University of Beira Interior
Department of Computer Science

Classification of Opinionated Texts by Analogy

Sebastião Pais

A Thesis submitted to the University of Beira Interior to require the Degree of Master of Computer Science Engineering

Supervisor: Prof. Dr. Gaël Harry Adélio André Dias
University of Beira Interior

Covilhã, Portugal
August 2008

Acknowledgments

We never know where the future takes us to, but, in the present, I know I have the chance to prove myself I can do what I have always dreamed of.

In fact, I have always liked investigation work. So, I decided to do my master thesis called "Classification of Opinionated Texts by Analogy". It was not easy. I worked a lot, but I liked very much what I did. In the more difficult moments, I searched for forces around me, in the people who care about me.

I am grateful to my mother who supported me all the time, and to my girlfriend, who always was on my side to give me forces to continue and to reach my dream. I dedicate this master thesis to them, because they are the most important people in my life. I cannot forget my father and my grand-mother, who already let us, but I know both are looking to me, there, where they are . . .

Finally, I want to thank my friends, all the people of Hultig, Professors João Paulo Cordeiro and Guillaume Cleuziou for their help in this work and Dinko, who is developing his PhD thesis (Temporal Opinion Detection) in the same area as me with the same supervisor. Obviously, I also have to thank Professor Gaël Dias who gave me the taste for the area of Human Language Technology, and besides being my advisor, he revealed to be a great friend.

Thank you for everyone!

Sebastião Pais

Abstract

With the disproportionate increase of the World Wide Web and the quantity of information services and their availability, we have an excessive accumulation of documents of various kinds. Despite the positive aspects this represents and the potential this causes, a new problem arises as we need capable tools and methodologies to classify a document as to its quality.

Assessing the quality of a Web page is not easy. For the technical evaluation of the structure of Web pages, many are the works that have emerged. This thesis follows a different course. It seeks to evaluate the content of pages according to the opinions and feelings they highlight. The adopted basis criterion to assess the quality of Web pages is to examine the absence of opinions and feelings in the texts.

When we consult information from the Web, how do we know exactly that the information is reliable and does not express opinions which are made available to the public feelings?

How can we ensure when we read a text that we are not being misled by the author who is expressing his opinion or, once again, his feelings?

How can we ensure that our own assessment is free from any judgment of value that we can defend?

Because of these questions, the area of "Opinion Mining", "Opinion Retrieval", or "Sentiment Analysis", is worth being investigated as we clearly believe that there is much to discover yet.

After a lot of research and reading, we concluded that we do not want to follow the same methodology proposed so far by other researchers. Basically, they work with objective and subjective corpora manually annotated. We think it is a disadvantage because these are limited corpora, once they are small, and cover a limited number of subjects.

We disagree with another point. Some researchers only use one or several morphological classes, or specific words as predefined attributes. As we want to identify the degree of objectivity/subjectivity of sentences, and not documents, the more attributes we will have, the more accurate we expect our classification to be.

We want to implement another innovation in our method. We want to make it as automatic as possible or, at least, the least supervised as possible.

Assessed some gaps in the area, we define our line of intervention for this dissertation.

As already mentioned, as a rule, the corpora used in the area of opinions are manually annotated and they are not very inclusive. To tackle this problem we propose to replace these corpora with texts taken from Wikipedia and texts extracted from Weblogs, accessible to any researcher in the area. Thus, Wikipedia should represent objective texts and Weblogs represent subjective texts (which we can consider that is an opinion repository).

These new corpora bring great advantages. They are obtained in an automatic way, they are not manually annotated, we can build them at any time and they are very inclusive.

To be able to say that Wikipedia may represent objective texts and Weblogs may represent subjective texts, we assess their similarity at various morphological levels, with manually annotated objective/subjective corpora. To evaluate this similarity, we use two different methodologies, the Rocchio Method and the Language Model on a cross-validation basis. By using these two different methodologies, we achieve similar results which confirm our hypothesis.

With the success of the step described above, we propose to automatically classify sentences (at various morphological levels) by analogy. At this stage, we use different SVM classifiers and training and test sets built over several corpora on a cross-validation basis, to, once again, have several results to compare to draw our final conclusions.

This new concept of quality assessment of a Web page, through the absence of opinions, brings to the scientific community another way of research in the area of opinions. The user in general is also benefited, because he has the chance, when he consults a Web page or uses a search engine, to know with some certainty if the information is true or if this is only one set of opinions/sentiments expressed by the authors, excluding thus their own judgments of value about what he sees.

Resumo

Com o aumento desmedido da World Wide Web e da quantidade de serviços de informação e respectiva disponibilização, deparamo-nos actualmente com uma acumulação excessiva de textos de diversas naturezas. Apesar dos aspectos positivos que isto representa e do potencial que acarreta, surge uma nova problemática que consiste na necessidade de existirem ferramentas e metodologias capazes de classificar um documento, quanto à sua qualidade.

Avaliar a qualidade de uma página Web não é tarefa fácil. Relativamente às técnicas de avaliação da estrutura das páginas, muitos são os trabalhos que têm surgido. Esta tese segue um rumo diferente, com ela pretende-se avaliar o conteúdo das páginas segundo as opiniões e os sentimentos nelas evidenciados. O critério de base adoptado para avaliar a qualidade das páginas Web é a análise da ausência de opiniões e sentimentos nos textos.

Quando consultamos informação proveniente da Web, como sabemos exactamente que essa informação é fiável e que não retrata meras opiniões ou expressa sentimentos de quem a disponibilizou ao público?

Como podemos garantir que ao estarmos a ler um texto não estamos a ser induzidos em erro pelo seu autor que está a expressar a sua opinião ou mais uma vez os seus sentimentos?

Como podemos garantir que a nossa própria avaliação é isenta de qualquer juízo de valor que possamos defender?

Por surgirem estas perguntas, entendemos ser necessário investigar e trabalhar numa área que se denomina "Opinion Mining", "Opinion Retrieval", ou ainda "Sentiment Analysis", onde julgamos existir muito ainda por descobrir.

Depois de muita pesquisa e leitura sobre a área em discussão, concluímos que não queríamos seguir a mesma metodologia que outros seguem. Basicamente trabalham com corpora objectivos e corpora subjectivos anotados de forma manual. Pensamos que é uma desvantagem, porque esses corpora são limitativos, uma vez que são pequenos e por isso abrangem um número restrito de assuntos.

Outro aspecto acerca do qual discordamos é que alguns investigadores utilizam apenas uma(s) classe(s) morfológica(s), ou palavras predefinidas como características. Como queremos identificar frases, e não só textos, quanto mais características tivermos, mais exacta deverá ser a nossa classificação.

Uma outra inovação que queremos implementar é tornar o nosso método o mais automático possível ou, pelo menos, o menos supervisionado possível.

Avaliadas algumas lacunas existentes na área, definimos a nossa linha de intervenção para a realização desta dissertação.

Como já foi mencionado, por norma, os corpora utilizados na área das opiniões são anotados manualmente e pouco abrangentes. Para combatermos esse problema propomos que para substituir esses mesmos corpora podemos utilizar textos extraídos do Wikipedia e textos extraídos de Weblogs, acessíveis a qualquer investigador na área. Deste modo, o Wikipedia representa textos objectivos e os Weblogs representam textos subjectivos (que podemos considerar que são um repositório de opiniões).

Estes novos corpora por nós definidos trazem grandes vantagens: são obtidos de forma automática, não são anotados manualmente, podemos construí-los em qualquer altura, para qualquer língua e são bastante abrangentes.

Para podermos afirmar que o Wikipedia representa textos objectivos e que os Weblogs representam textos subjectivos, avaliamos a sua similaridade, a vários níveis morfológicos, com os corpora (objectivos/subjectivos) anotados manualmente. Para avaliar essa similaridade, utilizamos duas metodologias diferentes, o Método de Rocchio e o Modelo da Linguagem, usando em ambos conjuntos de treino e de teste de todos os corpora e o conceito de validação cruzada. Ao utilizarmos estas duas metodologias diferentes, obtivemos resultados diferentes, que foi necessário compararmos para tirarmos as nossas conclusões, que resultaram na aprovação da nossa hipótese.

Com o sucesso do passo acima descrito, passamos à classificação de frases (também a vários níveis morfológicos) que podem conter poucas ou muitas palavras. Nesta fase, utilizamos vários classificadores SVM, conjuntos de treino e de teste dos vários corpora e o conceito de validação cruzada, para mais uma vez podermos ter vários resultados que comparamos para tirar as nossas conclusões.

Este novo conceito de avaliação da qualidade de uma página Web, através da ausência de opiniões, traz à comunidade científica um outro caminho de investigação na área das opiniões. O utilizador em geral também é beneficiado, pois tem a possibilidade de, ao consultar uma página Web ou efectuar uma pesquisa num motor de busca, saber com alguma certeza se a informação que visualiza é verídica ou se é apenas um conjunto de opiniões/sentimentos expressos pelos autores, excluindo, desta forma, os seus próprios juízos de valor acerca do que está a visualizar.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Natural Language Processing

We cannot imagine a world without communication. Every living being must communicate to survive. For us, human beings, language is a fundamental aspect and it is a crucial component of our life. In written form it serves as a long-term record of knowledge from one generation to the next. In spoken form it serves as our primary means of coordinating our day-to-day behavior with others. So, producing language is above all a social activity.

Natural Language [1] [2] requires many kinds of expertise: knowledge of the domain (what to say, relevance), knowledge of the language (lexicon, grammar, semantics), strategic rhetorical knowledge (how to achieve communicative goals, text types, style), etc. Moreover, building successful Natural Language systems requires engineering knowledge (how to decompose, represent and orchestrate the processing of all this information) as well as knowledge about the habits and constraints of the end user as an information processor (sociolinguistic and psychological factors).

Every user of Natural Language uses them in different ways. We consider Natural Language systems as a help to unravel the mystery how language works in our mind. Others see Natural Language systems as an approach to solving practical problems - such as contributing to the synthesis side of machine translation, to text summarization and to multilingual presentation of information in general.

Producing language is a knowledge-intensive, flexible and highly context-sensitive process. This context sensitivity reveals itself best when we consider connected texts rather than isolated sentences.

The objective of various researches is to create computational models of language and specify models that approach human performance in the linguistic tasks of reading, writing, hearing and speaking. These researches are concerned with the processes of comprehending and using language once the words are recognized. Computational models are useful for scientific purposes (for exploring the nature of linguistic communication) and for practical purposes (for enabling effective human-machine communication).

Language is studied in several different academic disciplines. Each one defines its own set of problems and has its own methods for addressing them. The linguist studies the structure of language itself; the psycholinguist studies the processes of human language production and comprehension; the philosopher considers how words can mean anything at all and how they identify objects in the world. The objective of computational linguist is to develop a computational theory of language, using the notions of algorithms and data structures from computer science. Obviously, to build a computational model, we must take advantage of the knowledge of the others disciplines.

There are two motivations for developing computational models. The scientific motivation is to obtain a better understanding of how language works. The other disciplines do not have the tools to completely address the problem of how language comprehension and production work. Even if we combine all the different approaches, a comprehensive theory would be very complex to be studied using traditional methods. But we may be able to realize such complex theories as computer programs, then we test them by observing how they perform and we can improve them where they fail. Psycholinguist can explore specific predictions about human behavior that computational models may provide. If we continue this process, we may eventually acquire a deep understanding of how human language processing occurs. To realize this dream, it is necessary to combine the efforts of linguists, psycholinguists, philosophers and computer scientists. This common objective creates a new area of interdisciplin ary research called cognitive science.

The practical, or technological, motivation is that Natural Language processing capabilities would revolutionize the way computers are used. Since most of human knowledge is recorded in linguistic form, computers that could understand Natural Language could access all this information. In addition, Natural Language interfaces to computers would allow complex systems to be accessible to everyone. Such systems would be significantly more intelligent and flexible than is possible with current computer technology. For technological purposes is more important that this works that the model used reflects the way humans process language.

James Allen defends a middle ground between the scientific and technological objectives. He believes that Natural Language is so complex that an ad hoc approach without a well-specified underlying theory will not be successful. Thus the technological objective cannot be realized without using sophisticated underlying theories on the level of those being developed by linguists, psycholinguists and philosophers. Besides, the present state of knowledge about Natural Language processing is so preliminary that attempting to build a cognitively correct model is not feasible. The objective of his book is to describe work that aims to produce linguistically motivated computational models of language understanding and production that can be shown to perform well in specific example models.

1.2 Applications of Natural Language Processing

According to James Allen, a good way to define Natural Language research is to consider the different applications that researchers work on. The applications can be divided into two major classes: text-based applications and dialogue-based applications.

Text-based applications involve the processing of written text, such as books, newspapers, reports, manuals, e-mail messages, etc. These are all reading-based tasks. Text-based Natural Language research is ongoing in applications such as finding appropriate documents on certain topics from a database of texts (for example, finding relevant books in a library); extracting information from messages or articles on certain topics (for example, building a database of all stock transactions described in the news on a given day); translating documents from one language to another (for example, producing automobile repair manuals in many different languages) and summarizing texts for certain purposes (for example, producing a 3-page summary of a 1000-page government report).

Dialogue-based applications involve human-machine communication. Typical potential applications include question-answering systems, where Natural Language is used to query database (for example, a query system to a personnel database); automated customer service over the telephone (for example, to perform banking transactions or order items from a catalogue); tutoring systems, where the machine interacts with a student (for example, an automated mathematics tutoring system); spoken language control of a machine (for example, voice control of a VCR or computer) and general cooperative problem-solving systems (for example, a system that helps a person plan and schedule freight shipments).

The objective of this master thesis, placed in the category of text-based applications, is to learn a classifier which evaluates the degree of objectivity or subjectivity of sentences. The corpora used up to now by the Scientific Community are manually annotated and so limited as they are small and cover a little number of subjects. We propose to replace these manual resources by corpora extracted from Wikipedia and Weblogs, automatically classified as objective corpora and subjective corpora respectively. Once we want to classify sentences and not only texts, we work with several classes of words, because it is necessary to have a great number of features.

1.3 Motivation

Opinions and sentiments are a very ample area, which is not still deeply explored. This means, there is a lot of investigations to do. So, this work pretends to contribute to the development of this interesting area.

It is necessary to point out that there are already important investigations that allowed to realize this work. In the same way, we hope that this work will be used by other investigators and propose new research directions .

At present, there are many sets of supervised documents, separated in objective and subjective, which allow the classification of texts. In particular, we can mention the Subjectivity v1.0[1] corpus built by Pang and Lee [3]. Based on our motivation, we compare them to Wikipedia and Weblogs corpora so that we can automatically build objective/subjective corpora by analogy. If we can prove that Wikipedia and Weblogs corpora are respectively comparable to the objective and subjective part of the Subjectivity v1.0 corpus, we aim at learning classifiers from these new resources. As a consequence, we pretend to propose a new "unsupervised" methodology for classification usefull for many languages, domains and genre.

1.4 Contribution

As we already said, this work contributes to define the quality of a Web page by absence of opinions or sentiments. In future work, we want to increase this issue to other techniques, like structure of the page or syntactic difficulty.

Up to now, supervised methodologies are usually used in this area. That means there are defined corpora (a set of texts manually classified like objective or subjective) used to classify texts. These corpora limit the classification as they are very small and present a little set of subjects. Our objective is to make a method as less supervised as possible, with corpora extracted for Wikipedia and Weblogs, which would give us a better chance to classify texts or sentences.

Our automatic method classifies not only texts, but simple sentences, which can have a little number of words. So, we must accept bigger numbers of possible attributes and use different morphological classes. In fact, the methods used so far privilege a specific morphological class of words: adjectives.

Finally, we must point out that our work will be exported to an ongoing project (VIP-ACCESS[2]), which we want to improve. Indeed, it aims at

---

[1] http://alias-i.com/lingpipe/demos/tutorial/sentiment/read-me.html

[2] VIP-ACCESS - Ubiquitous Web Access for Visually Impaired People, project funded by the Portuguese Foundation for Technology and Science with Reference PTDC/PLP/72142/2006.

developing an educational meta-search engine for which the great objective is the retrieval of high quality results.

1.5 Our Methodology

As already mentioned in this dissertation, we will adopt a new methodology, which we believe to be of great value to the Scientific Community working in the area of opinions or feelings.

Corpora used by Scientific Community are limited. As such, our first concern is to find a solution to this problem.

Therefore, instead of using objective/subjective corpora, we propose to replace them with automatically built corpora which consist of texts taken from Wikipedia and Weblogs.

To do so, we first need to make an assessment of similarity between the corpora. So, we compare the objective part of the Subjectivity v1.0 corpus with a sample of Wikipedia and the subjective part of the Subjectivity v1.0 corpus with a sample of Weblogs automatically extracted from the Web.

After assessing the similarity between corpora using the Rocchio Method and Language Modeling, we can propose the classification of sentences using various Support Vector Machine classifiers (i.e. with different kernels) based on two types of training and test sets (i.e. manually annotated and automatically built). Finally, we compare the results and interpret them to draw conclusions.

1.6 Plan of the Thesis

Chapter two refers to some work done in this area describing some already developed and used methodologies.

In chapter three, we describe the process of constructing the different corpora and the way we sample them (the Wikipedia corpus and the Weblogs corpus) to successfully achieve the objective of this thesis.

Regarding chapter four, we relate the whole process involving the evaluation of similarity between corpora i.e. methodologies, results and conclusions.

Chapter five presents the whole process for the classification of sentences where we use several Support Vector Machine classifiers with two different training and test sets.

Finally, chapter six concludes the work and proposes possible future improvements and research directions.

Chapter 2

Related Work

Natural Language is a very extensive area, still with a lot to discover. Along the last years, we assisted to a great deal of researches in the related areas of opinion extraction, sentiment analysis, semantic orientation and polarity classification, and subjectivity analysis.

For this work, it was necessary to do a research thorough all the existing publications in this area. In this section, we will highlight what we consider more relevant for the development of our work.

2.1 Information Retrieval

The meaning of the term Information Retrieval [4] [5] [6] can be very extensive. Just getting a credit card out of our wallet so that we can type in the card number is a form of Information Retrieval. However, in academic context, Information Retrieval might be defined thus: Information Retrieval is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually on local computer servers or on the internet).

Information Retrieval consists of retrieving information from stored data through queries formulated by the user or through preformulated user profiles. This information can be in any medium or format, e.g. text, image, video, speech, databases, and often combines media. The field of Information Retrieval has a well-established history, and has already reached an initial level of maturity that is deployed in industry and business. Recently and with the advent of the World Wide Web, the need for effective Information Retrieval techniques has reached the public in an unprecedented way. Whereas in past years, Information Retrieval was primarily required by subspecialties, such as business, law, and medicine, now users who simply want effective Internet searching are pushing the research community to solve information-finding needs. Increasing network transmission speed and capacity promise to bring even more impetus to this field. Finally, globalization adds yet another dimens ion to the need for powerful Information Retrieval across languages.

The first stage in the Information Retrieval task is to determine the limits of the corpus over which subsequent processing will be performed. In the case of specialized corpora, document collection properties can affect performance. For example, a corpus of email messages has different properties from a corpus transcribed speech. Furthermore, corpora could consist of single-subject documents, such as medicine or law, in which case the use of metadata as well as full text data can be helpful. This also holds for semi-structured text, such as memoranda or text associated with databases. Other collections could be selected based on user group, e.g. articles for medical providers across topics, or articles for patients. When highly heterogeneous corpora, such as web articles across sites, are to be processed, techniques geared towards generic documents are preferred. From the perspective of computational linguistics, key research areas include tailored processing for specific domains, user groups, genres, or other text characteristics; complementary research includes dealing with widely heterogeneous text, coupled with image, sound, and video retrieval.

Classical models in Information Retrieval consider that each document can be modeled by a set of indexed keywords. In principle, keywords give an indication of the document content. Before indexing, several text operations are performed in order to save space at indexing time. Among the main parts of speech (e.g. determiners, nouns, adjectives, verbs, adverbs, pronouns, etc.), content words, such as nouns, verbs, and adjectives, are the ones which carry most of the semantics, whereas function words such as prepositions, pronouns, and determiners have less impact on the determining of what an article is about. Thus, function words are often ignored when constructing representations. Furthermore, function words tend to be frequent, so eliminating them also contributes to efficiency, but researches tend to disagree about the impact of keeping or removing them. Among the document preprocessing tasks are: elimination of function words, stemming, which for English consists of stripp ing the end of words generally morphologically related to their common stem or root, see selection of index terms, and the representation of synonymic or taxonomic relations. Controversies about stemming techniques vis-à-vis collection types and language types still have not been resolved. In order to capture relationships between words (e. g. walk, walks, walking for English), stemming is applied. Stemming conflates morphologically related words to the same root, either by a traditional stemmer such as Porter's or Lovins's, or by a linguistically based morphological analyzer. The former tends to be efficient and robust but prone to error, whereas rule-based linguistic analyzers tend to be more accurate.

Two main issues are at stake in using stemming. The first involves the concepts of recall and precision. Recall and precision are two independent matrices traditionally used to assess the performance of Information Retrieval systems. Recall refers to the percentage of relevant documents that are classified as relevant, and precision refers to the percentage of documents classified as relevant which were correctly classified. In practice, most systems exhibit a trade-off between recall and precision: an improvement along one dimension typically forces a drop in performance along the other dimension. Depending on the target application, it may be desirable to buy high precision at the price of recall, or one may prefer to optimize recall and settle for low precision. A general Information Retrieval controversy lies in the fact that while stemming increases recall, it has a negative effect on precision. Second, two main errors occur while stemming: under stemming and over stemm ing. Over stemming is caused by relating forms that are not in fact morphologically related, for example magnesia, magnesium, magnet, magnetic, etc. are conflated by Lovins's stemming algorithm and reduce to one common stem magnes. On the other hand, under stemming is the non-conflation of related words, such as acquire, acquiring, acquired, and acquisition. The first three are correctly related to acquire, but the stem for acquisition is acquis. Thus, acquisition and acquire are not modeled as related.

