John Kaustinen

Sentiment analysis of Finnish movie reviews

—extracting sentiment from texts in a morphologically rich language

Master's Thesis in Information Systems
Supervisor: Markku Heikkilä
Faculty of Social Sciences, Business and Economics
Åbo Akademi University

Åbo 2018

ABSTRACT

Subject: Information Systems

Writer: John Kaustinen

Title: Sentiment analysis of Finnish movie reviews - extracting sentiment from texts in a morphologically rich language

Supervisor: Markku Heikkilä	**Supervisor:**

Abstract:

Sentiment analysis is by now a well-researched subject within the field of Natural Language Processing. Due to technological advancements in computer strength and an explosion of freely available textual data in a semi-structured format online, the costs for conducting research on sentiment analysis have drastically decreased. A majority of the research has been conducted on data in English, and hence the best practices within the field are being formed around data in English. That also includes resources like programming libraries, sentiment lexicons and annotated corpuses for model training. Hence, it is important to fully understand if the same practices can be applied to all languages or if there are fundamental differences between languages regarding how to find the most sentiment-loaded features of a text.

The purpose of this thesis is to examine whether a morphologically complex language, such as Finnish, requires different preprocessing methods than a morphologically simple language, such as English, in order to find the most sentiment-loaded features to be used in a sentiment analysis. Finnish has several properties that make it different than English. The number of unique words, the lack of adpositions and the much more frequent use of case endings are all features of the Finnish language that may have an impact on the result of a sentiment analysis system.

To answer this question, two datasets (one in Finnish and the other in English) were used to create an experiment where it could be observed if the two datasets respond differently to different preprocessing techniques. The most common preprocessing techniques used in prior research were used to create five different preprocessing settings. The two datasets were processed five times each with the different preprocessing settings chosen and a Naïve Bayes Classifier was trained and tested on each of the five versions of both datasets.

The results show that the only notable difference between the Finnish and the English dataset in terms of classification results, was that using only lemmatized adjectives as features are more fruitful when dealing with textual data in English. This can be a result

<table>
<tr><td colspan="2">of the fact that adjectives are used more extensively in English to describe sentiment than in Finnish. Otherwise, the experiments performed for this thesis did not indicate that data in Finnish would require different preprocessing techniques than data in English in order to perform a successful sentiment analysis.</td></tr>
<tr><td colspan="2">Keywords:

Sentiment analysis, opinion mining, natural language processing, NLP, morphology, morphological complexity, morphologically complex language</td></tr>
<tr><td>Date: 31.10.2018</td><td>Number of pages: 70</td></tr>
</table>

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

INTRODUCTION

1.1 Background

Gaining insights into the aggregated opinion of a large group of people in a fast and affordable way has long been merely a dream for many companies, political parties and governments, since their success is dependent on what consumers think about their brand, products or services. Before the age of the internet, trying to understand what a large group of people think about something was a task that required a great deal of time and money, since data had to be gathered manually and PCs were generally not powerful enough to handle large datasets. In the age of the internet, the amount of freely available textual data in electronic format has exploded through the rise of social media platforms, which serve as mediums for storing people's opinions. News articles, product reviews, online discussions and the ability for people to express an opinion through a comment or a "like", all provide companies, political parties and governments with a possibility to gather the opinions of people in a fast and inexpensive way.

During the recent years, the rise of cloud computing as well as advancements in CPU speed and the size of RAM for personal computers, have paved the way for machine learning. Through machine learning, we can instruct a computer to learn how to make decisions on its own given some input. The input in this case is data in the form of natural language, which is now available in large quantities on the internet. Research in the field of Natural Language Processing has benefited considerably from these technological advancements and several very useful application areas have emerged from the research. Sentiment analysis is one of the fields that has shown to be a useful application in many areas, which is the concept of classifying a text according to its opinion, usually as negative or positive.

