guthrie. Using A Probabilistic Model Of Context To Detect Word Obfuscation. In *Artificial Intelligence*, pages 1624–1628, 2008.

[KH04] Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. *Proceedings of the 20th international conference on Computational Linguistics COLING 04*, 4(28 January):1367–es, 2004.

[KH11] Sepandar D Kamvar and Jonathan Harris. We feel fine and searching the emotional web. *Design*, 66(1):117–126, 2011.

[KIM07] Nozomi Kobayashi, Kentaro Inui, and Yuji Matsumoto. Opinion Mining from Web Documents: Extraction and Structurization. *Transactions of the Japanese Society for Artificial Intelligence*, 22(1):227–238, 2007.

[KN06] Hiroshi Kanayama and Tetsuya Nasukawa. Fully automatic lexicon expansion for domain-oriented sentiment analysis. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing EMNLP 06*, (July):355–363, 2006.

[Kop05] Moshe Koppel. Using neutral examples for learning polarity. *Proceedings of International Joint*, pages 1–2, 2005.

[LA07] Jure Leskovec and LA Adamic. The dynamics of viral marketing. *ACM Transactions on the Web*, 1(May):1–46, 2007.

[LH05] Bing Liu and Minqing Hu. Opinion observer: analyzing and comparing opinions on the Web. *international conference on World Wide Web*, pages 342–351, 2005.

[LJL08] Dongjoo Lee, Ok-Ran Jeong, and Sang-goo Lee. *Proceedings of the 2nd international conference on Ubiquitous information management and communication ICUIMC 08*, page 230, 2008.

[LP01] Dekang Lin and Patrick Pantel. DIRT @SBT@discovery of inference rules from text. *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '01*, pages 323–328, 2001.

[LS07] Changwei Liu and Sid Stamm. Fighting unicode-obfuscated spam. *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit on - eCrime '07*, pages 45–59, 2007.

[LWWH06] Wei-hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann. Which side are you on ? identifying perspectives at the document and sentence levels. *American Heritage*, (June):109–116, 2006.

[LZ08] Yue Lu and Chengxiang Zhai. Opinion integration through semi-supervised topic modeling. *Proceeding of the 17th international conference on World Wide Web - WWW '08*, page 121, 2008.

[MYTF02] Satoshi Morinaga, Kenji Yamanishi, Kenji Tateishi, and Toshikazu Fukushima. Mining product reputations on the web. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining KDD 02*, pages 341–349, 2002.

REFERENCES

[NIK10] Tetsuji Nakagawa, Kentaro Inui, and Sadao Kurohashi. Dependency Tree-based Sentiment Classification using CRFs with Hidden Variables. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 786–794. Association for Computational Linguistics, 2010.

[OA09] Alexander Osherenko and Elisabeth Andre. *Differentiated semantic analysis in lexical affect sensing*, pages 1–6. Ieee, 2009.

[PE05] Ana-Maria Popescu and Oren Etzioni. Extracting product features and opinions from reviews. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing HLT 05*, 5(October):339–346, 2005.

[Pla98] John C Platt. *Advances in Kernel MethodsSupport Vector Learning*, 208(MSR-TR-98-14):1–21, 1998.

[PLV02] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. *Language*, 10(July):79–86, 2002.

[Sch12] Dr. Corine van Erkom Schurink. Explosive growth of Social media trends draw renewed interest in 'Sentiment Analysis' says PBT. http://alturl.com/z9hv3, 2012.

[Smi08] Justin Smith. Mapping Facebook Growth Over Time. http://www.insidefacebook.com/2008/08/19/mapping-facebooks-growth-over-time/, Aug 2008.

[Soo00] Ehsan S Soofi. Principal information theoretic approaches. *Journal of the American Statistical Association*, 95(452):1349–1353, 2000.

[SR03] C. Silva and B. Ribeiro. The importance of stop word removal on recall values in text categorization. In *Neural Networks, 2003. Proceedings of the International Joint Conference on*, volume 3, pages 1661–1666. IEEE, 2003.

[SWB06] D Sculley, G.M. Wachman, and C.E. Brodley. Spam filtering using inexact string matching in explicit feature space with on-line linear classifiers. In *The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*, number Section 2, 2006.

[Tan99] Ah-hwee Tan. Text mining: The state of the art and the challenges. In *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, pages 65–70, 1999.

[Tur02] Peter D Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. *Computational Linguistics*, 20(July):8, 2002.

[VW12] Pranjal Vachaspati and Cathy Wu. Sentiment Classification using Machine Learning Techniques, 2012.

[WA08] Gw Wang and Kj Araki. An unsupervised opinion mining approach for japanese weblog reputation information using an improved so-pmi algorithm. *Ieice Transactions On Information And Systems*, E91D(4):1032–1041, 2008.

[Whi04] Simon White. Matching Strings and Algorithms - Equivalence Methods, 2004.

REFERENCES

[WR05] Janyce Wiebe and Ellen Riloff. volume 3406, pages 486–497. Springer, 2005.

[WWH04] Theresa Wilson, Janyce Wiebe, and Rebecca Hwa. Just how mad are you? finding strong and weak opinion clauses. *Science*, 04:761–769, 2004.

[XTW10] Huosong Xia, Min Tao, and Yi Wang. Sentiment text classification of customers reviews on the Web based on SVM. *2010 Sixth International Conference on Natural Computation*, (Icnc):3633–3637, August 2010.

[Yel] Yelp. Yelp's Academic Dataset. http://www.yelp.com/academic_dataset/.

[YMTR08] Lei Yu, Jia Ma, Seiji Tsuchiya, and Fuji Ren. Opinion mining : A study on semantic orientation analysis for online document. *Text*, pages 4548–4552, 2008.

[Zha04] Harry Zhang. The Optimality of Naive Bayes. *Machine Learning*, 2004.