Different studies have shown inconsistent results of the effect of using stemmers. Harman [7] showed that stemming provides no improvement over no stemming at all, and that different stemming algorithms do not affect performance. Krovetz [8] showed that, with stemming, improvement ranges between 1.3% and 45.3% for different test collections and stemmers. Frakes and Baeza-Yates [9] compare eight distinct studies and they all conclude that there are beneficial aspects of using stemming techniques. A large-scale analysis by Hull [10] compared five algorithms for stemming: s.-plural removal, Porter [11], Lovins [12], and two linguistic algorithms for inflectional and derivational morphology. The conclusion was that stemming always helped, but the improvements range from just 1% to 3%.

Once the index terms are determined for a document, it is clear that not all terms are relevant to the document content. In other words, if the same word appears in the10.000 documents that form the collection, this index term is nearly useless in that it does not discriminate one document over another, and thus may not satisfy the user's request. On the other hand, if an index term appears only in five documents of the collection, it is very likely a discriminating term for a given query. Therefore, assigning a weight to index terms provides a discriminatory value.

2.2 Machine Learning

Machine learning [5] [6] [13] is another important concept in our work which needs to be explained.

What is learning anyway? What is Machine Learning? These are philosophic questions, and we will not be much concerned with philosophy in this work; our emphasis is firmly on the practical. However, it is worth spending a few moments at the outset on fundamental issues, just to see how tricky they are, before rolling up our sleeves and looking at Machine Learning in practice. We define "to learn" as follows:

- To get knowledge of by study, experience, or being taught;
- To become aware by information or from observation;
- To commit to memory;
- To be informed of, ascertain;
- To receive instruction.

These meanings have some shortcomings when it comes to talking about computers. For the first two, it is virtually impossible to test whether learning has been achieved or not. How do we know whether a machine has got knowledge of something? We probably cannot just ask these questions. Even if we could, we would not be testing its ability to learn but would be testing its ability to answer questions. How do we know whether it has become aware, or conscious, is a burning philosophic issue.

As for the last three meanings, although we can see what they denote in human terms, merely "committing to memory" and "receiving instruction" seem to fall far short of what we might mean by Machine Learning. They are too passive and we know that computers find these tasks trivial. Instead, we are interested in improvements in performance, or at least in the potential for performance, in new situations. We can "commit something to memory" or "be informed of something" by rote learning without being able to apply the new knowledge to new situations. We can receive instruction without benefiting from it at all.

This ties learning to performance rather than knowledge. You can test learning by observing the behavior and comparing it with past behavior. This is a much more objective kind of definition and appears to be far more satisfactory.

But there is still a problem. Learning is a rather slippery concept. Lots of things change their behavior in ways that make them perform better in the future, yet we would not want to say that they have actually learned. A good example is a comfortable slipper. Has it learned the shape of our foot? It has certainly changed its behavior to make it perform better as a slipper! Yet we would hardly want to call this learning. In everyday language, we often use the word "training" to denote a mindless kind of learning. We train animals and even plants, although it would be stretching the word a bit to talk of training objects such as slippers that are not in any sense alive. But learning is different. Learning implies thinking. Learning implies purpose. Something that learns has to do so intentionally. That is why we would not say that a vine has learned to grow round a trellis in a vineyard. We would say it has been trained. Learning without purpose is merely training. Or, more to the point, in learning the purpose is the learner's, whereas in training it is the teacher's.

Thus on closer examination the second definition of learning, in operational, performance-oriented terms, has its own problems when it comes to talking about computers. To decide whether something has actually learned, you need to see whether it intended to or whether there was any purpose involved. That makes the concept moot when applied to machines because whether artifacts can behave purposefully is unclear. Philosophic discussions of what is really meant by "learning", like discussion of what is really meant by "intention" or "purpose", are fraught with difficulty. Even courts of law find intention hard to grapple with.

2.3 Learning Subjective and Objective Language

Most of the research so far has been handled at the adjective level.

Wiebe [14] performs a statistical analysis, which shows that adjectives are correlated with subjective sentences. If there is at least one adjective in the sentence, there is 56% of probability of the sentence being subjective, even thinking there are more objective than subjective sentences in the corpus.

Turney and Littman [15] start to define sets of objective and subjective adjectives. By doing so, they also determine the orientation of a word, based on Pointwise Mutual Information and Latent Semantic Analysis.

Whitelaw et al. [16] propose a method for heuristically extracting adjectival appraisal groups, which consist of an appraising adjective (e.g. "beautiful") and optional modifiers (e.g. "very"). They developed a number of taxonomies of appraisal attributes by semi-automatically classifying 1329 adjectives and modifiers.

Adjectives are not the only parameter we must consider in a research on subjective language. Bethard et al. [17] propose to work with propositional phrases and experiment using SVMs to extract them.

Wiebe et al. [18] obtain a variety of subjectivity cues (frequencies of unique words in subjective-element data; collocations with one or more positions filled by a unique word; distributional similarity of verbs and adjectives) from corpora and show their effectiveness on classification tasks. They determine a relationship between low frequency words and subjectivity and they discover that their method for extracting subjective n-grams is improved by examining those that occur with unique words.

Esuli and Sebastiani [19] use their gloss definitions from online dictionaries to present a semi-supervised method for the semantic orientation identification of words. A seed set of words with positive and negative connotation is composed manually and provided as input, which is expanded with the words synonyms from a online dictionaries. A text classifier is trained to predict the polarity of words on the basis of their glosses.

Kim and Hovy [20] propose a method to automatically expand a set of seed words (verbs and adjectives) which are manually tagged as having positive, negative and neutral polarity. Their synonyms are identified using WordNet of Miller [21]. They recognize that synonyms of a word may not have the same polarity and proposed a method to calculate the closeness of a synonym to each polarity category to determine the most probable one. The method was evaluated on a corpus of German emails and achieved 77% accuracy on verbs and 69% on adjectives.

In order to determine adjectives polarity, Chesley et al. [22] present a method using verbs information and an online resource i.e. the Wikipedia dictionary. They use verb-class information in the sentiment classification task, since exploiting lexical information contained in verbs has shown to be a successful technique for classifying documents.

There are many interesting works about sentiment polarity using reviews as resource. Pang et al. [23] evaluate several machines learning algorithms to classify film reviews as either containing positive and negative opinions. Turney [24] proposes an unsupervised algorithm for classifying reviews as positive or negative. This method, evaluated on 410 reviews, showed accuracy between 66% and 84% depending on the domain. Hu and Liu [25] propose a method of identification of frequent features of a specific review item, and finding opinion words by extracting adjectives most proximate to the terms representing frequent features.

Yi et al. [26] propose to extract positive and negative opinions about specific features of a topic.

Hurst and Nigam [27] propose a method of identifying sentences that are relevant to a topic and express opinion on it. To determine if a document is relevant to a topic, they use a Machine Learning approach (Winnow classifier). For the sentences predicted topically relevant, they apply sentiment analyser. They evaluated their classification method on a set of 982 messages from online resource and their evaluation results show overall precision of 72%.

To conclude, we can say that many of the methodologies intend to identify subjectivity in the texts, based on verbs and adjectives.

Chapter 3

Construction of Corpora

In our methodology we use two types of corpora. On one side, we work with manually annotated corpora for which we know exactly what the classification of a text (objectivity or subjectivity) is. These corpora have been reviewed, approved and used by the Scientific Community.

On the other side, we build two corpora through an automatic process that is not supervised, also divided into objectivity corpus and subjectivity corpus based on the common sense feeling that Wikipedia texts are more objective than subjective and Weblogs are more subjective than objective.

For many applications, corpus data are the raw fuel of Natural Language Processing, and/or the test bed on which a Natural Language Processing application is evaluated. In this chapter the history of corpus linguistics is briefly considered. Following on from this, corpus annotation is introduced as a prelude to a discussion of some of the uses of corpus data in Natural Language Processing. But before any of this can be done, we need to define a corpus.

3.1 What is a Corpus?

A corpus [6] [28] (pl. corpora) is simply described as a large body of linguistic evidence typically composed of attested language use. One may contrast this form of linguistic evidence with sentence created not as a result of communication in context, but rather upon the basis of metalinguistic reflection upon language use, type of data common in the generative approach to linguistics. Corpus data is not composed of the ruminations of theorists. It is composed of such varied material as everyday conversations (e.g. the spoken section of the British National Corpus[1]), radio news broadcasts (e.g. the IBM/Lancaster Spoken English Corpus[2]), published writing (e.g. the major-

---

[1] http://www.natcorp.ox.ac.uk/

[2] http://icame.uib.no/lanspeks.html

ity of the written section of the British National Corpus[3]) and the writing of young children (e.g. the Leverhulme Corpus of Children's Writing[4]). Such data are collected together into corpora which may be used for a range of research purposes. Typically these corpora are machine readable-trying to exploit a paper-based linguistic resource or audio recording running into millions of words is impractical. So while corpora could be paper based, or even simply sound recording, the view taken here is that corpora are machine readable.

Corpora have uses in both linguistics and Natural Language Processing, and are of interest to researchers from other disciplines, such as literary stylistics. Corpora are multifunctional resources.

With this started, a slightly more refined definition of a corpus is needed than that which has been introduced so far. It has been established that a corpus is a collection of naturally occurring language data. But is any collection of language data from three sentences to three million words of data a corpus? The term corpus should properly only be applied to a well-organized collection of data, collected within the boundaries of a sampling frame designed to allow the exploration of a certain linguistic feature (or set of features) via the data collected. A sampling frame is of crucial importance in corpus design. Sampling is inescapable. Unless the object of study is a highly restricted sublanguage or a dead language, it is quite impossible to collect all of the utterances of a natural language together within one corpus. As a consequence, the corpus should aim for balance and representativeness within a specific sampling frame, in order to allow a particular variety of language to be studied or modeled. The best way to explain these terms is via an example. Imagine that a researcher has the task of developing a dialogue manager for a planned telephone ticket selling system and decides to construct a corpus to assist in this task. The sampling frame here is clear-the relevant data for the planned corpus would have to be drawn from telephone ticket sales. It would be quite inappropriate to sample the novels of Jane Austen or face-to-face spontaneous conversation in order to undertake the task of modeling telephone-based transactional dialogues. Within the domain of telephone ticket sales there may be a number of different types of tickets sold, each of which requires distinct questions to be asked. Consequently, we can argue that there are various linguistically distinct categories of ticket sales. So the corpus is balanced by including a wide range of types of telephone ticket sales conversations within it, with the types organized into coherent subparts (for example, train ticket sales, plane ticket sales, and theater ticket sales). Finally, within each of these categories there may be little point in recording one conversation, or even the conversations of only one operator taking a call. If one records

---

[3] http://www.natcorp.ox.ac.uk/

[4] http://www.lancs.ac.uk/fass/projects/lever/index.htm

only one conversation it may be highly idiosyncratic. If one records only the calls taken by one operator, one cannot be sure that they are typical of all operators. Consequently, the corpus aims for representativeness by including within it a range of speakers, in order that idiosyncrasies may be averaged out.

3.2 A History of Corpus Linguistics

Outlining a history of corpus linguistics is difficult. In its modern, computerized, form, the corpora have only existed since the late 1940s. The basic idea of using attested language use for the study of language clearly pre-dated this time, but the problem was that the gathering and use of large volumes of linguistic data in the pre-computer age was so difficult as to be almost impossible. There were notable examples of it being achieved via the deployment of vast workforce (Kaeding is a notable example of this). Yet in reality, corpus linguistics in the form that we know it today, where any PC user can, with relative ease, exploit corpora running into millions of words, is a very recent phenomenon.

The crucial link between computers and the manipulation of large bodies of linguistic evidence was forgotten by Bussa in the late 1940s. During the 1950s the first large project in the construction of comparable corpora was undertaken by Juilland, who also articulated clearly the concepts behind the ideas of the sampling frame, balance, and representativeness. English corpus linguistic took off in the late 1950s, with work in America on the Brown Corpus[5] and work in Britain on the Survey of English Usage. Work in English corpus linguistics in particular grew throughout the 1960s, 1970s, and 1980s, with significant milestones such as a corpus of transcribed spoken language, a corpus with manual encodings of parts-of-speech information, and corpus with reliable automated encodings of parts of speech being reached in this period. During the 1980s, the number of corpora available steadily grew as did the size of those corpora. This trend became clear in the 1990s, with corpora such as the British National Corpus and the Bank of English reaching vast sizes (100,000,000 words and 300,000,000 words of modern Britsh English respectively) which would have been for all practical purposes impossible in the pre-electronic age. The other trend that became noticeable during the 1990s was the increasingly multilingual nature of corpus linguistics, with monolingual corpora becoming available for a range of languages, and parallel corpora coming into widespread use.

In conjunction with this growth in corpus data, fueled in part by expanding computing power, came a range of technical innovations. For example, schemes for systematically encoding corpus data came into be-

---

[5] http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/corpora/list/private/brown/brown.html

ing, programs were began in earnest to represent the audio recording of a transcribed spoken corpus text in tandem with its transcription. The range of future developments in corpus linguistics is too numerous to mention in detail here. What can be said, however, is that as personal computing technology develops yet further, we can expect that research questions not addressable with corpus data at this point of time will become possible, as new types of corpora are development, and new programs to exploit these new corpora are written.

One area which has only been touched upon here, but which has been a major area of innovation in corpus linguistics in the past and which will undoubtedly remain so in the future, is corpus annotation. In the text section corpus annotation will be discussed in some depth, as it is an area where corpus linguistics and Natural Language Processing interact.

3.3 The Subjectivity v1.0 Corpus

The Subjectivity v1.0 corpus [3] contains 5000 subjective and 5000 objective sentences collected from movie reviews data for objective sentences and customer review snippets for subjective sentences. To gather subjective sentences (or phrases), 5000 customer review snippets (e.g., bold, imaginative, and impossible to resist) were collected from the url http://www.rottentomatoes.com. To obtain (mostly) objective data, they took 5000 sentences from plot summaries available from the Internet Movie Database[6].

3.4 Wikipedia Corpus

To build the Wikipedia corpus we downloaded the English static version of Wikipedia in XML format (freely available at http://download.wikimedia.org/enwiki/20071018). We extracted all the sentences from this file giving rise to a corpus of 40Gb of text divided into several files.

3.5 Weblogs Corpus

The Weblogs corpus was built differently than the previous, because there are no repository where one could download several Weblogs. To overcome this problem, we implemented a spider for which we gave an initial domain and downloaded all Weblogs from this domain (see appendix A). After spidering the domain and extracting all sentences from it, we obtained a corpus of texts with 12Gb divided into several files.

---

[6] http://www.imdb.com

3.6 Sampling

Based on our assumption, we have the objectivity corpus and the subjectivity corpus with sentences respectively extracted from Wikipedia and Weblogs. Yet, we still do not have our final corpora (Wikipedia/Weblogs) due to the overwhelming amount of texts.

For efficiency reasons, we decided to use a significant random sample of the corpora, maintaining statistical significance.

But how do we guarantee that the sample is properly significant? How is it possible to determine the amount of enough data to be part of the sample?

Because, in our case, very large samples delay the process as the data is too big, it is impossible to draw conclusions in useful time. On the other side, too small samples may induce in error, or we will have imprecise results.

To solve this problem, we will determine the error when we test several dimensions of the necessary sample to continue with our methodology. That error can be seen as the maximum difference between the sample and our total volume of data.

The error is represented as in [29] by the equation 3.1 where n is the size of the sample, p, the probability of observing one event from the probability distribution, d the maximum diameter of the allowed error and $Z_(\alpha/2)$ the inverse of the normal distribution, being α the significance level, that is additional to the trust level.

$$d = \sqrt{\frac{p(1-p)(Z_{\alpha/2})^2}{n}} \tag{3.1}$$

This equation can also be rewritten as in equation 3.2. In particular, for our test case, we will use $\alpha = 0.001$, and $p = 0.5$ which gives us a trust of 99.9%.

$$n = p(1-p)\left[\frac{Z_{\alpha/2}}{d}\right]^2 \tag{3.2}$$

Presented the equation and its parameters, we vary the dimension of our sample to obtain a significant and representative sample of our total volume of data.

As shown in many studies of referenced corpora, on average, a word has five characters, which corresponds to five bytes. As a consequence, we know that:

- 1 Megabyte ≈ 209 715 Words;
- 100 Megabytes ≈ 20 971 500 Words;

- 1 GigaByte ≈ 214 748 365 Words;
- 10 GigaBytes ≈ 2 147 483 650 Words.

For these possible dimensions of the sample, we obtain the following errors based on equation 3.2:

- 1 Megabyte ≈ 0.00359264;
- 100 Megabytes ≈ 0.0003592637;
- 1 GigaByte ≈ 0.0001122699;
- 10 GigaBytes ≈ 0.0000355028.

We concluded that we could use a dimension sample equal to 100 Megabytes to build the Wikipedia corpus and the Weblogs corpus. Indeed, a deviation of 0.0003592637 from the mean of the normal distribution $\mathcal{N}(0, 1)$ is negligible, given us confidence that the sample is representative of our corpus.

With the Subjectivity v1.0 corpus and the samples of the Wikipedia corpus and the Weblogs corpus, we now study the characteristics of each corpus to assess their difference and evidence to what extent our methodology may improve classification by gathering more information as shown in table 3.1[7].

Corpora	Wikipedia	Weblogs	Objectivity	Subjectivity
Unique Sentences	411 293	984 682	5 000	5 000
Unique Words	224 112	79 680	15 065	14 146

Table 3.1: *Dimensions of the Corpora.*

---

[7] Objectivity (resp. Subjectivity) refers to the set of 5000 objective (resp. subjective) sentences from the Subjectivity v1.0 corpus.

Chapter 4

Similarity between the Corpora

Our purpose in this chapter is to present our methodology which allows comparing the Subjectivity v1.0 corpus to the Wikipedia, the Weblogs and the Reuters[1] corpora in terms of similarity.

4.1 The Vector Space Model

In vector space models [28], documents and queries or sentences are represented in terms of vectors. In vector space models, non-binary weights are assigned to index terms in queries and documents. Eventually, these term weights are used to compute a degree of similarity between documents and queries or sentences, thus returning information that goes from full match to partial match to the user's request. Index term weight can be computed using many different techniques. In our implementation of the vector space model, similarity is computed by using the TF/ISF weight through the Cosine similarity measure.

4.2 The TF/ISF Weight

In the vector space model, we first need to attribute a weight to the words of any single sentence. These words can also be called features. The measure used to attribute a weight [28] to each sentence feature is adapted from the well-known TF*IDF, and we call it the TF/ISF:

$$w_{ij} = tf_j \times \log_2(\frac{N}{n}) \tag{4.1}$$

---

[1] In 2000 Reuters released a corpus of Reuters News stories for use in research and development of natural language-processing, information-retrieval or machine learning systems. Reuters stopped distributing the corpus in 2004. Instead, the Reuters corpus is now available from NIST, the National Institute of Science and Technology. Application forms are available at http://trec.nist.gov/data/reuters/reuters.html.

where:

- w_{ij} is the weight of the word T_j in the sentence i,
- tf_j is the frequency of the word T_j in the corpus,
- N is the number of sentences in the text,
- n is the number of sentences where the word T_j occurs at least once.

4.3 The Cosine Measure

The Cosine similarity measure [28] has extensively been used in Information Retrieval within the framework of the vector space model. The Cosine measure is defined as follows:

$$CosSim(d_j, q) = \frac{\overrightarrow{d_j}.\overrightarrow{q}}{|\overrightarrow{d_j}|.|\overrightarrow{q}|} = \frac{\sum_{i=1}^{t}(w_{ij}.w_{iq})}{\sqrt{\sum_{i=1}^{t} w_{ij}^2 . \sum_{i=1}^{t} w_{iq}^2}} \tag{4.2}$$

This is the scalar product of the normalized unit-length vectors produced by the *weight* values. For schema, the joint operation can be interpreted as the product of the two corresponding *weight* values, whereas the norm factor is the product of the lengths of the two *weight* vectors. When combined these two operations entail a high Cosine value if corresponding *weight* values are overall similar (obtaining the effect of a comparison), giving a higher impact for larger *weight* values. In the following figure, we give an example to explain the Cosine similarity measure.

$D_1 = 2T_1 + 3T_2 + 5T_3$
$D_2 = 3T_1 + 7T_2 + 1T_3$
$Q = 0T_1 + 0T_2 + 2T_3$
$CosSim(D_1, Q) = 10/\sqrt{(4+9+25)(0+0+4)} = 0.81$
$CosSim(D_2, Q) = 2/\sqrt{(9+49+1)(0+0+4)} = 0.13$

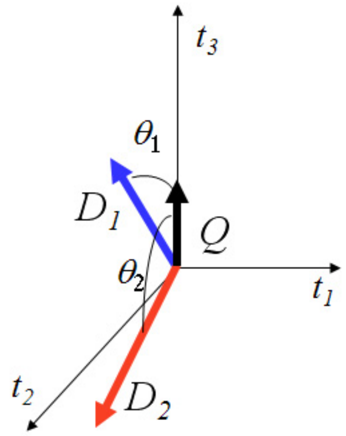


Figure 4.1: *Graph that illustrates the Cosine Measure.*

With the example we identify that Q is more similar to D_1 than D_2.

4.4 Methodologies

4.4.1 Rocchio Method

Text categorization is the process of grouping documents in different categories or classes. The amount of information available online grows every day, so the need for reliable automatic text categorization has increased.