Businesses of all sorts can benefit from sentiment analysis in one way or another. Analyzing what people's opinions are about a specific brand, a product or a political figure is perhaps the most common areas where sentiment analysis is used. Sentiment analysis can also be used internally, for example, by analyzing employer emails to gain insights on the level of negativity within the organization, which companies can use to take action to improve their own culture. Another possible application which Microsoft

and Google are working on to deploy in their search engines, is to identify sentiment related to links to other pages. Currently, search engines use, among other things, the "web" of links between pages to produce search results. Pages that have more links pointing to them are given a higher probability to be ranked higher in search results than pages that only have a small number of links pointing to it. The problem here that sentiment analysis could solve is that links on a webpage might not always be put on a page in a positive sense, which is something that search engines preferably should be taking into consideration. If I am searching the web for a cleaning service, I probably do not want to find cleaning services which have a large amount of negative content related to them. Thus, by conducting a sentiment analysis on the text related to the link itself, search engines can classify links as negative and positive, which could have a positive impact on the accuracy of search results. (Crowl, 2018)

However, a majority of the research regarding how to set up a sentiment analysis, including how to preprocess the textual data and what kinds of features to use in the sentiment classifier, has been conducted on data in English. While English is certainly one of the most important languages in the world and there are many languages specifically in the Germanic language family that share the same characteristics, there is also a very large number of languages that do not share these same characteristics and therefore pose additional challenges on sentiment analysis systems. Because of fundamental differences in the language's grammatical structure, one will need to process the data in a different way to achieve the best possible sentiment analysis results (Abdul-Mageed et al., 2011).

In a supervised sentiment analysis, a classification algorithm is trained with a large amount of example data. The data points are marked with the sentiment class to which they belong. The data fed into the classifier is in the form of a feature vector, which can be made up of, for example, the presence of certain words or the count of how many times certain words are occurring in the text. The classification algorithm then learns what constitutes a positive text and what constitutes a negative text based on the examples given in the training data. The data usually requires preprocessing in order to isolate the most sentiment-loaded aspects of the text, since many words used in a human-written text are not good indications of the direction of sentiment at all. The performance of the classification algorithm is greatly dependent on finding the correct features and therefore,

much research in this area has circled around finding the correct level of preprocessing to achieve as good results as possible. Since most of this research is done on data in the English language, I want to investigate if data in Finnish, which as a language has quite many grammatical differences compared to English, requires a different set of preprocessing steps to find the most sentiment-loaded features.

1.2 Objective of the thesis

This thesis studies the impact of language when creating a sentiment analysis system. The objective of this thesis is to provide a comprehensive study by comparing how the results of a sentiment classifier differ for different levels of preprocessing. I will use one dataset, which consists of Finnish movie reviews and another dataset that consists of English movie reviews. The experiments will be run separately on both datasets and the performance of the sentiment classifiers will be measured to understand if different feature setups work better for different languages.

The experiments are set up similarly to Abdul-Mageed et al. (2011), except that I will use two datasets in different languages and compare the results. I will have five different levels of preprocessing setups and will record the performance of the classifiers for each setup and for each data set, which will allow me to observe if the language of the data has an impact on the performance of the sentiment classifier.

Finnish is a morphologically complex language while English is a morphologically simple language. Morphology is about to what extent the words in a language are inflected when producing sentences. For example, Finnish has almost no prepositions at all and morphemes are therefore used instead to produce the same grammatical functionality that prepositions do in English. In English though, prepositions are used to a large extent when producing sentences. Now, in a sentiment analysis system created for English data, prepositions are usually removed from the data since they provide very little information about the sentiment of a text. If we want to remove the same aspects from the data in a Finnish text, we cannot simply have a list of words that we remove, because the syntactical function of prepositions is handled by morphemes which are merged to other words, such as nouns, verbs and adjectives, that may be the sentiment-loaded ones and that we therefore want to include in our model. Thus, if we want to remove the syntactical

function of prepositions in Finnish, we need to process the data differently. This thesis will investigate if this is relevant or not to consider in a sentiment analysis system.