Relevance feedback methods can be adapted for the categorization of texts. Rocchio [5] relevance feedback algorithm is one of the most popular and widely used learning methods from Information Retrieval. This algorithm uses standard TF/IDF weighted vectors - also called Vector Space Model - to represent text documents (normalized by the maximum term frequency). For each category it computes a prototype vector, summing the vectors of the training documents in the same category. It also assigns test documents for the category with the closest prototype vector, based on a similarity measure (e.g. the Cosine similarity) that calculates the cosine of the angle between two vectors. This is the most used similarity measure in the vector space classification and the one we have used in our work as well. Notice that Rocchio's relevance feedback is designed to distinguish two classes: relevant and not relevant.

In the vector space model, each document is a vector with one component for each term. A set of documents is a set of points in the dimensional space. In this classification, a basic hypothesis is used: the contiguity hypothesis in which documents of the same class form a contiguous region in the space. In the Rocchio classification (vector space classification algorithm), the documents are represented as points in N dimensions. We can interpret these points as end points of normalized document vectors in three dimensions. The main work we must do in the vector space classification is to define the boundary lines, since they determine the classification decision.

4.4.2 Language Model

Language Models [2] [6] are used in speech recognition to estimate the probability of word sequences. Grammatical constraints can be described using a context-free grammar (for small to medium-size vocabulary tasks these are usually manually elaborated) or can be modeled stochastically. The most popular statistical methods are n-gram models, which attempt to capture the syntactic and semantic constraints by estimating the frequencies of sequences of n words. The assumption is made that the probability of a given word string $(w_1, w_2, ..., w_k)$ can be approximated by $\prod_{i=1}^{k} Pr(w_i|w_{i-n+1}, ..., w_{i-2}, w_{i-1})$, thereby reducing the word history to the preceding $n-1$ words. A back-off mechanism is generally used to smooth the estimates of the probabilities of rare n-grams by relying on a lower order n-gram when there are insufficient training data, and to provide a means of modeling unobserved word sentences. While trigram Language Models are the most widely used, higher-order $(n > 3)$ and word-cross-based (counts are based on sets of words rather than individual lexical items) n-grams, and adapted Language Models, are recent research areas aimed at improving Language Model accuracy.

Given a large text corpus, it may seem relatively straightforward to construct n-gram language models. Most of the steps are pretty standard and make use of tools that count word and word sequence occurrences. The main differences arise in the choice of the vocabulary and in the definition of words, such as the treatment of compound words or acronyms, and the choice of the back-off strategy. There is, however, a significant amount of effort needed to process the texts before they can be used.

A common motivation for text normalization in all languages is to reduce lexical variability so as to increase the coverage for a fixed-size task vocabulary. Normalization decisions are generally language specific. Much of speech recognition research for American English has been supported by ARPA and has been based on text materials which were processed to remove upper/lower-case distinction and compounds. Thus, for instance, no lexical distinction is made between Gates, gates or Green, green. In the French Le Monde corpus, capitalization of proper names is distinctive, with different lexical entries for Pierre, pierre or Roman, roman.

The main conditioning steps are text mark-up and conversion. Text mark-up consists of tagging the texts (article, paragraph, and sentence markers) and garbage bracketing (which includes not only corrupted text materials, but all text material unsuitable for sentence based language modeling, such as tables and lists). Numerical expressions are typically expanded to approximate the spoken form ($150-> one hundred and fifty dollars). Further semi-automatic processing is necessary to correct frequent errors inherent in the texts (such as obvious misspellings million, officials) or arising from processing with the distributed text processing tools. Some normalization can be considered as "decompounding" rules in that they modify the word boundaries and the total number of words. These concern the processing of ambiguous punctuation markers (such as hyphen and apostrophe), the processing of digit strings, and treatment of abbreviations and acronyms (ABCD -> A.B.C.D.). Another example is the treatment of numbers in German, where decompounding can be used in order to increase lexical coverage. The data 1991, which in standard German is written as neunzehnhunderteinundneunzig, can be represented by word sequence neunzehn hundert ein und neunzig. Other normalizations (such as sentence-initial capitalization and case distinction) keep the total number of words unchanged, but reduce graphemic variability. In general the choice is a compromise between producing an output close to the correct standard written form of the language and the lexical coverage, with the final choice of normalization being largely application driven.

Better language models can be obtained using texts transformed to be closer to observed reading styles, where the transformation rules and corresponding probabilities can be automatically derived by aligning prompt texts with the transcriptions of the acoustic data.

In practice, the selection of words is done so as to minimize the systems OOV rate by including the most useful words. By useful we mean that the words are expected as an input to the recognizer, but also that the Language Model can be trained given the available text corpora. In order to meet the latter condition, it is common to choose the n most frequent words in the training data. This criterion does not, however, guarantee the usefulness of the lexicon, since no consideration of the expected input is made. Therefore it is common practice to use a set of additional development data to select a word list adapted to the expected test conditions.

There are sometimes conflicting needs for sufficient amounts of text data to estimate Language Model parameters, and for ensuring that the data are representative of the task. It is also common that different types of Language Model training material are available in differing quantities. One easy way to combine training material from different sources is to train a language model per source and to interpolate them. The interpolation weights can be directly estimated on some development data with the EM algorithm. An alternative is to simply merge the n-gram counts and train a single language model on these counts. If some data sources are more representative than others for the task, the n-gram counts can be empirically weighting to minimize the perplexity on a set of development data. While this can be effective, it has to be done by trial and error and cannot easily be optimized. In addition, weighting the n-gram counts can pose problems in properly estimating the back-off coefficients. The relevance of a language model is usually measured in terms of test-set perplexity defined as $Px = Pr(text|LanguageModel)^{-\frac{1}{n}}$, where n is the number of words in the text. The perplexity is a measure of the average branching factor, i. e. the vocabulary size of a memory less uniform language model with the same entropy as the language model under consideration.

The Carnegie Mellon University Statistical Language Modeling (CMU-SLM) Toolkit [30] is a set of Unix software tools, designed to facilitate language modeling work in the research community. Its version 1.0 [2] was written by Roni Rosenfeld and released in 1994.

In the course of time, the available corpora are every time larger and the available computers for processing are more powerful. Interest has grown in moving beyond trigram language model towards tetragram and pentagram models. Besides, some of version 1's inefficiencies, which are tolerable when we have small corpora, became a real problem when we have hundreds of millions of words.

Version 2[3] of the toolkit has been developed in order to solve these problems. It tries to conserve the structure of Version 1, to include all of the functionality of Version 1 and to improve Version 1 in terms of functionality and efficiency.

In this work, we have used Version 2 of Toolkit to apply the concept of language model.

4.4.3 N-fold Cross Validation

Test and training sets are, ideally, independent on each trial. However, this would require too much labeled data. To solve this problem, we run n-trials, each time using a different segment of the data for testing and training on the remaining n-1 segments. This way, at least test-sets are independent.

We then average classification accuracy over the n trials. Usually, n is equal to 10.

4.5 Results

4.5.1 Evaluation Scheme

Knowing each word weight in the several corpora, it is also necessary to know their morphological class to be able to make an evaluation at several morphological levels. For that purpose, we used the part-of-speech tagger TreeTagger[4] [31] [32]. In particular, we want to understand to which extent

---

[2] http://www.speech.cs.cmu.edu/SLM/CMU-Cam_Toolkit_v1.tar.gz

[3] http://www.speech.cs.cmu.edu/SLM/CMU-Cam_Toolkit_v2.tar.gz

[4] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger

each part-of-speech plays a role in the similarity between corpora.

In our evaluation, we used the 10-fold cross validation technique. To apply it, we built ten different training sets and an equal number of test sets containing randomly selected sentences from the corpora both for the case of Subjectivity and Objectivity analysis. As a consequence, we built ten training sets containing 90% of Wikipedia corpus sentences, ten other training sets containing 90%Weblogs corpus sentences and ten other training sets containing 90% Reuters sentences. Similarly, we built ten test sets each one containing 10% of objective sentences of the Subjectivity v1.0 corpus and the other containing 10% of subjective sentences of the Subjectivity v1.0 corpus. Indeed, we are interested in knowing to what extent the manually classified (objective/subjective) sentences approximate the Wikipedia, Weblogs and Reuters sentences.

Created the sets for each model, we applied the Rocchio Method and the Language Model for our training and test sets for a 10-fold cross-validation scheme.

4.5.2 Rocchio Method

When we applied the Rocchio Method, we used the Cosine similarity measure to calculate the cosine of an angle between two vectors. These vectors are constituted by training sets and test sets sentences features respectively.

In the following tables, we present the results obtained at several morphological levels. In Table 4.1[5], we present the results in which the training vector is the set of Wikipedia sentences and the testing vectors are the subjective and objective sentences from the Subjectivity v1.0 corpus. The Cosine similarity is performed ten times, one for each test set and averaged.

Morphological Level	**Subjective**	**Objective**	**Class**
All Words	0.76	0.79	Objective
All ADJ	0.54	0.61	Objective
All V	0.71	0.67	Subjective
All N	0.66	0.69	Objective
All ADJ + All V	0.65	0.66	Objective
All ADJ + All N	0.65	0.68	Objective
All N + All V	0.70	0.69	Subjective
All ADJ + All N + All V	0.68	0.69	Objective

Table 4.1: *Results with the Wikipedia Model.*

The graph shown below illustrates what happens when the class is Objective.

---

[5] For example: if the line the of table is **All Words** and the column the of table is **Objective**, then $CosSim(Wikipedia\_Objective) = 0.79$

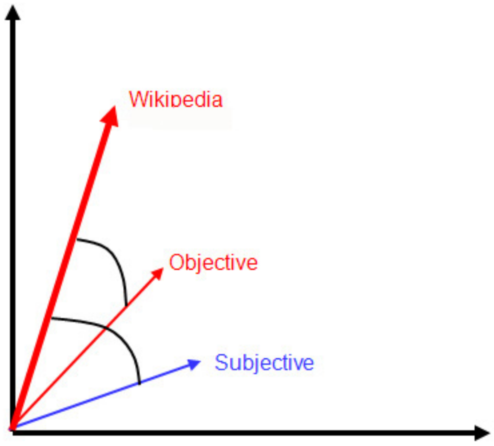


Figure 4.2: *Graph that illustrates the Table 4.1.*

On the other hand the next graph presented below illustrates what happens when the class is Subjective.

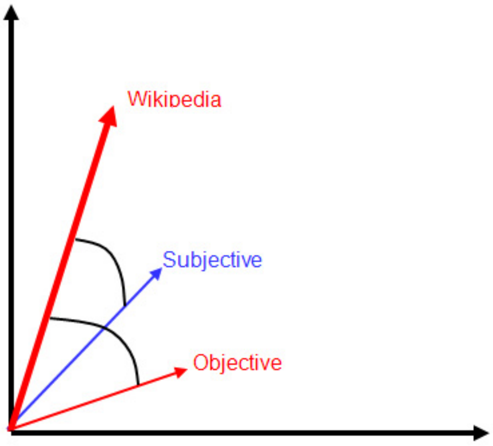


Figure 4.3: *Graph that illustrates the Table 4.1.*

In Table 4.2[6], we present the results in which the training vector is the set of Weblogs sentences and the testing vectors are the subjective and objective sentences from the Subjectivity v1.0 corpus. The Cosine similarity is performed ten times, one for each test set and averaged.

---

[6] For example: if the line the of table is **ADJ** and the column the of table is **Subjective**, then $CosSim(Weblogs\_Subjective) = 0.52$

Morphological Level	Subjective	Objective	Class
All Words	0.60	0.56	Subjective
All ADJ	0.52	0.49	Subjective
All V	0.53	0.48	Subjective
All N	0.47	0.43	Subjective
All ADJ + All V	0.49	0.48	Subjective
All ADJ + All N	0.48	0.44	Subjective
All N + All V	0.50	0.45	Subjective
All ADJ + All N + All V	0.47	0.46	Subjective

Table 4.2: *Results with the Weblogs Model.*

The graph shown below illustrates what happens when the class is Subjective.

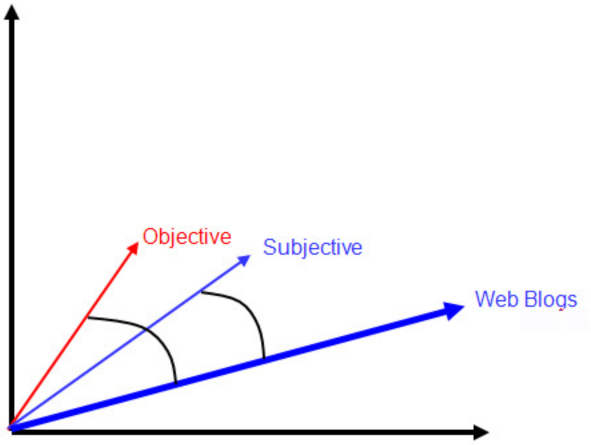


Figure 4.4: *Graph that illustrates the Table 4.2.*

In Table 4.3, we present the results in which the training vector is the set of Reuters sentences and the testing vectors are the subjective and objective sentences from the Subjectivity v1.0 corpus. The Cosine similarity is performed ten times, one for each test set and averaged.

Morphological Level	**Subjective**	**Objective**	**Class**
All Words	0.64	0.68	Objective
All ADJ	0.30	0.40	Objective
All V	0.38	0.37	Subjective
All N	0.34	0.47	Objective
All ADJ + All V	0.36	0.38	Objective
All ADJ + All N	0.35	0.49	Objective
All N + All V	0.36	0.47	Objective
All ADJ + All N + All V	0.37	0.47	Objective

Table 4.3: *Results with the Reuters Model.*

The values presented in the Table 4.1, 4.2 and 4.3 are the values of the cosine between a training vector (Wikipedia or Weblogs or Reuters) and a testing vector (Objective or Subjective). We know that the greater the value of the Cosine is, the less the angle between the two carriers is, and as a consequence the closest the vectors are.

As we can see above, the results are satisfactory, but we will discuss them in more detail in the last section of this chapter.

4.5.3 Language Model

As we already referred, we used the tool CMU-ToolKit to apply the Language Model. In this context, the minor the perplexity (Px) and the entropy (H) are between the training and the testing models, the more the test model fits to the training model. We present the results in Table 4.4.

		Model		
		Wikipedia	Weblogs	Reuters

Text	Objective	Px = 691.27
		H = 9.43
			Px = 2027.06
			H = 10.99
				Px = 1104.03
				H = 10.11

	Subjective	Px = 880.67
		H = 9.75
			Px = 1991.09
			H = 10.96
				Px = 1226.34
				H = 10.26

Table 4.4: *Results obtained by the Language Model.*

In the Language Model, our training models contain sentences extracted from Wikipedia, Weblogs and Reuters respectively, giving rise to three language models. The testing models contain sentences extracted from the Subjectivity v1.0 corpus, i.e. objective and subjective sentences.

What we can constat in Table 4.4 is that the training model Wikipedia shows lower perplexity and entropy for the objective sentences than the subjective sentences. Likely, the same happens when using the training model Weblogs. In that case, lower perplexity and entropy are shown for the subjective sentences than the objective sentences. These results meet our expectations.

4.6 Conclusion and Discussion

As it has already been referred, the obtained results are satisfactory. In fact, there is a strong similarity between the Wikipedia corpus and objective sentences and between the Weblogs corpus and subjective sentences.

The Reuters Corpus was introduced in the similarity evaluation between the corpora to see if it was objective or subjective. Though it was not in our plans to use this corpus for the evaluation of sentences, we thought it would be interesting to see if we could classify a journalistic corpus of reference. Our initial suspicion was that Reuters corpus could be considered a mixture of objective and subjective sentences. From the beginning, the classification of the Reuters corpus would be objective, once it is a journalistic corpus, nonetheless we also have evidence of events or situations that are reported by the interviewed people. These people give their opinion and they exercise their own judgment on something, so these sentences are considered subjective. We have shown this in our evaluation.

When we look at the results of the Rocchio method, we may conclude that the Reuters corpus is objective in his global context. But, it can be considered subjective when we only look at the analysis of the verbs - while in all the other morphological levels it is classified as objective. These data are easily verified in column "Class" of the table 4.3.

In a more detailed analysis, it is necessary to take the three tables of the evaluation into account. This analysis shows that the Wikipedia corpus is objectively stronger than the Reuters corpus. On the other hand, the Weblogs corpus is definitively subjective. In particular, we have shown that the use of verbs (alone or accompanied by another morphological class) is not of great interest in the classification of objective texts.

The use of the Reuters corpus gives more consistency to our methodology and our contribution to this scientific area, as - according to our evaluation - this corpus is objective in its global context; although not as objective as the Wikipedia corpus. This way, we got to mark the separation that exists between the Wikipedia corpus and the Weblogs corpus and, at the same time, we demonstrated that the Wikipedia corpus is more similar to the objective text that belongs to Subjectivity v1.0 Corpus than to the subjective text of the same Corpus, and as for the Weblogs corpus it is the opposite.

With this superficial conclusion, we may think that Wikipedia and Weblogs may replace the habitual manually annotated corpora.

However, we need to make a more rigorous evaluation of the results. There is to say that we are satisfied with the results obtained through the

Language Model, as we can draw the conclusion - with some evidence - that the similarity between the Wikipedia corpus and the objectivity sentences is larger than its similarity with the subjective ones. In a symmetrical way, we concluded that the similarity between the Weblogs corpus and subjective sentences is larger than the similarity between the Weblogs corpus and objective sentences. The similarity evaluation with the Language Model is possible through perplexity and entropy. We can interpret (in a rough way) that the perplexity and the entropy show the degree of surprise with which the training model "sees" the test model. So, the minor the degree of surprise is, the more similar to the training model the test model is.

With the Rocchio Method, we used several morphological levels (this cannot happen in the Language Model). That means that the similarity evaluation was done based on vectors which contained **all words** or **all nouns (All N)** or **all adjectives (All ADJ)** or **all verbs (All V)** or **all adjectives + all verbs (All ADJ + All V)** or **all adjectives + all nouns (All ADJ + All N)** or **all nouns + all verbs (All N + All V)** or, finally, **all adjectives + all of the nouns + all verbs (All ADJ + All N + All V)**.

When we use as training vector the one constituted by Weblogs features, the similarity evaluation is good. Indeed, in all cases and as referred in Table 4.2, the subjective sentences always approximate the training vector more than the objective ones. We get a good classification in all of the morphological levels. Unfortunately, it is not the same with the similarity evaluation using as training the vector constituted by the Wikipedia features. When we built the training and test vectors only with all verbs, or with all nouns and with all verbs, the results were not satisfactory as shown in Table 4.1. Thus, we could conclude that the inclusion of verbs does not benefit classification and may be avoided when trying to automatically learn classifiers.

There are other situations that are not as satisfactory as we would like them to be. In spite of good classifications, the values of the Cosine in the same morphological level between the training vector and the test vectors are usually very close. This does not give much confidence in the results. To avoid this problem we developed a metric to find out which of the morphological levels was the most selective. The metric is defined in Equation 4.3. This measure is easy to understand. For each morphological level, we look at its capacity to correctly classify subjective sentences when trained with Weblogs combined with its capacity to correctly classify objective sentences when trained with Wikipedia. It is also a way of combining the approximation to both training vectors instead of just one.

$$quality = \frac{CosSim(Weblogs\_Subjective)}{CosSim(Weblogs\_Objective)} + \frac{CosSim(Wikipedia\_Objective)}{CosSim(Wikipedia\_Subjective)} \tag{4.3}$$

Through the equation described above we obtained the following results for each morphological level in Table 4.5.

Morphological Level	$\frac{CosSim(Weblogs\_Subjective)}{CosSim(Weblogs\_Objective)}$	$\frac{CosSim(Wikipedia\_Objective)}{CosSim(Wikipedia\_Subjective)}$	**quality**
All Words	$\frac{0.60}{0.56}$	$\frac{0.79}{0.76}$	2.25
All ADJ	$\frac{0.52}{0.49}$	$\frac{0.61}{0.54}$	2.19
All N	$\frac{0.47}{0.43}$	$\frac{0.69}{0.66}$	2.14
All ADJ + All V	$\frac{0.49}{0.48}$	$\frac{0.66}{0.65}$	2.04
All ADJ + All N	$\frac{0.48}{0.44}$	$\frac{0.68}{0.65}$	2.14
All ADJ + All N + All V	$\frac{0.47}{0.46}$	$\frac{0.69}{0.68}$	2.04

Table 4.5: *Results of the metric measure.*

With these values, we calculated the average and considered relevant morphological levels, which results were above the average. The average is 2.13 and as a consequence, relevant morphological levels are: **all words**, **all nouns**, **all adjectives** and **all adjectives + all nouns**.

In this chapter, we showed that there is a great similarity between Wikipedia sentences and objective sentences as well as between Weblogs sentences and subjective sentences. For this purpose, we used two different methodologies: the Rocchio Method and the Language Model. Therefore, we can affirm - with some degree of certainty - that the community, which investigates in the area where this master thesis intervenes, may replace habitual manually annotated corpora.

Besides this positive similarity evaluation, we also got to know which the most relevant morphological levels for the classification of sentences were. However, we still need to prove to which an extent these results can embrace real world classification of sentences. This is our aim in the next chapter.

Chapter 5

Classification of Sentences

As already advanced in the previous chapter, the classification of sentences becomes more complicated as the number of features in a sentence is quite reduced. In this chapter we will show some fairly intriguing results even if they may be considered little satisfactory. Nonetheless, we believe that they are very important to continue our investigation in this area.

5.1 Support Vector Machine

In the last decades, improving classifier performance has been an area of exhaustive machine-learning research. On the other hand, we attended to a new generation development of state-of-the-art classifiers, such as Support Vector Machines [5] [33] [34] [35], boosted decision trees, regularized logistic regression, neural net-works and random forests.

Support Vector Machines are not necessarily better than other methods, but they have much current theoretical and empirical appeal.