1.3 Structure of the thesis

The second chapter of this thesis consists of the theoretical framework. The chapter starts with briefly presenting artificial intelligence, which is the parent field from which the first ideas relating to machine learning and computational processing of language emerged. Natural language processing is then presented that can be considered a subfield of artificial intelligence and is the field out of which sentiment analysis emerged. Here I also go through the computational language processing methods which will be used to conduct the preprocessing of the data in my experiment. This is followed by a thorough presentation of what sentiment analysis is, how it can be performed and where it can be used in our society. I continue by presenting prior research from which I have taken much inspiration and many ideas for this thesis. The key focus point here is the research about how language can impact a sentiment analysis and specifically what kinds of challenges the morphology of a language poses when conducting a sentiment analysis. I also present the linguistic differences between Finnish and English, which serve as a foundation for the objective of this thesis. Lastly, in section 2.3 I form the hypothesis for my research.

The third chapter is focused on presenting the model that I will use to conduct the experiment and the data that I use in my experiment. The third chapter explains in detail why and how the methods are used to produce the results I need to accept or decline my hypothesis. In this chapter, I also present descriptive statistics about the data and already here we can see some evidence regarding the differences between the two languages.

The fourth chapter presents in detail the experiment setup, how the results are validated and the actual results of the experiment together with a discussion about the possible reasons for the result. Chapter 5 provides a summary and a conclusion of the thesis.

2 THEORETICAL FRAMEWORK

In this chapter I will lay out the theoretical framework for this thesis. The theoretical framework is focused on presenting the theoretical material to support my hypothesis. This chapter will include a presentation of the fields of artificial intelligence, natural language processing, sentiment analysis as well as how differences between languages can have an impact on the result of a sentiment analysis.

2.1 Mining opinions from natural language

I will start this chapter by presenting artificial intelligence, which is the research field that first started to deal with the problem of how computers can understand natural language spoken by humans. I will also present natural language processing, which is the parent-field of sentiment analysis.

2.1.1 Artificial intelligence

2.1.1.1 Definition

Creating something equal to the human mind has been discussed by philosophers and scientists for a long period of time. According to Poole and Mackworth (2010), artificial intelligence (AI) is the field where "the synthesis and analysis of computational agents that act intelligently," are studied. The Association for the Advancement of Artificial Intelligence defines the field of AI as "the scientific understanding of the mechanisms underlying thought and intelligent behavior and their embodiment in machines." Wikipedia defines artificial intelligence as "intelligence exhibited by machines."

The basic idea of AI is that we have an agent that refers to something that exists and acts intelligently in an environment. An agent is acting intelligently when it does things that are appropriate for its goals and it can adjust flexibly to a changing environment. An intelligent agent is also able to learn from its experience and the choices it makes are appropriate to its "perceptual and computational limitations." (Poole & Mackworth, 2010) The central scientific goal in the field is to understand what it is that makes intelligent behavior possible in agents, artificial or natural. The central engineering goal

of artificial intelligence is to be able to create agents that act intelligently, which would be a benefit in many applications.

2.1.1.2 Past and present

Scientists and philosophers have for 400 years discussed what thought and reason really is. The work on artificial intelligence as we think of it today began in the 1940s and 1950s with a handful of scientists from different fields starting to explore the possibility of creating a machine that is able to reason. In 1950, a now quite famous man named Alan Turing published a paper about computing machinery (computing was at that time only performed by humans) and intelligence where he introduced a "game", which later came to be known as the "Turing test". (Poole and Mackworth, 2010) The Turing test is simply a textual communication between a human being and a machine. If the human being cannot tell based on the answers that the machine gives, if it really is a machine or a human being, then the machine has passed the Turing test. (Turing, 1950)

When actual real computers as we know them today were built, artificial intelligence applications were among the first to be created for these machines. Already in 1952, a man named Arthur Samuel created a computer program that could play checkers against human beings. Discoveries in how the human brain works laid the groundwork for neural networks applications and there were lots of other attempts in creating intelligent applications. The most apparent problem was how to express the knowledge needed to solve any given problem. (Poole & Mackworth, 2010)

One of the goals of AI has long been to create a machine or application that is able to understand natural language. At first, language understanding and translation was thought to be a trivial problem because of the incredible capacity and power that computers possess to store large dictionaries and retrieve words easily. Translation applications that worked by simply looking up words in a table failed heavily and after that there were not much research done in the field of language understanding and translation for a long period of time. (Buchanan, 2005) There was occasional success in the area though, for example a program called STUDENT, created by D. Bobrow in 1964, was able to solve algebra word problems given to the computer in natural language but overall, understanding natural language proved to be quite a difficult problem. (Poole and Mackworth, 2010; McCorduck, 2004)

Since the first attempts of creating artificial intelligence in the 1950s and 1960s, we have come a long way. The first annual research conference devoted to artificial intelligence was established in 1965 and in 1980 the Association for the Advancement of Artificial Intelligence was founded to provide "annual conferences for the North American AI community".(Buchanan, 2005) One milestone in the research and development of artificial intelligence was achieved in the famous chess game between the then world champion Garry Kasparov and the IBM computer named Deep Blue. Garry Kasparov had won the first game in 1996 but eventually lost the second game in 1997, where he also accused IBM of cheating by using human beings to make some moves and not the computer. The victory of Deep Blue was seen as a landmark in artificial intelligence because a machine beat one of the greatest human minds ever. (Krauthammer, 1997)

Today artificial intelligence is more actual than it has ever been. Only in 2015, companies around the world invested about $8.5 billion in artificial intelligence. Today's most ambitious projects include self-driving cars and virtual personal assistants that can understand not only language in textual format but also speech. The industry has come so far that a concern over an "intellectual monopoly" has been raised, meaning that one company is able to create such an advanced technology that rapidly creates industry barriers that will be impossible to overcome for other companies. For this reason, OpenAI was founded by Elon Musk and other tech leaders in 2015 to ensure that the development on AI is not going in a direction that will be potentially harmful in the future. (The Economist, 2016)

2.1.1.3 Machine learning

As the researchers within artificial intelligence preferred to focus on knowledge in intelligence and not on learning-related issues, a new field called *machine learning* appeared in the 1980s. (Langley, 2011) This was also the time period called the "AI winter", when the research done in the AI field did not have any substantial breakthroughs. (McCorduck, 2004) Machine learning is a field that lies in the intersection between computer science and statistics. (Jordan & Mitchell, 2015) The area of machine learning blends with an area called statistical learning and there are only minor differences between the two but the central themes in both areas is to derive real world knowledge from data. This can be seen by reading through the table of contents of the books "Introduction to Statistical Learning" and "An Introduction to Machine Learning", where

the contents are partially overlapping. (James et al., 2013; Smola and Vishwanathan, 2008)

When describing machine learning, Smola and Viswanathan (2008) uses a couple of examples of applications where learning is necessary in order to find the answer to a specific problem. Automatic translation between two languages, named entity recognition and speech recognition is a couple of applications that are all machine learning problems and also part of natural language processing, which will be presented more in depth later in this thesis. Some of the problems in machine learning that Smola and Viswanathan (2008) mentions are binary classification, multiclass classification, structured estimation, regression and novelty detection. All of these areas have different algorithms that work better or worse depending on the circumstances (the data).