Support Vector Machines are based on Statistical Learning Theory proposed by Vladimir Vapnik and Alexey Chernovemkis, between 1960 and 1970. This theory seeks to find mathematical conditions for a function choice that separates data apprehended in categorization problems. This separation should consider the smallest training error and, at the same time, it should maximize the capacity of classifier generalization.

There are several favorable conditions to use SVM's. They have high generalization capacity, avoiding over training. They are robust for the data categorization with high dimensions, which tend to be over trained in other classifiers, because a lot of micro-features can discriminate very little. The function objective is convex, because it is a quadratic function with just a great global. Finally, we can increase that SVM's underlying theory is very established in mathematics and statistics areas.

While some learning methods, such as the perceptron algorithm, just find any linear separator, others search for the best linear separator in agreement with certain criterion, and the SVM in particular defines this as looking for a decision surface that is maximally far away from any data points. This distance from the decision surface to the closest data point determines the margin of the classifier. A classifier with a large margin makes no very uncertain classification decisions.

A SVM is built to maximize the margin around the separating hyperplane. This inevitably means that the decision function for a SVM is completely specified by a subset of the data which defines the position of the separator. We call support vectors these points.

Data sets that are linearly separable are well-handled, but if data set is too hard it just doesn't allow classification by a linear classifier. One way to solve this problem is to map the data on to a higher dimensional space and then to use there a linear classifier. Kernels can make a non-separable problem separable, and they can map data into a representational space. In this work, we used some Kernel functions: Linear function, Polynomial function and Radial Basis function.

5.2 Results

5.2.1 Evaluation Scheme

The evaluation scheme that will be described in this chapter is, in some points, similar to the evaluation scheme described in chapter 4.2.1. Thus, we will take advantage of some work already developed there.

In the process of classification of sentences we used the same measure that was used in the evaluation of similarity of the corpora to attribute a weight to a word. We also used the concept of cross-validation with the same models used previously, but each model was increased with two more training sets: one contains 90% of the sentences contained in Objectivity Corpus and the other contains 90% of the sentences contained in Subjectivity Corpus. Therefore we have ten models having each one a Wikipedia training set, a Weblogs training set, a Objectivity training set, a Subjectivity training set, a Objectivity test set and a Subjectivity test set. As it was already referred, these groups hold a certain number of sentences taken randomly from the respective corpora.

We added two more training sets - Objectivity and Subjectivity - to our models so to compare the results obtained by classifying the sentences of our test sets with the training model created from the Wikipedia and Weblogs training sets and with the training model created from Objectivity and Subjectivity training sets.

One of the differences we have in the classification of sentences in relation to the evaluation of the similarity corpora is that now we know exactly which morphologic classes are more relevant to us in order to trying a good classification of sentences. Thus, in this chapter we will only use the morphologic classes we considered relevant in the previous chapter.

Due to the reduced number of features the sentences have and to the vast number of features our corpora - not manually annotated - possesses, we thought we could be more meticulous and demanding with our methodology. As input, we first began to give our classifier all the sentences for it to classify; then we gave it only sentences that contained more than one feature; later on we built the input for the classifier only with sentences that possessed more than two features. We concluded that decreasing the number of sentences successively to input it is going to increase the number of features that the sentence needs to have to be part of the input file of the classifier. This way, we get to know, for example, which the number of features is that a sentence should have to be well classified and as the number of sentences well classified with the increase of features there in contained varies.

Until now we have described the first part of our evaluation scheme, which is, we demonstrated how we dealt with our training and test data and gave the approaches we used to make a good classification of the sentences.

In the second and last part of our evaluation system we will focus on the way how the results - obtained through the classification of the sentences in its turn achieved through the classifier SVM^{Ligth}[1] [35], will be represented. The results will be represented through the confusion matrix [2] [13].

In the field of artificial intelligence, a confusion matrix is a visualization tool typically used in supervised learning (in unsupervised learning it is typically called a matching matrix). Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. One benefit of a confusion matrix is that it is easy to see if the system is confusing two classes (i.e. commonly mislabeling one as another).

When a data set is unbalanced (when the number of samples in different classes vary greatly) the error rate of a classifier is not representative of the true performance of the classifier.

The entries in the confusion matrix have the following meaning in the context of our study:

- a is the number of correct **predictions** that an instance is **Wikipedia**,
- b is the number of incorrect **predictions** that an instance is **Weblog**,
- c is the number of incorrect **predictions** that an instance in **Wikipedia** and
- d is the number of correct **predictions** that an instance is **Weblog**.

---

[1] http://svmlight.joachims.org/

		Predicted	
		Wikipedia	**Weblogs**
Actual	**Objectivity**	**a**	**b**
	Subjectivity	**c**	**d**

Table 5.1: *Confusion Matrix.*

Or.

- a is the number of correct **predictions** that an instance is **Objectivity**,
- b is the number of incorrect **predictions** that an instance is **Subjectivity**,
- c is the number of incorrect **predictions** that an instance in **Objectivity** and
- d is the number of correct **predictions** that an instance is **Subjectivity**.

		Predicted	
		Objectivity	**Subjectivity**
Actual	**Objectivity**	**a**	**b**
	Subjectivity	**c**	**d**

Table 5.2: *Confusion Matrix.*

Several standard terms have been defined for the 2 class matrix:

The accuracy (AC) is the proportion of the total number of predictions that were correct. It is determined using the equation:

$$AC = \frac{a+d}{a+b+c+d} \tag{5.1}$$

The recall or true positive rate (TP) is the proportion of positive cases that were correctly identified, as calculated using the equation:

$$TP = \frac{d}{c+d} \tag{5.2}$$

Finally, precision (P) is the proportion of the predicted positive cases that were correct, as calculated using the equation:

$$P = \frac{d}{b+d} \tag{5.3}$$

The accuracy determined using equation 1 may not be an adequate performance measure when the number of negative cases is much greater than the number of positive cases. Suppose there are 1000 cases, 995 of which are negative cases and 5 of which are positive cases. If the system classifies them all as negative, the accuracy would be 99.5%, even though the classifier missed all positive cases. Other performance measures account for this by including TP in a product: for example F-Measure, as defined in equation 9.

$$F = \frac{(\beta^2 + 1) * P * TP}{\beta^2 * P + TP} \tag{5.4}$$

The representation of the results on the form of Confusion Matrix is very useful as we have just see, because starting from her we got to know, among other data, the accuracy, the recall and the precision of our results.

Now, we will analyze and discuss the achieved results, in the several perspectives: of the morphologic class that best result obtains; of the minimum number of features with which is gotten better results. We will still analyze the Confusion Matrix and all that it derives (accuracy, recall, precision, etc.), not forgetting obviously to refer the kernel functions that were used.

All of the referred analyses approach two similar situations: when the test sentences are submitted to a classification using as training set the Wikipedia/Weblogs and also using as training set Objectivity / Subjectivity. Thus, we can compare these two classifications and to see which the differences between both.

We presented the Confusion Matrix, when we used all of the words as features for the classification process, in which a certain kernel function is used. Then, we presented the resulting graphs of the classification of that function with the parameters that can be calculated from the Confusion Matrix.

5.2.2 Support Vector Machine

We will present the results of the entire process of classification of sentences. These results will be presented in the confusion matrix and graphics.

We believe that it will only be necessary to represent the confusion matrix when all sentences are classified and not represent the confusion matrix when classifying sentences that have more than 1 feature, or more than 2 features,...

In contrast to the confusion matrices, graphics are represented in the process of complete classification of sentences, when all the sentences are classified, if there are only classified sentences that have more than 1 feature, or more than 2 features,...

Linear Function

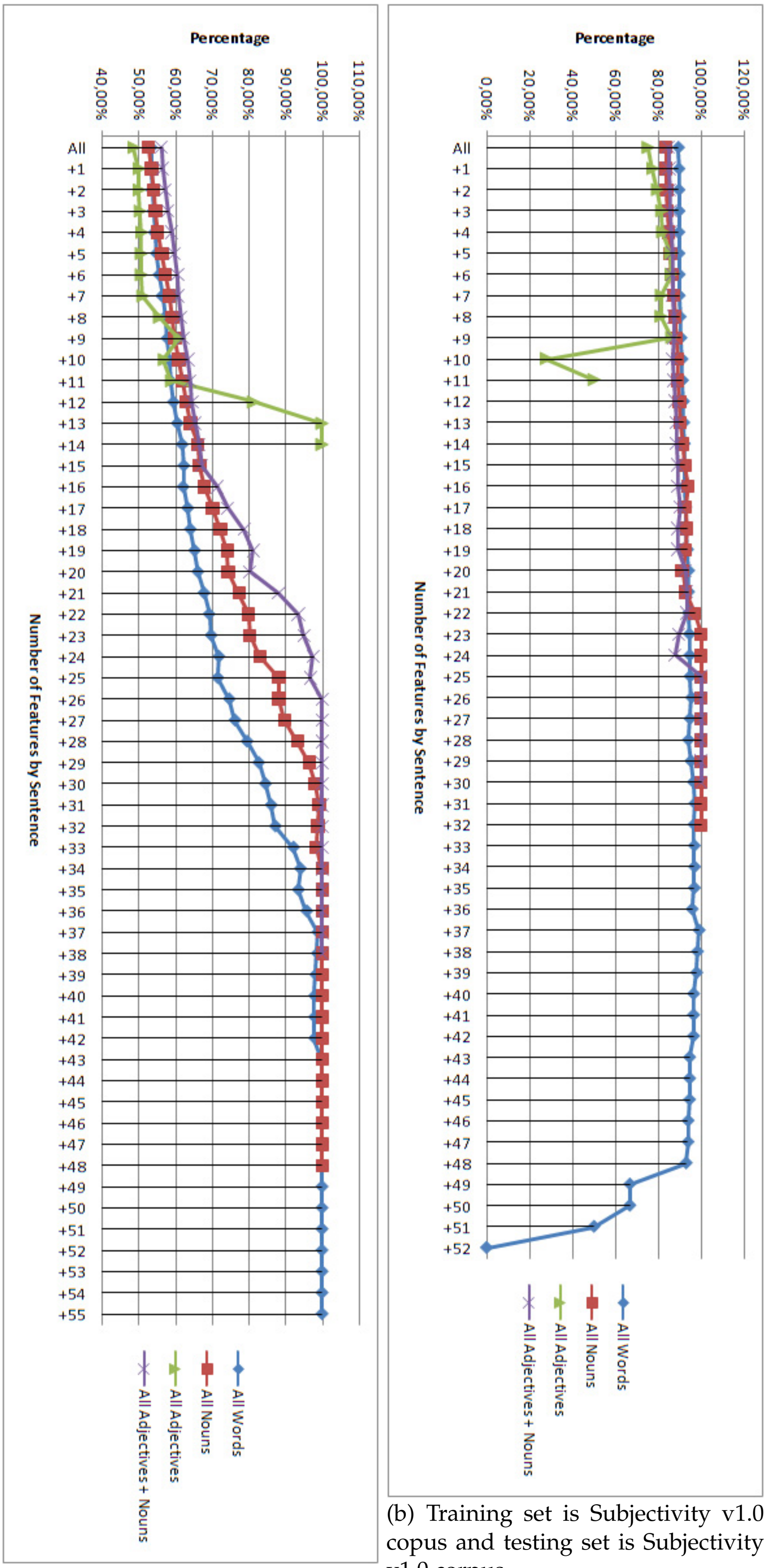
Next there are the results presented in the form of confusion matrix and graphs that allow us to draw conclusions from the entire process of classification of sentences to use a linear function in SVM^{Ligth}.

		Predicted	
		Wikipedia	**Weblogs**
Actual	**Objectivity**	455.23	13.01
	Subjectivity	417.90	4.40

Table 5.3: *Confusion Matrix for all Words and for Linear Function, where the predicted class is Wikipedia and Weblogs and actual class is Objectivity and Subjectivity.*

		Predicted	
		Objectivity	**Subjectivity**
Actual	**Objectivity**	407.43	59.24
	Subjectivity	65.45	360.12

Table 5.4: *Confusion Matrix for all Words and for Linear Function, where the predicted class is Objectivity and Subjectivity and actual class is Objectivity and Subjectivity.*



(a) Training set is Wikipedia copus and Weblogs corpus. Testing set is Subjectivity v1.0 corpus.

(b) Training set is Subjectivity v1.0 copus and testing set is Subjectivity v1.0 corpus.

Figure 5.1: *Graphs of the Linear function and Accuracy.*

In the analysis of the Figure 5.1, we will consider the accuracy definition, represented in this chapter by the Equation 5.1.

We observe the graph a) of Figure 5.1 and we conclude that each representative line of a morphologic level has an identical behavior, in other words, the more features are considered for the classification of the sentences (objectivity vs. subjectivity) the bigger is the success percentage - more proportion of the total number of predictions that were correct.

Guidelines of the graph:

- in the classification of sentences with all the features, the success percentage for all of the morphologic levels is between 45% and 60%;
- in the classification of sentences with more than 20 features, the success percentage is between 60% and 80% for all of the morphologic levels, except for the adjectives, because there are no sentences with more than 13 adjectives;
- in the classification of sentences with more than 35 features the success percentage varies between 90% and 100%.

In terms of accuracy, the morphologic level that achieves better success percentage is "All Adjectives + Nouns", because its representative line has a bigger inclination - the same can't be applied to "All Words" or "All Nouns" - and it includes a considerable number of features (more than 33). On the contrary, the level morphologic "All Adjectives" has a bigger inclination in comparison to all the other levels, but in compensation it includes a more reduced number of features per sentence, which we consider an aspect that stands against the use of only adjectives to make the classification of sentences.

We observe the graph b) of Figure 5.1 and we conclude that each representative line of a morphologic level has a different behavior.

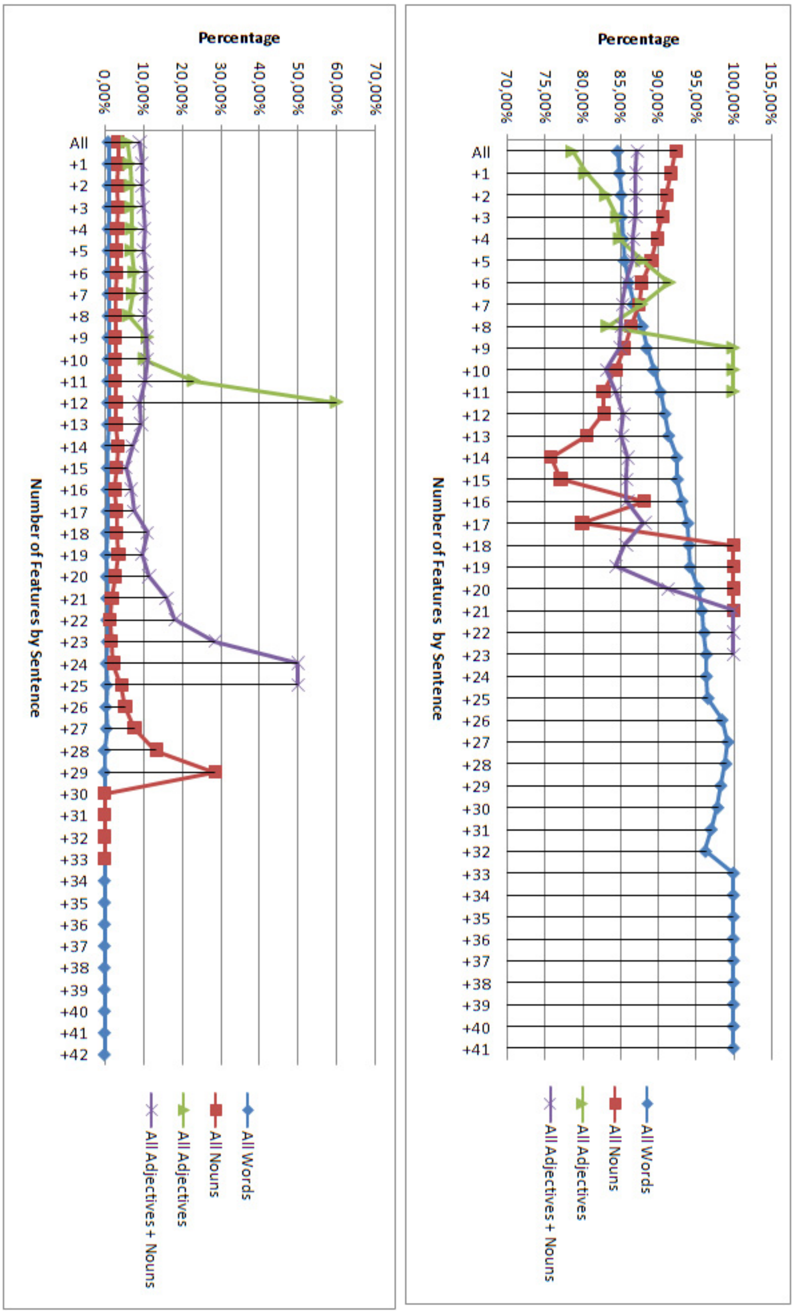
Guidelines of the graph:

- in the classification of sentences with a maximum number of 48 features, the success percentage for the morphologic level "All Words" is between 80% and 100%; though decreasing afterwards to 0%;
- the morphologic levels "All Nouns" and "All Adjectives + Nouns" have a similar behavior and their success percentages are between 80% and 90%;

- the morphologic level "All Adjectives" is, once again, characterized by its small number of features; its success percentage is between 70% and 90% even when the sentences have more than 9 features, then it suffers oscillations. When the sentences have more than 10 features, the success percentage is of approximately 25% and when the sentences have more than 11 features, the success percentage is of approximately 50%.

In terms of accuracy, the morphologic levels that achieve better success percentages are "All Nouns" and "All Adjectives + Nouns", as their representative lines achieve to have a success percentage of 80% to 100%. Besides, they include a considerable number of features (more than 32 and more than 30 respectively), unlike "All Adjectives" which always has a smaller success percentage relatively to the other morphologic levels. "All Words" is also punished, because with a high number of features for sentence its success percentage decreases to reach 0%.

When comparing the two graphs of Figure 5.1, we can conclude that, in terms of accuracy, graph a) achieves better classification results with a great number of features, unlike graph b) that declines to 0% starting from the 48 features. On the other hand, graph a) presents almost always the same standard for all of the morphologic levels, while graph b) presents some variations. Finally, the morphologic level, common to both graphs, which achieves a better success percentage, is "All Adjectives + Nouns".



(a) Training set is Wikipedia copus and Weblogs corpus. Testing set is Subjectivity v1.0 corpus.

(b) Training set is Subjectivity v1.0 copus and testing set is Subjectivity v1.0 corpus.

Figure 5.2: *Graphs of the Linear function and True Positive or Recall*

In the analysis of Figure 5.2, we will consider the True Positive or Recall definition, represented in this chapter by the Equation 5.2.

We observe graph a) of the Figure 5.2 and we conclude that each representative line of a morphologic level has a different behavior, although the success percentage - more proportion of the total number of predictions that were correct - is between 0% and 10%, independently of the number of features that are necessary to the sentences classification (objectivity vs. subjectivity).

Guidelines of the graph:

- in the morphologic level "All Words" the success percentage is very close to 0%;
- in the morphologic level "All Nouns" the success percentage is also very close to 0%, except when the sentences have more than 25, more than 26, more than 27, more than 28 and more than 29 features;
- the morphologic levels "All Adjectives" and "All Adjectives + Nouns" have a similar behavior, because the respective lines are going up until they reach a pick, from which there are not sentences with more than 12 Adjectives or sentences with more than 25 Adjectives + Nouns for the respective morphologic levels.

In terms of Recall, the morphologic level that has better success percentage is "All Adjectives + Nouns", because its corresponding line has a greater inclination - the same does not apply to "All Words" or "All Nouns" - and it includes a considerable number of features (more than 25). On the contrary, the morphologic level "All Adjectives" has a bigger inclination in comparison with all the other levels, but in compensation, once again, it includes a much reduced number of features for sentence, which we have already considered as being an aspect against the use of the adjectives only to make the classification of sentences.

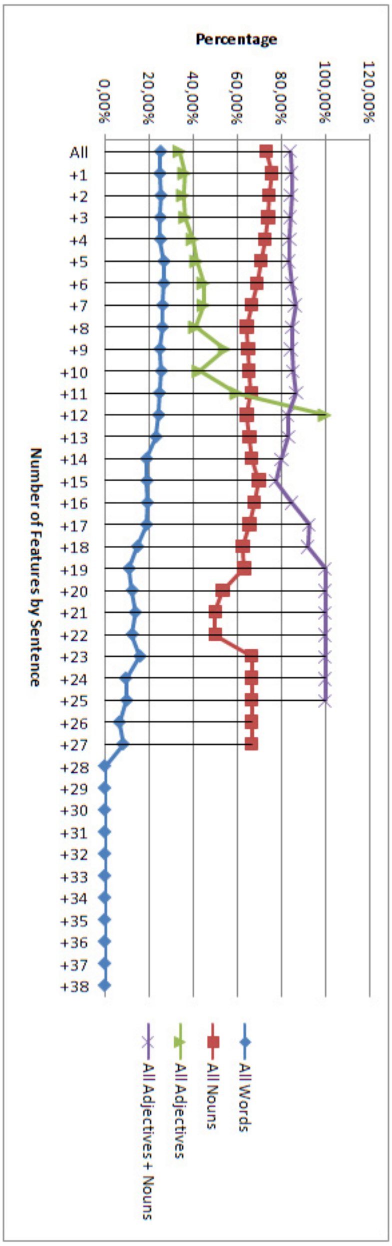
We observe graph b) of Figure 5.2 and we conclude that each representative line of a morphologic level has a different behavior and there are lots of oscillations.

Guidelines of the graph:

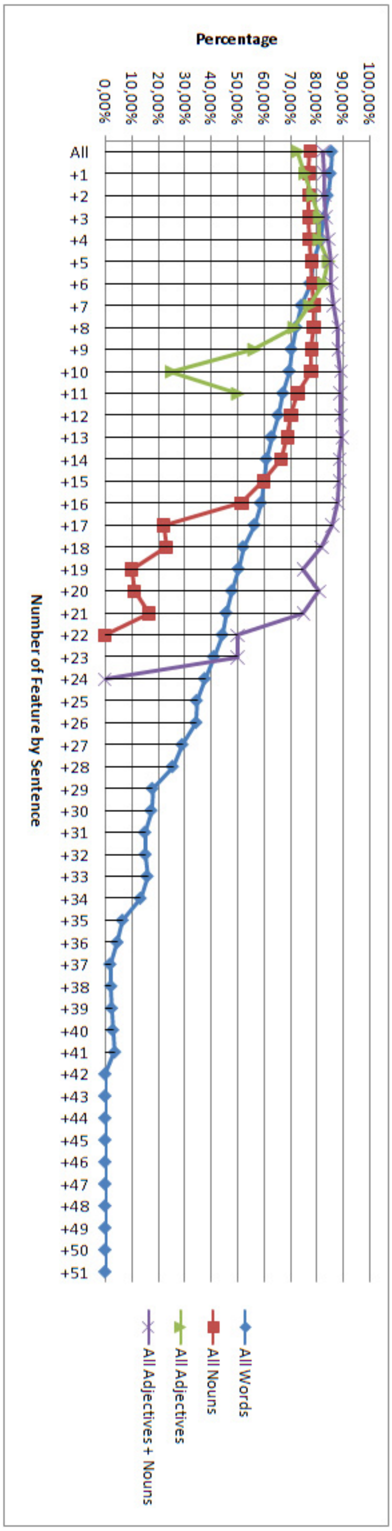
- there are a lots of oscillations;
- the only morphologic level that has less oscillations is "All Words"; in the classification of all sentences had a percentage of success of approximately 85%. This percentage increases (it suffers few oscillations) until reaching the 100%.

In terms of Recall, the morphologic levels have many oscillations, except for "All Words" that has a better success percentage and doesn't have so many oscillations.

When comparing the two graphs of Figure 5.2 we may conclude that, in terms of Recall, graph b) shows better success percentages than graph a). On the other hand, graph a) presents almost always the same pattern of "All Adjectives" and "All Adjectives + Nouns". Finally, the morphologic level "All Words" in graph b) has few oscillations and its success percentage increases gradually to reach the 100%.



(a) Training set is Wikipedia copus and Weblogs corpus. Testing set is Subjectivity v1.0 corpus.



(b) Training set is Subjectivity v1.0 copus and testing set is Subjectivity v1.0 corpus.

Figure 5.3: *Graphs of the Linear function and Precision*

In the analysis of Figure 5.3, we will consider the precision definition, represented in this chapter by the Equation 5.3.

We observe graph a) of Figure 5.3 and we conclude that each representative line of a morphologic level has an identical behavior. Although the success percentages - more proportion of the predicated positive cases that were correct - are different, some levels achieve to maintain the same inclination.

Guidelines of the graph:

- in the morphologic level "All Words" the more features there are in the sentences that are classified, the smaller the success percentage is decreasing until reaching 0%;
- in the morphologic level "All Nouns" with the increasing number of features there are in the sentences that are classified, the smaller is the success percentage (although it also suffers some oscillations). That success percentage is between 40% and 80%;
- in the morphologic level "All Adjectives" there are no sentences with more than 12 adjectives (reduced number of features). In the classification of all the sentences, the success percentage is approximately 30%. But, for the moment it is necessary that the sentences have a certain number of adjectives for them to be classified, as its success percentage goes up reaching 100%;
- clearly, in terms of Precision, our best morphologic level is "All Adjectives + Nouns", because its success percentage is between 80% and 100%.

In terms of precision, the morphologic level that gets the best success percentage is "All Adjectives + Nouns", because its line always achieves to be between 80% and 100% - the same does not happen to the other morphologic levels - and it is becoming a habit as it includes a considerable number of features (more than 25). On the contrary, the morphologic level "All Adjectives" has a larger inclination when compared with all the other levels. In compensation it has a more reduced number of features per sentence, but we consider it to be an aspect that stands against the use of only adjectives when making the classification of sentences.

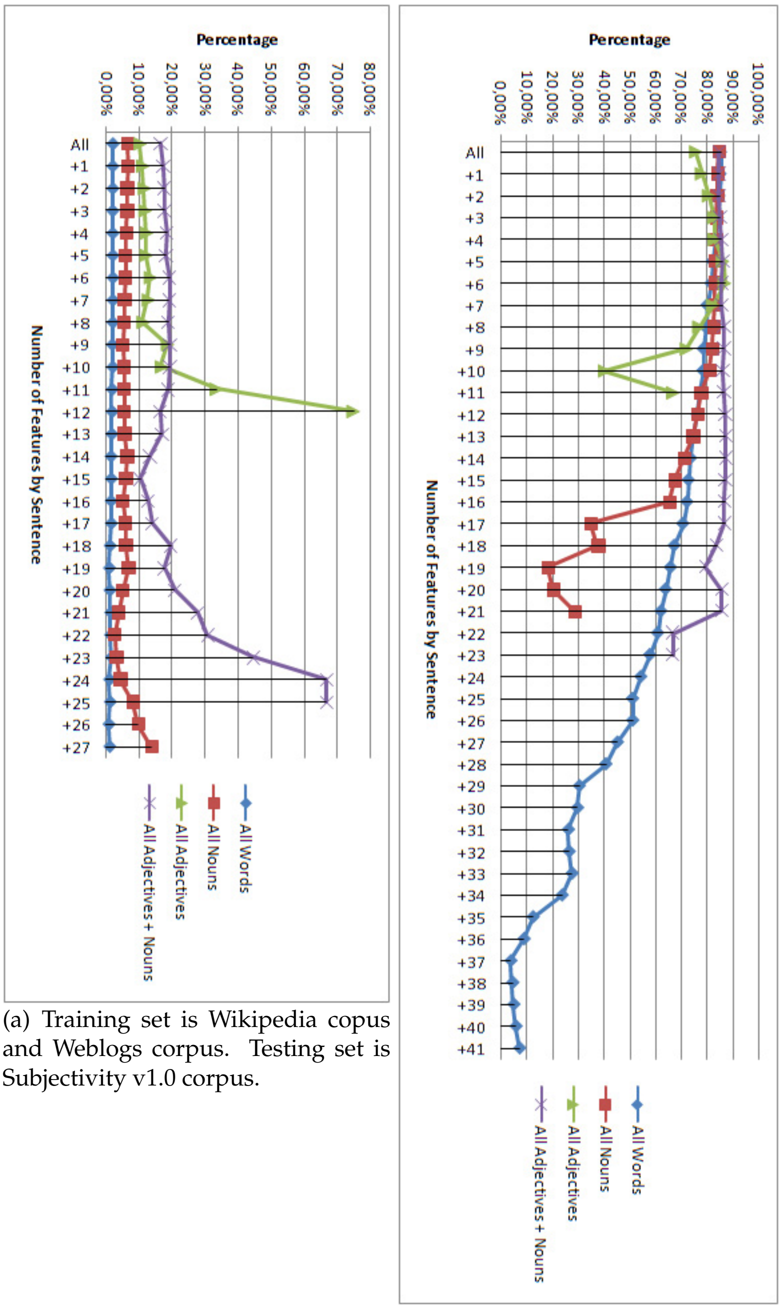
We observe graph b) of Figure 5.3 and we conclude that each representative line of a morphologic level has a identical behavior.

Guidelines of the graph:

- in the classification of all the sentences in all morphologic levels the success percentage is between 70% and 90%;
- in the classification of sentences, where the number of necessary features for its classification increases, the success percentage for "All Words", "All Nouns" and "All Adjectives + Nouns" decreases to 0%;
- in the morphologic level "All Adjectives" the success percentage is approximately 72% and it lowers when we only classify sentences that respect a certain number of features; but when we classify sentences with more than 11 features, the success percentage goes up again approaching 50%.

In terms of Precision, the morphologic level that gets better performance is "All Adjectives", as it never reaches 0%, in spite of having a reduced number of features.

By comparing the two graphs of Figure 5.3 we may conclude that, in terms of Precision, graph a) achieves better classification results - especially in the morphologic level "All Adjectives + Nouns" - in opposition to graph b) that reaches 0% in almost all morphologic levels. On the other hand, graph b) and graph a) almost always present the same pattern for all the morphologic levels, although both have oscillations. Finally, the morphologic level that achieves a better performance in graph b) is "All Adjectives" and in graph a) is "All Adjectives + Nouns."



(a) Training set is Wikipedia copus and Weblogs corpus. Testing set is Subjectivity v1.0 corpus.

(b) Training set is Subjectivity v1.0 copus and testing set is Subjectivity v1.0 corpus.

Figure 5.4: *Graphs of the Linear function and F-Measure*

In the analysis of Figure 5.4, we will consider the F-Mesure definition, represented in this chapter by the Equation 5.4.

We observe graph a) of Figure 5.4 and we conclude that each representative line of a morphologic level has a different behavior, although the success percentage in both morphologic levels and in most cases is of 0% to 20% - independently of the number of features that are considered to the classification of the sentences (Objectivity vs. Subjectivity).

Guidelines of the graph:

- in the morphologic level "All Words" the success percentage is very close to 0%;
- in the morphologic level "All Nouns" the success percentage is very close to 10%, with the exception of the sentences that have more than 27 features whose success percentage is close to 15%;
- the morphologic levels "All Adjectives" and "All Adjectives + Nouns" have a similar behavior, as the respective lines go up until they reach a pick in which the success percentage for "All Adjectives" is approximately 75% and 70% for "All Adjectives + Nouns". From these picks, there are no sentences with more than 12 adjectives or with more than 25 Adjectives + Nouns in the respective morphologic levels.

In terms of F-Measure, the morphologic level that achieves a better success percentage is "All Adjectives + Nouns", because its success percentage manages to be better than the one of the other morphologic levels and it includes a considerable number of features (more than 25). On the contrary, the morphologic level "All Adjectives" has a bigger inclination in comparison to all the other levels, but in compensation it includes a more reduced number of features per sentence, which we have considered as being an aspect that stands against the use of only adjectives when making the classification of sentences.

We observe graph b) of Figure 5.4 and we conclude that each representative line of a morphologic level has an identical behavior and there are lots of oscillations.

Guidelines of the graph:

- there are lots of oscillations and we verified that for all the morphologic levels - when we classify the sentences with all the characteristics - the success percentage is between 70% and 90%. These percentages decrease when we classify sentences that must have a certain number of features;

- the morphologic level "All Adjectives + Nouns" has a better performance than all the other morphologic levels.

In terms of F-Measure, the morphologic levels have lots of oscillations, except for "All Adjectives + Nouns" that achieves a better success percentage and doesn't have so many oscillations.

By comparing the two graphs of Figure 5.4 we may conclude that, in terms of F-Measure, graph b) shows better success percentages than graph a). On the other hand, graph a) almost always presents the same standard for "All Adjectives" and "All Adjectives + Nouns". Finally, the morphologic level "All Adjectives + Nouns" in graph b) has few oscillations and its success percentage doesn't fall below 60%, in none of the cases.

Polynomial Function

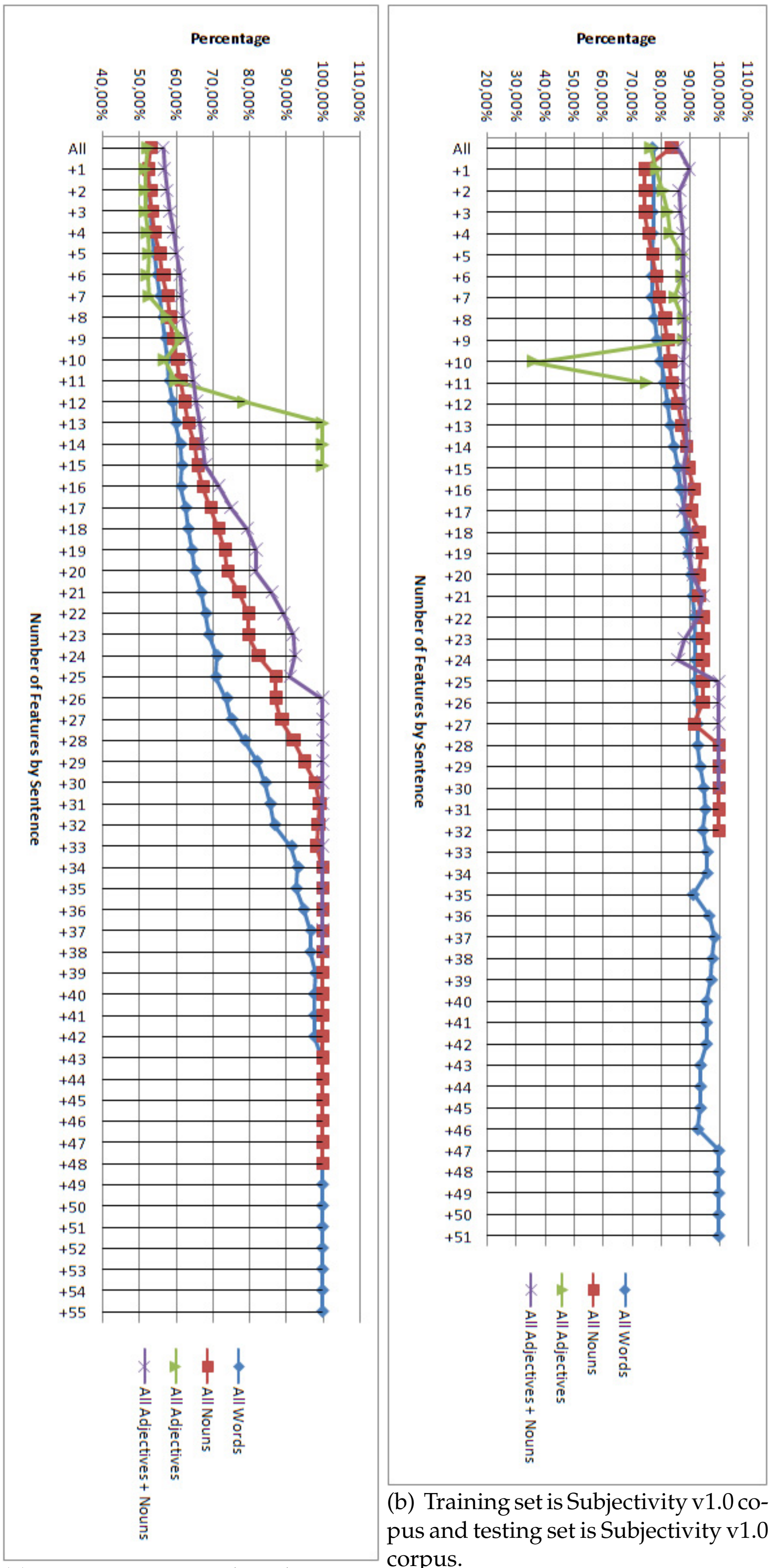
Following is the results presented in the form of confusion matrix and graphs that allow us to draw conclusions from the entire process of classification of sentences to use a Polynomial function in SVM^{Ligth}.

		Predicted	
		Wikipedia	**Weblogs**
Actual	**Objectivity**	466.32	1.56
	Subjectivity	425.84	0.24

Table 5.5: *Confusion Matrix for all Words and for Polynomial Function, where the predicted class is Wikipedia and Weblogs and actual class is Objectivity and Subjectivity.*

		Predicted	
		Objectivity	**Subjectivity**
Actual	**Objectivity**	433.35	35.36
	Subjectivity	176.83	249.30

Table 5.6: *Confusion Matrix for all Words and for Polynomial Function, where the predicted class is Objectivity and Subjectivity and actual class is Objectivity and Subjectivity.*



(a) Training set is Wikipedia copus and Weblogs corpus. Testing set is Subjectivity v1.0 corpus.

(b) Training set is Subjectivity v1.0 copus and testing set is Subjectivity v1.0 corpus.

Figure 5.5: *Graphs of the Polynomial function and Accurary.*

In the analysis of Figure 5.5, we will consider the accuracy definition, represented in this chapter by the Equation 5.1.

We observe graph a) of Figure 5.5 and we conclude that each representative line of a morphologic level has an identical behavior, in other words, the more features are considered for the classification of the sentences (objectivity vs. subjectivity), the bigger the success percentage is - more proportion of the total number of predictions that were correct.

Guidelines of the graph:

- in the classification of sentences with all the features the success percentage for all the morphologic levels is between 50% and 60%;
- in the classification of sentences with more than 20 features the success percentage is between 60% and 85% for all the morphologic levels, except for the adjectives, because there are no sentences with more than 13 adjectives;
- in the classification of sentences with more than 35 features the success percentage varies between 90% and 100%.

In terms of accuracy, the morphologic level that gets a better success percentage is "All Adjectives + Nouns", because its representative line achieves having a bigger inclination - the same does neither apply to "All Words" nor to "All Nouns" - and it includes a considerable number of features (more than 33). On the contrary, the morphologic level "All Adjectives" has a bigger inclination in comparison with all the other levels, but in compensation it includes a more reduced number of features per sentence, which we consider an aspect that stands against the use of only adjectives when making the classification of sentences.

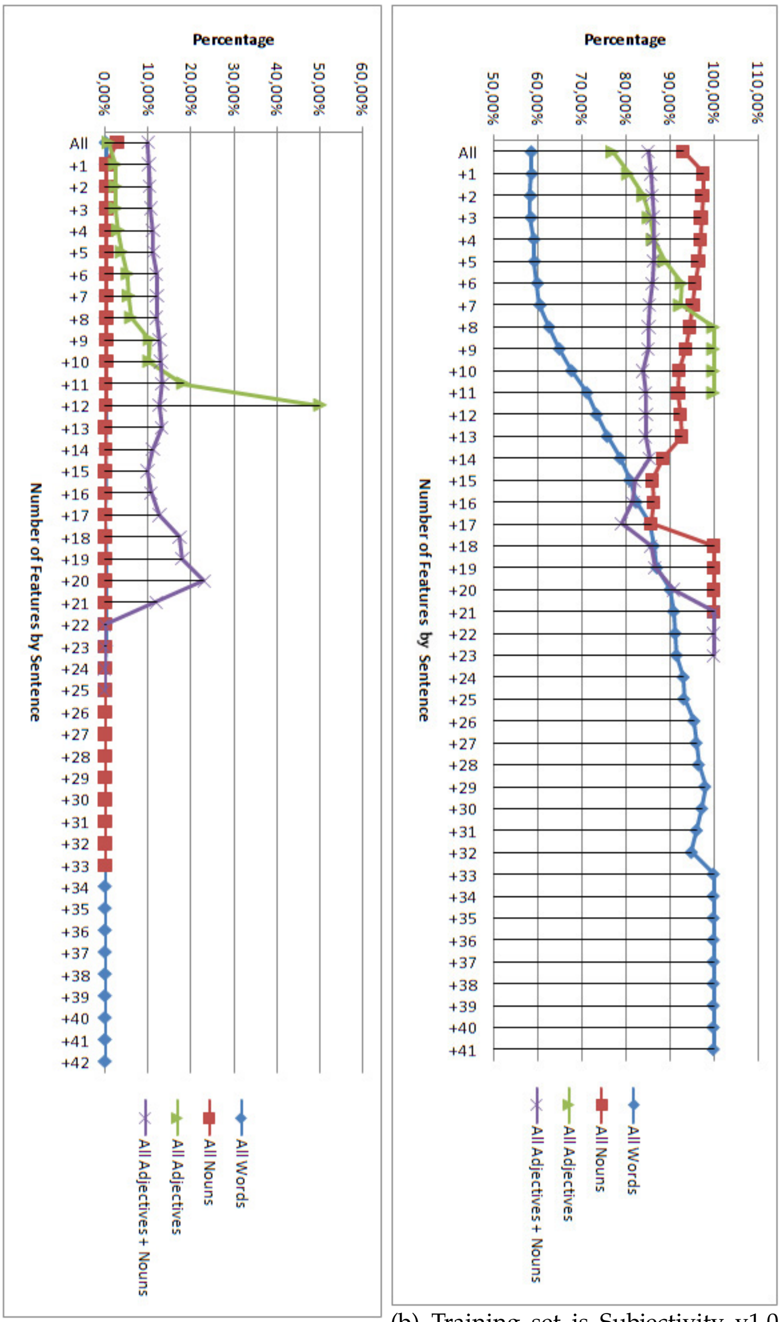
We observe graph b) of Figure 5.5 and we conclude that each representative line of a morphologic level has a different behavior.

Guidelines of the graph:

- the morphologic levels "All Words", "All Nouns" and "All Adjectives + Nouns" have a similar behavior and their success percentages stand between 70% and 100%;
- the morphologic level "All Adjectives" is, once again, characterized by its small number of features; its success percentage is between 70% and 90% even when the sentences have more than 9 features, then it suffers oscillations. When the sentences have more than 10 features, the success percentage is of approximately 35% and when the sentences have more than 11 features, the success percentage is of approximately 75%.

In terms of accuracy, the morphologic levels that achieve a better success percentage are "All Words", "All Nouns" and "All Adjectives + Nouns", because their representative lines manage to have a success percentage of 70% to 100%. Besides, they include a considerable number of features (more than 51, more than 32 and more than 27 respectively) in opposition to what happens to "All Adjectives" which has oscillations and thus is punished.

By comparing the two graphs of Figure 5.5 we may conclude that, in terms of accuracy, graph a) achieves better classification results with a great number of features than graph b) as it doesn't include a large number of features. On the other hand, graph a) almost always presents the same standard for all the morphologic levels, while graph b) presents some variations. Finally, the morphologic level, common to both graphs, which achieves a better success percentage, is "All Adjectives +Nouns".



(a) Training set is Wikipedia copus and Weblogs corpus. Testing set is Subjectivity v1.0 corpus.

(b) Training set is Subjectivity v1.0 copus and testing set is Subjectivity v1.0 corpus.

Figure 5.6: *Graphs of the Polynomial function and True Positive or Recall*

In the analysis of the Figure 5.6, we will consider the True Positive Recall definition, represented in this chapter by the Equation 5.2.

We observe graph a) of Figure 5.6 and we conclude that each representative line of a morphologic level has a different behavior, although the success percentage - more proportion of the total number of predictions that were correct - is between 0% and 10% - independently of the number of features that are necessary for the sentences classification (objectivity vs. subjectivity).

Guidelines of the graph:

- in the morphologic level "All Words" the success percentage is very close to 0%;
- in the morphologic level "All Nouns" the success percentage is also very close to 0%;
- in the morphologic levels "All Adjectives" the representative line goes up to 50%, a pick where there are no sentences with more than 12 Adjectives.

In terms of Recall, the morphologic level that achieves a better success percentage is "All Adjectives + Nouns" (in spite of the oscillations), because its corresponding line has a greater inclination - the same does not happen in "All Words" or "All Nouns" - and it includes a considerable number of features (more than 25). On the contrary, the morphologic level "All Adjectives" has a bigger inclination when compared with all the other levels, but in compensation it includes a more reduced number of features of sentence, which we have already considered as being an aspect that stands against the use of only adjectives to make the classification of sentences.

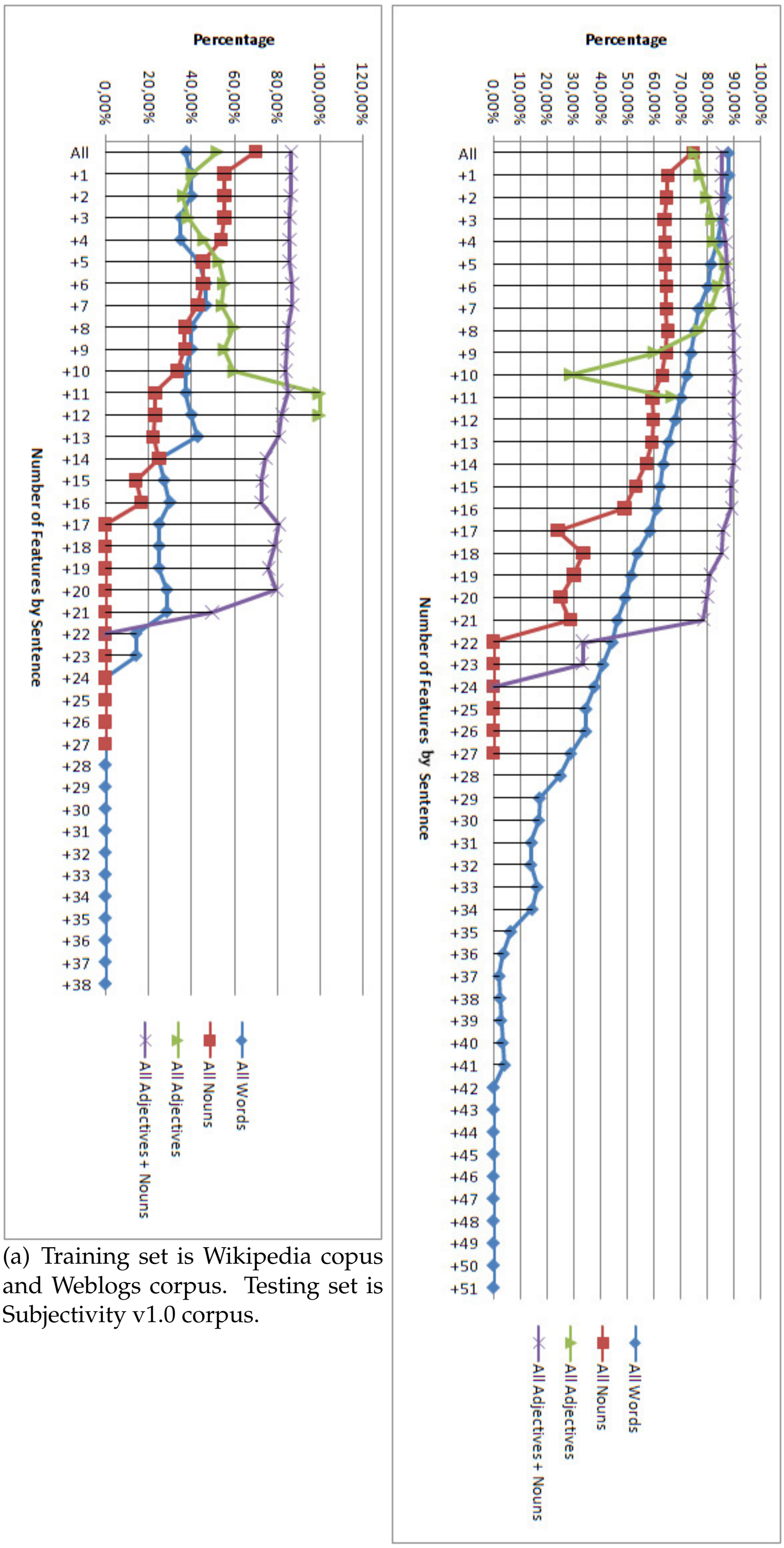
We observe graph b) of Figure 5.6 and we conclude that each representative line of a morphologic level has a different behavior and there are lots of oscillations.

Guidelines of the graph:

- there are lots of oscillations;
- the morphologic levels "All Words" and "All Adjectives" have less oscillations; in the classification of all the sentences we achieve a percentage of success of approximately 60% and 80% respectively . These percentages increase (they suffer few oscillations) until reaching both 100%.

In terms of Recall, the morphologic levels have many oscillations, except for "All Words" and "All Adjectives" that have a similar behavior; these levels have a better success percentage and they don’t have so many oscillations.

When comparing the two graphs of Figure 5.6 we may conclude that, in terms of Recall, graph b) shows better success percentages than graph a).



(a) Training set is Wikipedia copus and Weblogs corpus. Testing set is Subjectivity v1.0 corpus.

(b) Training set is Subjectivity v1.0 copus and testing set is Subjectivity v1.0 corpus.

Figure 5.7: *Graphs of the Polynomial function and Precision*

In the analysis of Figure 5.7 we will consider the precision definition, represented in this chapter by the Equation 5.3.

We observe graph a) of Figure 5.7 and we conclude that each representative line presents oscillations, although the success percentages - more proportion of the predicated positive cases that were correct - are different.

Guidelines of the graph:

- in the morphologic levels "All Words", "All Nouns" and "All Adjectives + Nouns" the more features there are in the sentences that are being classified, the smaller the success percentage is; reaching 0%;
- in the morphologic level "All Adjectives" there are no sentences with more than 12 adjectives (reduced number of features). In the classification of all the sentences the success percentage is of approximately 50%. But, for the moment it is necessary that the sentences contain a certain number of adjectives for them to be classified; the success percentage goes up reaching 100%;
- clearly, in terms of Precision, our best morphologic level is "All Adjectives", because its success percentage is between 50% and 100%.

In terms of precision, the morphologic level that gets the best success percentage is "All Adjectives", because its representative line achieves to be always between 35% and 100% - the same does not happen to the other morphologic levels. But we consider it to be an aspect against the fact that there are no sentences with more than 12 adjectives, because it is a reduced number of features.

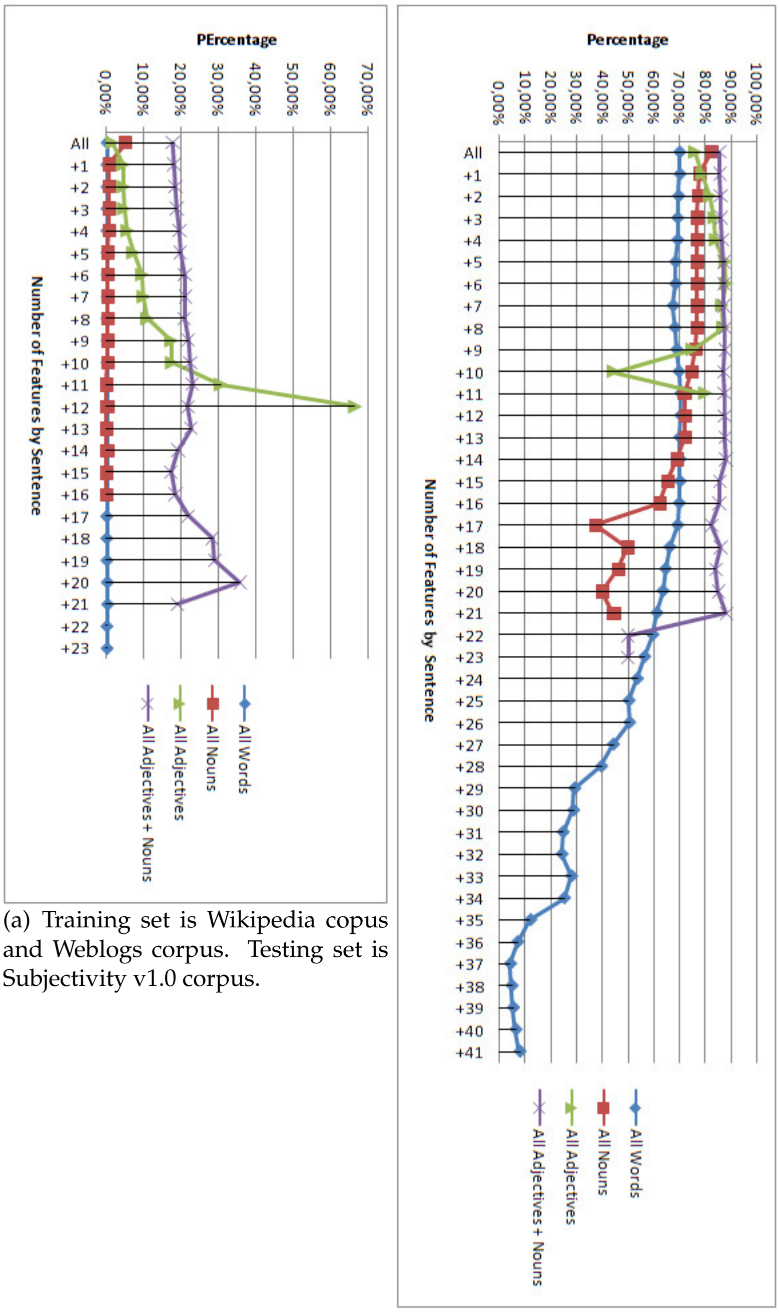
We observe graph b) of Figure 5.7 and we conclude that each representative line of a morphologic level has an identical behavior.

Guidelines of the graph:

- in the classification of all the sentences in all morphologic levels the success percentage is between 70% and 90%;
- in the classification of sentences, whose number of necessary features for its classification increases, the success percentage for "All Words", "All Nouns" and "All Adjectives + Nouns" decreases up to 0%;
- in the morphologic level "All Adjectives" the success percentage is approximately 72% and it lowers when we only classify sentences which respect a certain number of features. But, when we classify sentences with more than 11 features the success percentage goes up again approaching 70%.

In terms of Precision, the morphologic level that achieves better performance is "All Adjectives", as it never reaches 0% in spite of having a reduced number of features.

Comparing the two graphs of Figure 5.7 we may conclude that, in terms of Precision, the morphologic level "All Adjectives" has the best performance, despite its reduced number of features; its success percentage never approaches 0%. We can also consider that the morphologic level "All Adjectives" has a better performance in graph a), because it reaches up to 100%.



(a) Training set is Wikipedia copus and Weblogs corpus. Testing set is Subjectivity v1.0 corpus.

(b) Training set is Subjectivity v1.0 copus and testing set is Subjectivity v1.0 corpus.

Figure 5.8: *Graphs of the Polynomial function and F-Measure*

In the analysis of Figure 5.8 we will consider the F-Mesure definition, represented in this chapter by the Equation 5.4.

We observe graph a) of the Figure 5.8 and we conclude that each representative line of a morphologic level has a different behavior, although the success percentage of the morphologic levels is, in most cases, between 0% and 20% - independently of the number of features that are considered for the classification of the sentences (Objectivity vs. Subjectivity).

Guidelines of the graph:

- in the morphologic levels "All Words" and "All Nouns" the success percentage is very close to 0%;
- the morphologic levels "All Adjectives" and "All Adjectives + Nouns" have a similar behavior, because the respective lines go up until they reach a pick in that the success percentage is of approximately 68% for "All Adjectives" and of 38% for "All Adjectives + Nouns". From these picks, there are no sentences with more than 21 adjectives, but there are sentences with more than 21 Adjectives+Nouns. When we classified sentences with more than 21 features for the morphologic level "Adjectives + Nouns", the success percentage came down again to 20%.

In terms of F-Measure, the morphologic level that achieves a better success percentage is "All Adjectives + Nouns", because its success percentage achieves to be better than the one of the other morphologic levels, and it includes a considerable number of features (more than 21). On the contrary, the morphologic level "All Adjectives" has a bigger inclination when compared with all the other levels. But in compensation it includes a more reduced number of features per sentence, which we considered to be an aspect against the use of only adjectives to make the classification of sentences.

We observe graph b) of Figure 5.8 and we conclude that each representative line of a morphologic level has an identical behavior and there are lots of oscillations.

Guidelines of the graph:

- there are lots of oscillations and we verified that for all the morphologic levels the success percentage is between 70% and 90%, when we classify the sentences with all the characteristics. These percentages decrease when we classify sentences that must have a certain number of features;

- the morphologic level "All Adjectives + Nouns" has a better performance than all the other morphologic levels.

In terms of F-Measure, the morphologic levels have a lot of oscillations, except for "All Adjectives + Nouns" that achieves a better success percentage and it doesn't have so many oscillations.

When comparing the two graphs of Figure 5.8 we may conclude that, in terms of F-Measure, graph b) shows better success percentages than graph a). On the other hand, graph a) presents almost always the same standard for "All Adjectives" and "All Adjectives + Nouns". Finally, the morphologic level "All Adjectives + Nouns" in graph b) has few oscillations and its success percentage doesn't fall below 50% in none of the cases.

Radial Basis Function

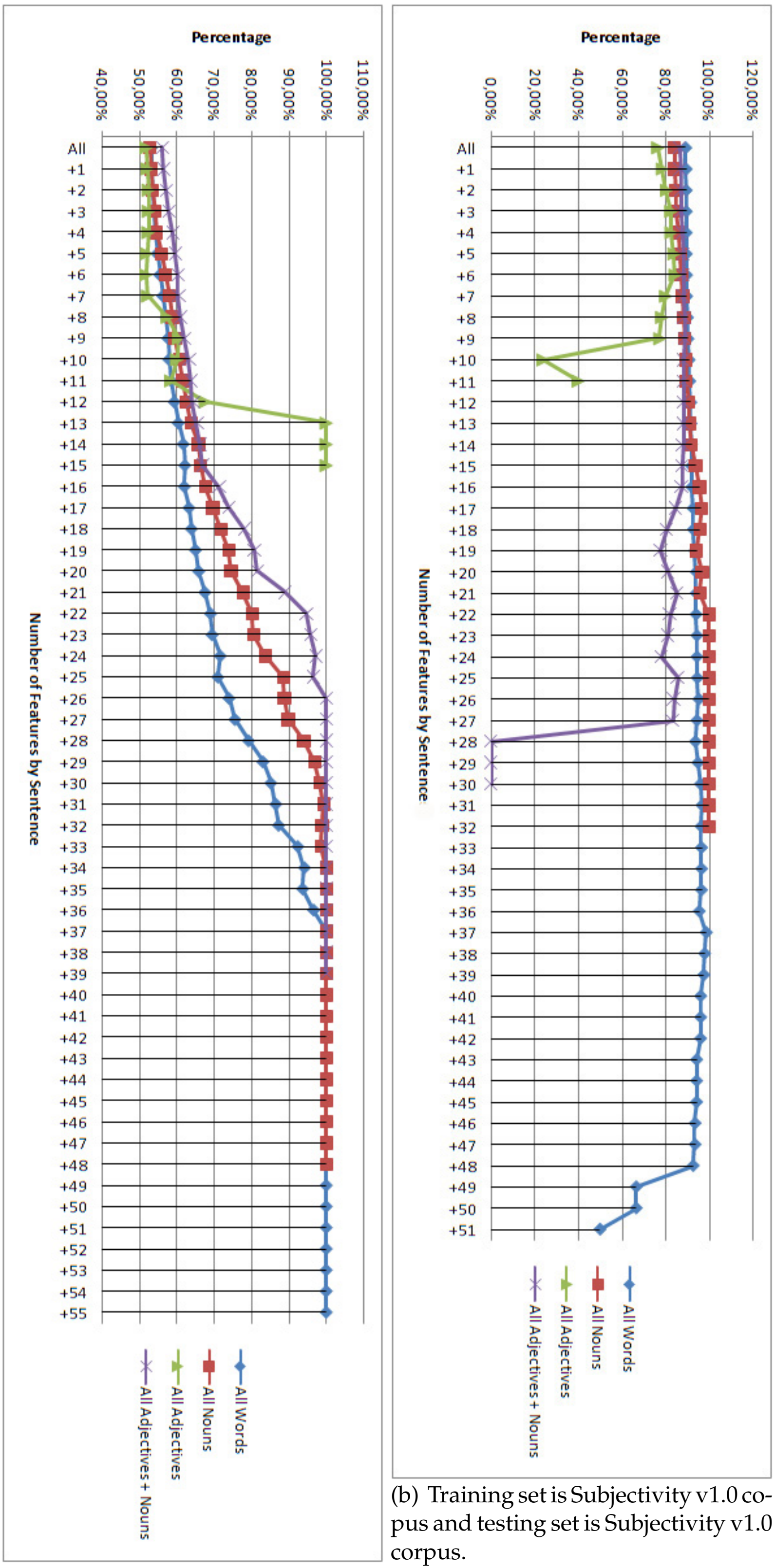
Following is the results presented in the form of confusion matrix and graphs that allow us to draw conclusions from the entire process of classification of sentences to use a radial basis function in SVM^{Ligth}.

		Predicted	
		Wikipedia	**Weblogs**
Actual	**Objectivity**	456.45	11.80
	Subjectivity	421.13	4.35

Table 5.7: *Confusion Matrix for all Words and for Radial Basis Function, where the predicted class is Wikipedia and Weblogs and actual class is Objectivity and Subjectivity.*

		Predicted	
		Objectivity	**Subjectivity**
Actual	**Objectivity**	408.58	58.24
	Subjectivity	65.24	363.35

Table 5.8: *Confusion Matrix for all Words and for Radial Basis Function, where the predicted class is Objectivity and Subjectivity and actual class is Objectivity and Subjectivity.*



(a) Training set is Wikipedia copus and Weblogs corpus. Testing set is Subjectivity v1.0 corpus.

(b) Training set is Subjectivity v1.0 copus and testing set is Subjectivity v1.0 corpus.

Figure 5.9: *Graphs of the Radial Basis function and Accurary*

In the analysis of Figure 5.9, we will consider the accuracy definition, represented in this chapter by the Equation 5.1.

We observe graph a) of Figure 5.9 and we conclude that each representative line of a morphologic level has an identical behavior, in other words, the more features are considered for the classification of the sentences (objectivity vs. subjectivity), the bigger the success percentage is - more proportion of the total number of predictions that were correct.

Guidelines of the graph:

- in the classification of sentences with all the features, the success percentage for all the morphologic levels is between 50% and 60%;
- in the classification of sentences with more than 20 features, the success percentage is between 60% and 85% for all of the morphologic levels; except for the adjectives, because there are no sentences with more than 13 adjectives;
- in the classification of sentences with more than 35 features the success percentage varies between 90% and 100%.

In terms of accuracy, the morphologic level that achieves a better success percentage is "All Adjectives + Nouns", because its representative line achieves to have a bigger inclination - the same does not apply to "All Words" or "All Nouns" - and it includes a considerable number of features (more than 33). On the contrary, the morphologic level "All Adjectives" has a bigger inclination in comparison to all the other levels. But in compensation it includes a more reduced number of features per sentence, which we consider to be an aspect that stands against the use of only adjectives when making the classification of sentences.

We observe graph b) of Figure 5.9 and we conclude that each representative line of a morphologic level has a different behavior.

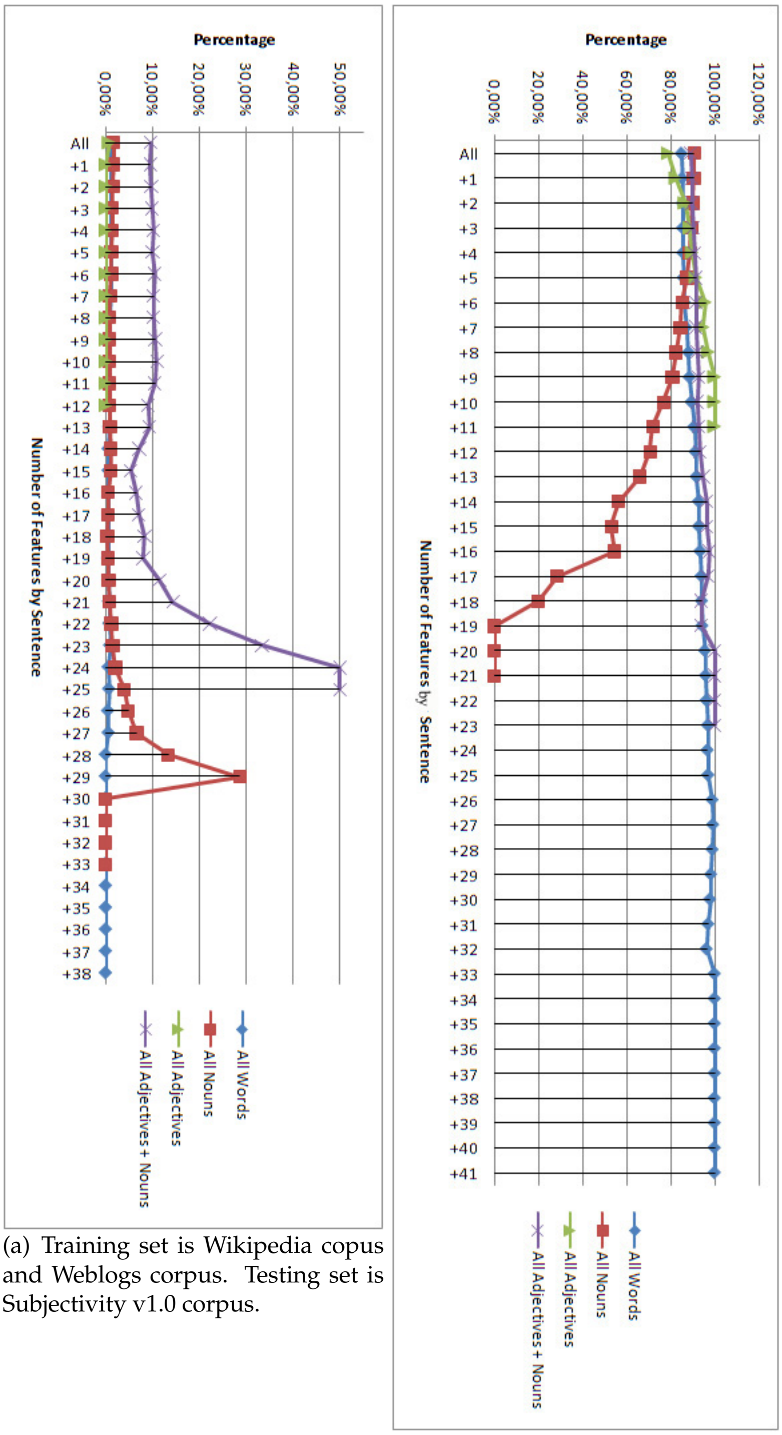
Guidelines of the graph:

- the morphologic levels "All Words", "All Nouns" and "All Adjectives + Nouns" have a identical behavior even when the classified sentences have more than 27 features; their success percentages are between 80% and 100%;
- in the classification of sentences and in the morphologic level "All Adjectives + Nouns" the success percentage is 0% for the sentences with more than 28, 29 and 30 features;

- the success percentage of the morphologic level "All Words" decreases a bit, because when the sentences have more than 49, 50 and 51 features their success percentages are proximally 65%, 65% and 50% respectively, though never reaching the 100%;
- the morphologic level "All Adjectives" is characterized by its small number of features, its success percentage is between 70% and 90% even when the sentences have more than 9 features; from then on it suffers oscillations. When the sentences have more than 10 features, the success percentage is of approximately 21% and when the sentences have more than 11 features, the success percentage is of approximately 40%.

In terms of accuracy, the morphologic levels that achieve a better success percentage are "All Nouns" and "All Adjectives + Nouns", because their representative lines get to have a success percentage between 80% and 100%. Besides, they include a considerable number of features (more than 32 and more than 30 respectively), in opposition to what happens with "All Adjectives" that always has a smaller success percentage in relation to the other morphologic levels. "All Words" is also punished, because with a high number of features for sentence its success percentage decreases to reach 0%.

Comparing the two graphs of Figure 5.9 we may conclude that, in terms of accuracy, graph a) achieves better classification results with a great number of features, unlike graph b) that declines until reaching 0% - starting from the 48 features. On the other hand, graph a) almost always presents the same standard for all the morphologic levels, while graph b) presents some variations. Finally, the morphologic level, common to both graphs, which achieves a better success percentage, is "All Adjectives + Nouns".



(a) Training set is Wikipedia copus and Weblogs corpus. Testing set is Subjectivity v1.0 corpus.

(b) Training set is Subjectivity v1.0 copus and testing set is Subjectivity v1.0 corpus.

Figure 5.10: *Graphs of the Radial Basis function and True Positive or Recall*

In the analysis of the Figure 5.10, we will consider the True Positive or Recall definition, represented in this chapter by the Equation 5.2.

We observe graph a) of Figure 5.10 and we conclude that each representative line of a morphologic level has a different behavior, although the success percentage - more proportion of the total number of predictions that were correct - of the morphologic levels is between 0% and 50%, independently of the number of features that are necessary for the sentences classification (objectivity vs. subjectivity).

Guidelines of the graph:

- in the morphologic levels "All Words" and "All Adjectives" the success percentage is very close to 0%;
- in the morphologic level "All Nouns" the success percentage is also very close to 0%, until it reaches the classification of sentences with 25 features; from this point on, the success percentage gets up to 30%. But afterwards it decreases to 0%;
- the representative line of morphologic level "All Adjectives + Nouns" goes up until reaching a pick, that is, when the sentences have more than 20 features.

In terms of Recall, the morphologic level that gets a better success percentage is "All Adjectives + Nouns" (in spite of the oscillations) , because its corresponding line has a greater inclination - the same does not happen to "All Words" or "All Nouns" - and it includes a considerable number of features (more than 20).

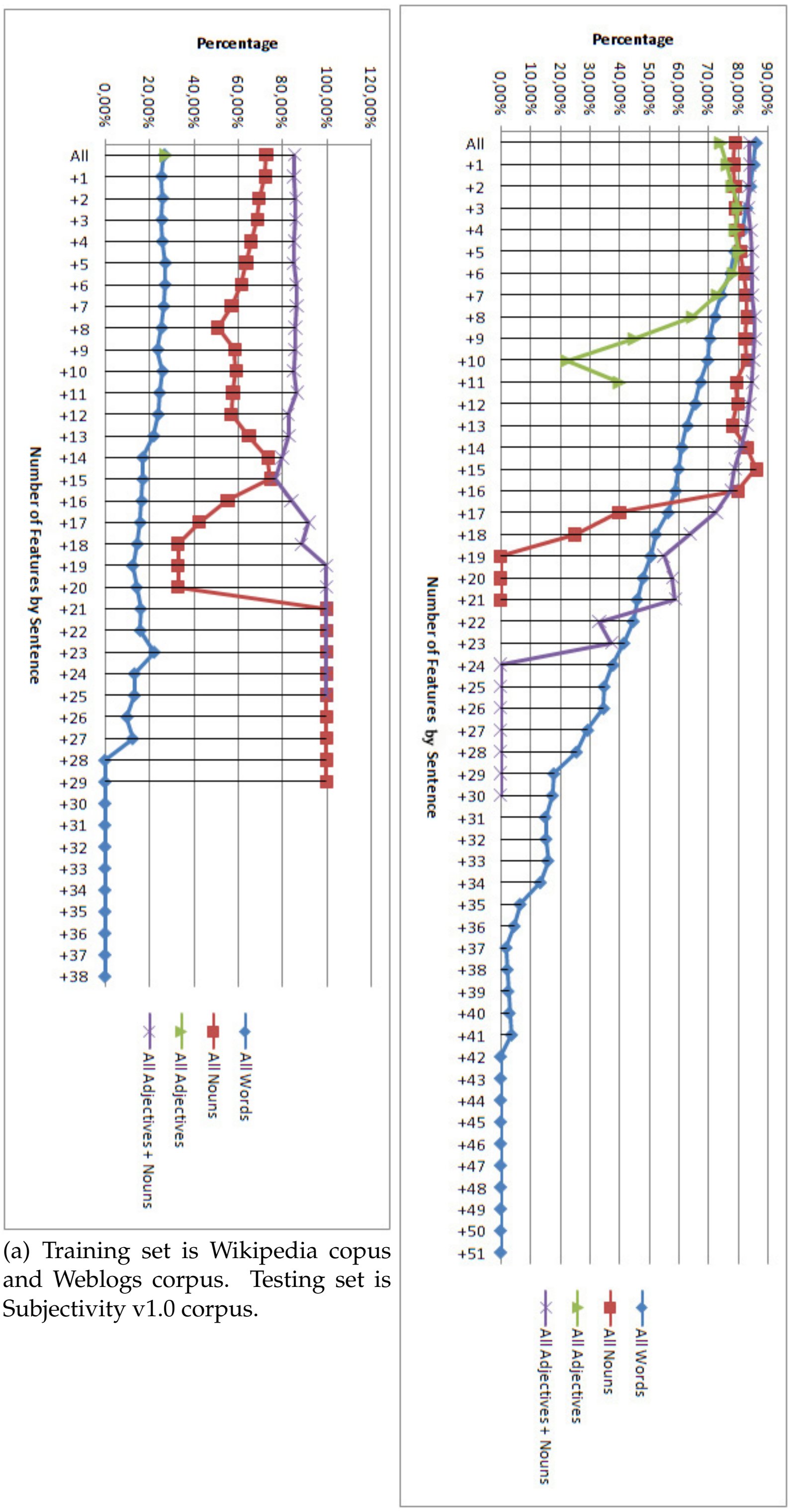
We observe graph b) of Figure 5.10 and we conclude that each representative line of a morphologic level has a different behavior and that there are lots of oscillations.

Guidelines of the graph:

- there are not lots of oscillations;
- in the morphologic levels "All Words", "All Adjectives" and "All Adjectives + Nouns", the success percentages are between 80% and 100%, so these levels have a similar behavior;
- in the morphologic level "All Nouns" we achieve to have a success percentage of proximally 90%; decreasing afterwards to 0%.

In terms of Recall, the morphologic levels don’t have many oscillations and the morphologic levels "All Words", "All Adjectives" and "All Adjectives + Nouns", give the best success percentage.

When comparing the two graphs of Figure 5.10 we may conclude that, in terms of Recall, graph b) shows better success percentages than graph a).



(a) Training set is Wikipedia copus and Weblogs corpus. Testing set is Subjectivity v1.0 corpus.

(b) Training set is Subjectivity v1.0 copus and testing set is Subjectivity v1.0 corpus.

Figure 5.11: *Graphs of the Radial Basis function and Precision*

In the analysis of Figure 5.11 we will consider the precision definition, represented in this chapter by the Equation 5.3.

We observe graph a) of Figure 5.11 and we conclude that each representative line of a morphologic level has a different behavior. Although the success percentages - more proportion of the predicated positive cases that were correct - are different, some levels achieve to maintain the same inclination.

Guidelines of the graph:

- in the morphologic level "All Words" the more features there are in the sentences that are being classified, the smaller the success percentage is - until reaching 0%;
- in the morphologic level "All Nouns" there are some picks - we found several oscillations. As far as we classify sentences with more than 8 characteristics, the line of the respective morphologic level goes down until reaching a success percentage of approximately 50%; afterwards it goes up a little and it reaches 80% approximately, but then the success percentage falls reaching, approximately, 30%. Finally, the percentage rises until reaching the 100%;
- in the morphologic level "All Adjectives", the success percentage is near to 30%;
- in the morphologic level "All Adjectives + Nouns" the success percentage is between 80% e 100%;
- clearly, in terms of Precision, our best morphologic level is "All Adjectives + Nouns".

In terms of precision, the morphologic level that achieves the best success percentage is "All Adjectives + Nouns", because the representative line is always between 80% and 100% - the same does not happen to the other morphologic levels - and, as always, it includes a considerable number of features (more than 25).

We observe graph b) of Figure 5.11 and we conclude that each representative line of a morphologic level has an identical behavior.

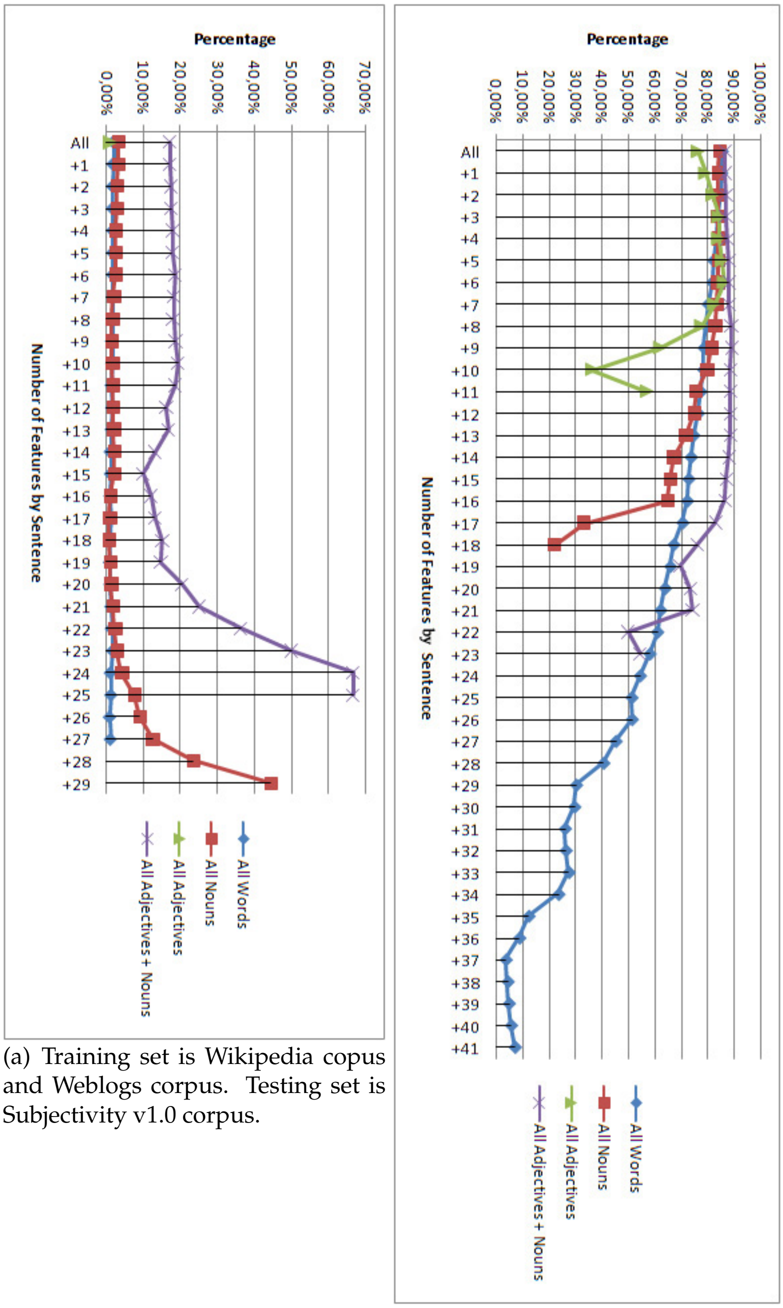
Guidelines of the graph:

- in the classification of all the sentences, in all the morphologic levels, the success percentage is between 70% and 90%;

- in the classification of sentences, whose number of necessary features for its classification increases, the success percentage for "All Words", "All Nouns" and "All Adjectives + Nouns" decreases until reaching 0%;
- in the morphologic level "All Adjectives" the success percentage is approximately 72% and it lowers when we only classify sentences that observe a certain number of features; but when we classify sentences with more than 11 features the success percentage goes up again approaching 40%.

In terms of Precision, the morphologic level that has the best performance is "All Adjectives", because it never reaches 0% - in spite of having a reduced number of features.

By comparing the two graphs of Figure 5.11 we may conclude that, in terms of Precision, graph a) achieves better classification results, especially in the morphologic level "All Adjectives + Nouns", unlike graph b) that reaches 0% in almost all the morphologic levels. On the other hand, graph b) and graph a) almost always present the same standard for all of the morphologic levels, although both have oscillations. Finally, the morphologic level that gets better performance in graph b) is "All Adjectives" and in graph a) is "All Adjectives + Nouns".



(a) Training set is Wikipedia copus and Weblogs corpus. Testing set is Subjectivity v1.0 corpus.

(b) Training set is Subjectivity v1.0 copus and testing set is Subjectivity v1.0 corpus.

Figure 5.12: *Graphs of the Radial Basis function and F-Measure*

In the analysis of the Figure 5.12 we will consider the F-Mesure definition, represented in this chapter by the Equation 5.4.

We observe graph a) of Figure 5.12 and we conclude that all the representative lines of a morphologic level have an identical behavior, although their success percentage is, in most cases, between 0% and 20% - independently of the number of features that are considered for the classification of the sentences (Objectivity vs. Subjectivity).

Guidelines of the graph:

- in the morphologic level "All Words" the success percentage is very close to 0%;
- in the morphologic level "All Nouns" the success percentage is very close to 0% when classifying sentences with more than 24 features; afterwards, the success percentage rises coming near to 50%;
- in the morphologic level "All Adjectives + Nouns" the representative line comes up until reaching a pick, which success percentage is near 65% - considering that there are no sentences with more than 25 Adjectives + Nouns.

In terms of F-Measure, the morphologic level that achieves a better success percentage is "All Adjectives + Nouns", because its success percentage is better than the one of the other morphologic levels, and it includes a considerable number of features (more than 25).

We observe graph b) of Figure 5.12 and we conclude that each representative line of a morphologic level has an identical behavior and there are lots of oscillations.

Guidelines of the graph:

- there are lots of oscillations and we verified that for all the morphologic levels, when we classify the sentences with all the characteristics, the success percentage is between 70% and 90%. These percentages decrease when we classify sentences that must have a certain number of features;
- the morphologic level "All Adjectives + Nouns" has a better performance than all the other morphologic levels.

In terms of F-Measure, the morphologic levels have lots of oscillations, except for "All Adjectives + Nouns" that achieves a better success percentage and it doesn't have so many oscillations.

When comparing the two graphs of Figure 5.12 we may conclude that, in terms of F-Measure, graph b) shows better success percentages than graph a). The morphologic level "All Adjectives + Nouns" in graph b) has few oscillations and its success percentage doesn't fall below the 60% in none of the cases.

5.3 Conclusion and Discussion

Before we begin reporting our conclusions, we will recover the tables 5.3, 5.4, 5.5, 5.6, 5.7 and 5.8 that represent the Confusion Matrix of the respective kernel functions. It is necessary not to forget that these matrices only concern the classification of all the sentences with the morphologic level "All Words".

			Predicted	
			Wikipedia	**Weblogs**
Actual	**Linear Function**	**Objectivity**	455.23	13.01
		Subjectivity	417.90	4.40
	Polynomial Function	**Objectivity**	466.32	1.56
		Subjectivity	425.84	0.24
	Radial Basis Function	**Objectivity**	456.45	11.80
		Subjectivity	421.13	4.35

Table 5.9: *Confusion Matrix for all Words, for Linear Function, Polynomial Function and for Radial Basis Function, where the predicted class is Wikipedia and Weblogs and actual class is Objectivity and Subjectivity.*

			Predicted	
			Objectivity	**Subjectivity**
Actual	**Linear Function**	**Objectivity**	407.43	59.24
		Subjectivity	65.45	360.12
	Polynomial Function	**Objectivity**	433.35	35.36
		Subjectivity	176.83	249.30
	Radial Basis Function	**Objectivity**	408.58	58.24
		Subjectivity	65.24	363.35

Table 5.10: *Confusion Matrix for all Words, for Linear Function, Polynomial Function and for Radial Basis Function, where the predicted class is Objectivity and Subjectivity and actual class is Objectivity and Subjectivity.*

In the analysis and comparison of the two previous tables, we have quickly noticed that, to classify objective sentences, we achieve better results when the training set is constituted by sentences of the Wikipedia corpus and of the Weblogs corpus than when the training set is constituted by sentences of Subjectivity v1.0 corpus. To classify subjective sentences we verified that the training set, constituted by sentences of Subjectivity v1.0 corpus, achieves better results than when the training set is constituted by sentences of the Wikipedia corpus and of the Weblogs corpus.

In this first analysis, we concluded that our methodology doesn't satisfy a hundred percent for the moment. However, we are not at all unhappy, as we achieved to identify that in all of the kernel functions we obtained better results of classification of objective sentences by using the training set constituted by sentences of the Wikipedia corpus and of the Weblogs corpus than by using the training set whose sentences belong to Subjectivity v1.0 corpus.

Thus, we verified that in the classification of all the sentences that are part of our test set (those sentences belong to Subjectivity v1.0 corpus), the objective sentences are well classified, but the subjective sentences are badly classified.

When we are classifying sentences whose number of necessary features to be classified increases (we cannot forget that the test set is going to decrease every time we increase the number of features) we are going to get an improvement in the results (in some morphologic levels) obtained in all the kernel functions, as have shown the graphs presented in this chapter.

A deeper analysis that reports to the terms which derive from the Confusion Matrix show that what we said before is reflected in the graph that belongs to the Accuracy, Recall, Precision and F-Measure.

It should be noted that, for the analysis of the graphs representing the precision in all of the kernel functions, we achieved better results when the training set was constituted by sentences of the Wikipedia corpus and Weblogs corpus than when the sentences were of Subjectivity v1.0 corpus. This is important because the precision shows that the sentences classified as objective are really objective and the sentences classified as subjective are really subjective; in other words, there is a great precision in the classifications we have made.

From the analysis of the graphs appeared the following tables:

		Kernel Functions		
		Linear	**Polynomial**	**Radial Basis**
Terms	**Accuracy**	All Adj + Nouns	All Adj + Nouns	All Adj + Nouns
	Recall	All Adj + Nouns	All Adj	All Adj + Nouns
	Precision	All Adj + Nouns	All Adj	All Adj + Nouns
	F-Measure	All Adj + Nouns	All Adj + Nouns	All Adjs + Nouns

Table 5.11: *Presentation of the highest levels morphological when training set is Wikipedia corpus and Weblogs corpus and testing set is Subjectivity v1.0 corpus. Adj is same which Adjectives.*

		Kernel Functions		
		Linear	**Polynomial**	**Radial Basis**
Terms	**Accuracy**	All Nouns	All Nouns	All Nouns
	Recall	All Words	All Nouns	All Adj + Nouns
	Precision	All Adj + Nouns	All Adj	All Adjectives
	F-Measure	All Adj + Nouns	All Adj + Nouns	All Adj + Nouns

Table 5.12: *Presentation of the highest levels morphological when training set is Subjectivity v1.0 corpus and testing set is Subjectivity v1.0 corpus. Adj is same which Adjectives.*

These tables show which is, for us, the best morphologic level conjugating the term that occurs of the Confusion Matrix and the kernel functions. We can verify that, when the training set contains sentences of the Wikipedia corpus and Weblogs corpus, the morphologic level that prevails is "All Adjectives + Nouns"; this is no longer the case when the training set contains sentences of Subjectivity v1.0 corpus, as we can see in table 5.12.

In this chapter, we achieved to show that the Wikipedia corpus and Weblogs corpus can replace the annotated corpora, although there are still some gaps that are necessary to deal with. In the chapter "Conclusion and Future Work" we point out some strategies to be taken to better deal with these gaps.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

Theory and Practice are two concepts that should walk side by side in order to contribute to the success of those who start a new project, instead of fighting each other in order to obtain a more important place.

At the beginning of a scientific activity, it is also necessary to create a solid base of knowledge. Having this in mind and for the development of this master's thesis we have: studied the scientific production in the areas of Information Retrieval, Opinion Mining, Opinion and Retrieval and Sentiment Analysis in detail; compiled all the information; deepened our knowledge; studied what is published in this area and aspired to the possibility of proposing new solutions.

We propose a new concept for the evaluation of the quality of a Web page, through the absence of opinion and where manually annotated corpora is not used but corpora that can be available and accessible in a fast and effective way. This new concept brings the Scientific Community another form of research in this area. The common user may also benefit from it, as he/she - when consulting a web page or making a research in a search motor-will get the possibility of knowing, with some certainty, if the information he/she visualizes is true or if it is just an opinion/sentiment set expressed by the authors, thus excluding his/her own judgment.

We built our not annotated manually corpora (Wikipedia corpus and Weblogs corpus) for the Scientific Community that researches in the area and replaced the usual annotated manually corpora (e. g. Subjectivity v1.0 corpus). We started by evaluating their similarity to be able to continue with our methodology. The results of that evaluation were very positive, as we achieved to show that the Wikipedia corpus is very similar to the objective sentences that Subjectivity v1.0 corpus contains, and very little similar to the subjective sentences of the same corpus. On the other hand, the Weblogs corpus is very similar to the subjective sentences contained in the Subjectivity v1.0 corpus and very little similar to the objective sentences it contains.

After the success described in the chapter "Similarity between the corpora", we passed to the classification of sentences, expecting that the training set constituted by sentences of the Wikipedia and Weblogs corpus was capable of correctly classifying sentences of the test sets constituted by sentences of Subjectivity v1.0 corpus.

In the classification process, we showed that we have achieved good results in some circumstances: namely when we used as features the morphologic level "All Adjectives + Nouns" to make a classification of sentences using as training set sentences those that belong to the Wikipedia corpus and Weblogs corpus.

The following figure shows, in three dimensions, how difficult it is to separate Wikipedia and Weblogs. We have 125 sentences belonging to the Wikipedia corpus and 125 sentences belonging to the Weblogs corpus; we used as characteristics the morphologic level "All Words" and we applied the measure TF/ISF (these values were calculated in Chapter 4).

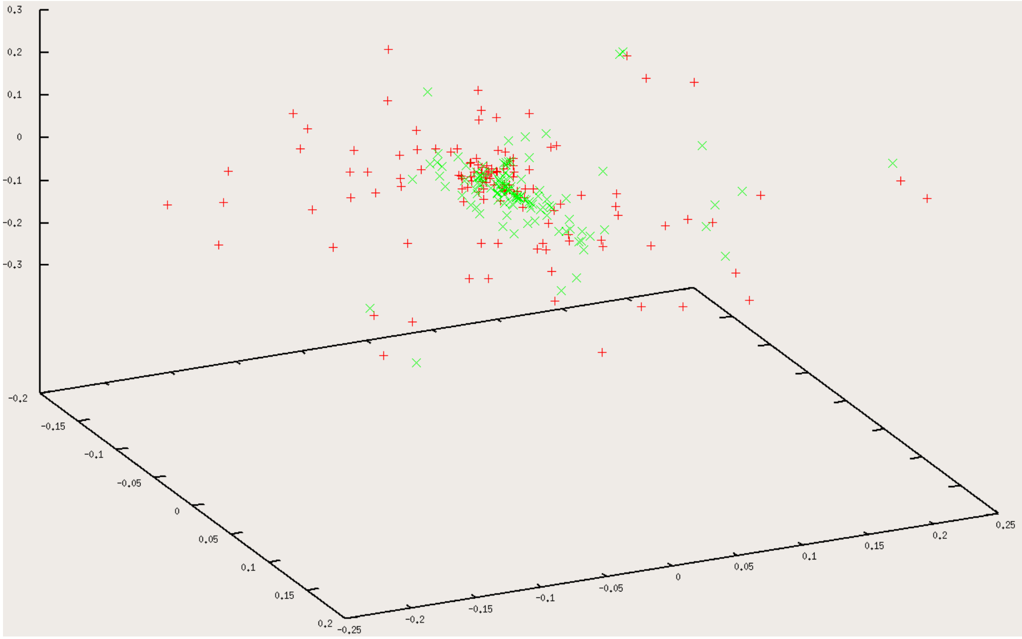


Figure 6.1: *Representation of set Wikipedia corpus and set Weblogs corpus*

The red symbol "+" represents the sentences of the Wikipedia corpus and the green symbol "x" represents the sentences of the Weblogs corpus.

Figure 6.1. shows that, at the moment, it is very difficult to separate the Wikipedia set from the Weblog set, and this means there is still much work to do.

As part of this master's thesis, we conclude this dissertation with the beginning of our scientific activity and therefore nothing ends where everything may begin.

6.2 Future Work

To accomplish this research work, we defined a plan that doesn't finish with the end of this master's thesis. We defined a set of ideas to improve the obtained results and to achieve the aim of this dissertation with success. Despite the results, this work can be much more developed and enriched.

The first suggestion is to apply a technique of selection of features (features selection techniques) called Odds Ratio [13].

Odds Ratio is able to find terms commonly included in messages belonging to a certain category. The meaning of this measure is the following: words that appear in both spam and legitimate classes are assigned an Odds Ratio score near to 1, otherwise, terms with are representative of a certain class present an Odds Ratio value higher than 1. Odds Ratio is computed as expression 6.1 shows.

$$OR(t_i, c_j) = \frac{p(t_i|c_j).[1 - p(t_i|\bar{c}_j)]}{[1 - p(t_i|c_j)].p(t_i|\bar{c}_j)} \tag{6.1}$$

The second suggestion is the use of other classifiers implemented in Software Weka[1] [13], such as the Naive Bayes Multinomial [2] [13] and the K-Nearest Neighbor [2] [13].

A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model".

In simple terms, a Naive Bayes classifier assumes that the presence (or lack of presence) of a particular feature of a class is unrelated to the presence (or lack of presence) of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 4 in diameter. Even though these features depend on the existence of the other features, a Naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple.

Depending on the precise nature of the probability model, Naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for Naive Bayes models uses the method of maximum likelihood; in other words, one can work with the Naive Bayes model without believing in Bayesian probability or using any Bayesian methods.

---

[1] http://www.cs.waikato.ac.nz/ml/weka/

In spite of their naive design and apparently over-simplified assumptions, Naive Bayes classifiers often work much better in many complex real-world situations than one might expect. Recently, careful analysis of the Bayesian classification problem has shown that there are some theoretical reasons for the apparently unreasonable efficacy of Naive Bayes classifiers. An advantage of the Naive Bayes classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

The K-Nearest Neighbor algorithm is amongst the simplest of all machine learning algorithms. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors. k is a positive integer, typically small. If $k = 1$, then the object is simply assigned to the class of its nearest neighbor. In binary (two class) classification problems, it is helpful to choose k to be an odd number as this avoids tied votes.

The same method can be used for regression, by simply assigning the property value for the object to be the average of the values of its k nearest neighbors. It can be useful to weight the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones.

The neighbors are taken from a set of objects for which the correct classification (or, in the case of regression, the value of the property) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. In order to identify neighbors, the objects are represented by position vectors in a multidimensional feature space. It is usual to use the Euclidean distance, though other distance measures, such as the Manhattan distance could in principle be used instead. The K-Nearest Neighbor algorithm is sensitive to the local structure of the data.

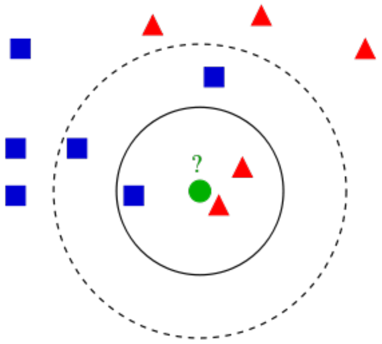


Figure 6.2: *Example of K-Nearest Neighbor classification.*

The test sample (green circle) should be classified either to the first class of blue squares or to the second class of red triangles. If $k = 3$ it is classified to the second class because there are 2 triangles and only 1 square inside the inner circle. If $k = 5$ it is classified to first class (3 squares vs. 2 triangles inside the outer circle).

To the third suggestion, we defend the use of another method to apply the classification concept according to "Language Model Kullback-Leibler divergence" description in [36] [37], use the following equations.

$$D(objective||subjective) = \sum_{w \in sentence} p(w|obj) log \frac{p(w|obj)}{p(w|sub)} \tag{6.2}$$

$$D(subjective||objective) = \sum_{w \in sentence} p(w|sub) log \frac{p(w|sub)}{p(w|obj)} \tag{6.3}$$

Appendix A

List of Domian of Weblogs.

- http://deepblog.com/
  - We extracted 14251 Weblogs.
- http://weblogs.about.com/
  - We extracted 1247 Weblogs.
- http://www.blogthings.com/
  - We extracted 9014 Weblogs.
- http://www.blogstakes.com/
  - We extracted 7785 Weblogs.
- http://neworleans.metblogs.com/
  - We extracted 15321 Weblogs.
- http://newslib.blogspot.com/
  - We extracted 7458 Weblogs.
- http://veganlunchbox.blogspot.com/
  - We extracted 1349 Weblogs.
- http://loveandcooking.blogspot.com/
  - We extracted 14261 Weblogs.
- http://www.vidblogs.com/
  - We extracted 9645 Weblogs.
- http://www.blogarama.com/

  - We extracted 1136 Weblogs.
- http://london.metblogs.com/
  - We extracted 8801 Weblogs.
- http://mydhaba.blogspot.com/
  - We extracted 15641 Weblogs.
- http://www.bloggercon.org/
  - We extracted 11290 Weblogs.
- http://libswithclass.blogspot.com/
  - We extracted 1434 Weblogs.
- http://www.globeofblogs.com/
  - We extracted 451 Weblogs.
- http://www.bloggersblog.com/
  - We extracted 20186 Weblogs.
- http://www.blogsontop.com/
  - We extracted 11984 Weblogs.
- http://www.topbloglists.com/
  - We extracted 4251 Weblogs.
- http://blog.itopsites.com/
  - We extracted 9642 Weblogs.
- http://www.top100bloggers.com/
  - We extracted 5218 Weblogs.
- http://bogbumper.blogspot.com
  - We extracted 5674 Weblogs.
- http://duanekeiser.blogspot.com/
  - We extracted 7254 Weblogs.
- http://www.blogcatalog.com/
  - We extracted 25154 Weblogs.

Appendix B

Glossary

annotation The process of annotating specific linguistic features, relationships, or structures in a text (usually in a corpus).

corpus A body of linguistic data, usually naturally occurring data in machine readable form, especially one that has been gathered according to some principled sampling method.

corpus linguistics A computer-assisted methodology that addresses a range of questions in linguistics by empirical analysis of naturally occurring speech and writing.

definition The explanation of the meaning of a term. Traditionally, definitions were assumed to state necessary and sufficient conditions for the correct use of a word, but modern lexicographers, following philosophers such as Wittgenstein and Putnam, object that a definition cannot set boundaries of this kind. For this reason, some lexicographers prefer to talk about the sense of a word rather than is definition. The term explanation is sometimes preferred to definition to describe what is actually said about the term.

dictionary A collection of words and phrases with information about them. Tradicional dictionaries contain explanations of spelling, pronunciation, inflection, word class (part of speech), word origins, word meaning, and word use. However, they do not provide much information about the relationships between meaning and use. A dictionaty for computational purpose (often called a lexicon) rarely says anything about word origin, and may say nothing about meaning or pronunciation either.

distribution The variety of different texts in a language or a corpus in which a particular word or phrase is used. Some terms tend to cluster in particular domains or text types.

entropy The degree of disorder or randomness in a system, often taken as a measure of how difficult it is to predict the outcome of a random variable.

feature (i)In lexical semantics, a formal property of a word or phrase that marks it as similar to one set of words and phrases on a particular dimension and distinguishes it from other sets. A feature is usually indicated by naming the dimension (e.g. number or gender) and specifying the value (e.g. singular or plural); (ii) in phonetics, a particular aspect of the articulation of a speech sound; (iii) the term feature is used in a number of other areas of linguistics or computational linguistics to denote a property or attribute.

generation (i) Automatic production of natural language texts by machine on the basis of some specified semantic, communicative, or syntactic input; (ii) in formal languages, the process of producing the strings that express a given set of meanings or grammatical relations.

grammar (i) The whole system and structure of a language (or of languages in general), in particular syntax and morphology, or a systematic analysis of the structure of a particular language; (ii) in the theory of formal grammars, a generating device consisting of a finite nonterminal alphabet, a finite terminal alphabet, an axiom, and a finite set of productions.

hyponym A word that has a more restricted meaning than another word with which it is in a hyponymy relation. For example, śparrowánd ćanaryáre hyponyms of b́ird.́

information retrieval The science of finding objects in any media relevant to user′s possible queries. For example, information retrieval over text takes a text query and retrieves documents relevant to that query; information retrieval over images might take a query in text or speech and retrieve images, or might take an image as query and retrieve related image.

inverse sentence frequency (isf) A measurement of the occurrence of a word within a collection of sentences in inverse relation to the number of sentences in the collection.

language (i) The system of communication used by human beings in general or a particular communicative system used by a particular community. A language may be natural (e.g. English or Bulgarian) or formal (e.g. computer programming language or a logical system); (ii) in the theory of formal grammars and languages, any subset of the infinite set of strings over an alphabet.

language model In statistical Natural Language Processing, a model used to estimate the probability of word sequences.

machine learning The use of computing systems to improve the performance of a procedure or system automatically in the light of experience.

morphology The internal structures and forms of words, or the branch of linguistics that studies these.

n-gram A sequence of n tokens.

part-of-speech Any of the basic grammatical classes of words, such as noun, verb, adjective, and preposition.

part-of-speech tagger A computer program for assigning labels for grammatical classes of words.

part-of-speech tagging Assigning labels for grammatical classes of words through a computer program.

phrase A sequence of words that can be processed as a single unit ia a text.

probability distribuition A distribuition that determines the mathematical properties of a random variable; it is the cumulative of the probability (density) function.

recall The number of correct responses divided by the total number of possibly correct responses.

recognition In the theory of formal grammars and languages, the process of determining whether a particular string belongs to the language (of a given grammar) accepted by a given automation.

relevance An important principle in pragmatics, according to which a header interprets a speaker's utterance, in part on the basis of its contextual impact.

segment (i) (Verb) the act of splitting up a dialogue into utterance units; (ii) (noun) any of the subunits into which a text may be divided; (iii) (noun) a unit of sound in phonetics; (iv) (noun) an alternative term for utterance unit, best avoided as it can easily be confused with (ii) or (iii).

semantics The study of linguistic meaning.

speech recognition Transcription of the speech signal into a sequence of words.

spoken corpus A corpus that seeks to represent naturally occurring spoken language. While this could in principle be simply a collection of tape recordings, it is much more common to find that such material has been orthographically transcribed. It may also be that the material has been phonemically transcribed either in addition to, or instead of, an orthographic transcription, sometimes with suprasegmental markings.

string Any sequence of letters from an alphabet, including numerals, punctuation marks, and spaces.

synonym A lexical item that has the same meaning as another lexical item.

tagging Assignment of tags to words or expressions in a text.

term A lexical unit, typically one validated for entry in an application-oriented terminological resource describing the vocabulary of a specialized subject field.

term frequency (tf) A measurement of the frequency of a word or term. Term frequency reflects how well that term describes the text contents.

text categorization The process of making decisions about whether a document is a member of a given class or category, e.g. in news, sports vs. finance, or in literature, poetry vs. prose.

user A human agent involved in some form of communication or interaction with a computer system.

utterance A unit of spoken text, typically loosely defined and used. On the structural level utterances may correspond to phrases or sentences uttred by a speaker, whereas on the functional level they may correspond to dialogue acts.

WordNet A database based on the psycholinguistic theories of George Miller at Princeton University, consistin of a semantic network relating synsets to one another, where synsets are sets of synonyms in a language.

Bibliography

[1] James Allen. *Natural Language Understanding*. The Benjamin/Cummings Publishing Company, Redwood City, 2 edition, 1998.

[2] Stuart Russell and Peter Norving. *Inteligência Artificial*. Editora Campus, Rio de Janeiro, 2 edition, 2004.

[3] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*, pages 271–278, 2004.

[4] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill International Book Company, Tokyo, 1 edition, 1983.

[5] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *An Introduction to Information Retrieval*. Preliminary draft (c), Cambridge University Press, Cambridge, 2006.

[6] Ruslan Mitkov. *The Oxford Hanbook of Computational Linguistics*. Oxford University Press, New York, 1 edition, 2003.

[7] D. Harman. How effective is suffixing. *Journal of the American Society for Information Science*, pages 7–15, 1991.

[8] R. Krovetz. Viewing morphology as an inference process. *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 191–202, 1993.

[9] W. Frakes and R. Baeza-Yates. *Information Retrieval: Data Structures and Algoritms*. Englewood Cliffs, NJ: Prentice Hall, 1 edition, 1992.

[10] D. A. Hull. Stemming algorithms: a case study for detailed evaluation. *Journal of the American Society for Information Science*, pages 70–84.

[11] M. F. Porter. An algorithm for suffix stripping. *Program*, pages 7–130, 1980.

[12] J. B. Lovins. Development of a stemming algorithm. *Translation and Computational Linguistics*, pages 22–31, 1968.

[13] Ian H. Witten and Eibe Frank. *Data Mining*. Morgan Kaufmann Publisher, San Francisco, 2 edition, 2005.

[14] J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin. Learning subjective language. *In Computational Linguistics, 30(3)*, pages 277–308, 2004.

[15] P. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002.

[16] C. Whitelaw, N. Garg, and S. Argamon. Using appraisal groups for sentiment analysis. *In Proceedings of CIKM, Bremen,Germany*, pages 625–631, 2005.

[17] S. Bethard, H. Yu, A. Thornton, V. Hatzivassiloglou, and D. Jurafsky. Automatic extraction of opinion propositions and their holders. *In Working Notes of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, 2004.

[18] J. Wiebe, M. Bruce, and P. O'Hara. Development and use of a gold standard data set for subjectivity classifications. *In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 1999.

[19] A. Esuli and F. Sebastiani. Determining the semantic orientation of terms through gloss analysis. *In Proceedings of the ACM SIGIR Conference on Information and Knowledge Management, Bremen, Germany*, 2005.

[20] S. Kim and E. Hovy. Determining the sentiment of opinions. *In Proceedings of the 20th International Conference on Computational Linguistics*, 2004.

[21] G. A. Miller. Wordnet: A lexical database. *Communications of the ACM 38*, 1995.

[22] P. Chesley, B. Vincent, L. Xu, and R. Srihari. Using verbs and adjectives to automatically classify blog sentiment. *In Proceedings of AAAI Spring Symposium*, 2006.

[23] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. *In Proceedings of the 2002 conference on empirical methods in natural language processing*, 2002.

[24] P. D. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. *In Proceedings of*

the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02), pages 417–424, 2002.

[25] M. Hu and B. Liu. Mining opinion features in customer reviews. *In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2004.

[26] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack. Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques. *In Proceedings of the 3rd IEEE International Conference on Data Mining*, 2003.

[27] M. Hurst and K. Nigam. Retrieving topical sentiments from online document collections. *In proceedings of the 11th conference on document recognition and retrieval*, 2004.

[28] Robert Dale, Herman Moisl, and Harold Somers. *Handbook of Natural Language Processing*. Marcel Dekker, New York, 1 edition, 2000.

[29] Gouri K. Bhattacharyya and Richard A. Johnson. *Statistical Concepts and Methods*. John Wiley & Sons, New York, 1 edition, 1977.

[30] Philip Clarkson and Ronald Rosenfeld. Statistical language modeling using the cmu-cambridge toolkit. *Proceedings ESCA Eurospeech*, 1997.

[31] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees.

[32] Helmut Schmid. Improvements in part-of-speech tagging with an application to german.

[33] Trevor Hastie, Robert Tibshirami, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics, New York, 1 edition, 2001.

[34] Vladimir Cherkass Ky and Filip Mulier. *Learning From Data*. John Wiley & Sons, New York, 1 edition, 1998.

[35] Thorsten Joachims. *Learning to Classify Text using Support Vector Machines, Methods, Theory, and Algorithms*. Kluwer Academic Publishers / Springer, Cornell City, 1 edition, 2002.

[36] John Lafferty and Chengxiang Zhai. Modelbased feedback in the language modeling approach to information retrieval.

[37] John Lafferty and Chengxiang Zhai. Document language models, query models, and risk minimization for information retrieval.