Machine learning is today used not only within research but has also found its way into corporations and the public sector. Some of the areas that machine learning techniques are used in today are marketing, financial modeling, policing, education, manufacturing and health care. The development in the field has been driven by the development in the algorithms used, the explosion of available data and the development of more and more powerful computers. (Jordan and Mitchell, 2015)

Function approximation is one of the main problems studied in machine learning and goes under the area of *supervised* learning, which basically means that we have a set of predictor measurements (observations) that all of them have an associated response measurement and then we try to fit the model so that the predictor measurements relates to the response. The goal is to create a model so good that we can to a certain accuracy predict the response of future observations or learn about the relationship between the predictors and the response. (James et al., 2013)

Unsupervised learning focuses on the situation where we have a set of measured observations but no related response. It is not possible to fit a function to these observations that would produce an accurate response. The reason for that is that there is no response model we can use to supervise our learning phase. For example, if we do not know anything about the outcomes of a set of different experiments that we have done, we cannot say anything about the outcomes of future experiments either. What we can do though, is to find relationships among these observations. One example of an

unsupervised learning technique is cluster analysis, which can determine the degree of similarity between these different observations that we have. (James et al., 2013)

There are also problems called *semi-supervised* learning problems. This is a situation where a part of the observations does have a response variable while the other part of the observations does not have a response variable. These kinds of situations occur in areas where it is very expensive to obtain labeled data and unlabeled data can be obtained easily. (James et al., 2013) Sentiment analysis is one example where labeled training data is difficult to obtain and there has in fact been some research within the field where the researchers have used semi-supervised learning approaches when classifying textual data. (Yang et al., 2015; Goldberg and Zhu, 2006)

In this thesis, I will use supervised machine learning in order to accept or reject my hypothesis.

2.1.2 Natural language processing

I mentioned natural language processing in the earlier section about artificial intelligence and I will now present it more in depth as it is a central subject to the purpose of this thesis.

2.1.2.1 Definition

Natural language is evolved and used by human beings in both spoken and written form for communication purposes. Natural language processing is a field where the use of natural language is studied in order to find a way for machines to use, manipulate and understand natural language as a way to solve real world problems. In some circumstances, Natural language processing is also called computational linguistics. (Chowdhury, 2003; Kumar, 2011)

Natural language processing is a product of many different fields including computer and information sciences, linguistics, mathematics, electrical and electronic engineering, artificial intelligence, robotics and psychology. The applications of natural language include machine translation, natural language text processing and summarization, user interfaces, multilingual and cross-language information retrieval (CLIR), speech recognition, artificial intelligence and expert systems.

Natural language processing is considered a subfield of artificial intelligence and is indeed a very important area of research in the quest of creating a machine that is interchangeable with a human being. Understanding natural language as input and being able to generate output in natural language is a requirement for a machine to pass the Turing test, which was mentioned earlier in the section about artificial intelligence. Kumar (2011) confirms what the first pioneers within the field of NLP realized through their experiments; natural language does not seem to be too complex at first but the "deep level processing of natural languages, understanding of implicit messages and intentions of a speaker are extremely difficult avenues."

2.1.2.2 History

One of the first applications in NLP was machine translation. (Kumar, 2011) As defined by Chowdhury (2003), "machine translation (MT) refers to computerized systems responsible for the production of translation with or without human assistance." As mentioned earlier in the section about artificial intelligence, the first approaches in machine translation were very disappointing because the machine could only do lookup based on words and was not able to handle any sort of deeper meaning or ambiguity. A well-known example of these early failures in the field, mentioned by Kumar (2011), is about a group of researchers at Georgetown University who had created a system that was able to translate English into Russian and vice versa. When they fed the sentence "The spirit is willing but, the flesh is weak" into the system and translated it into Russian and back to English, the translation that the system then outputted was "The Vodka is good, but the meat is rotten." This demonstrates how ambiguous natural language can be and why the early attempts in the field failed. Today, MT systems use something called the "knowledge-based technique" to incorporate semantics into the translation attempt. The knowledge-based approach seeks to understand the text before translating it by using AI techniques, but a flawless translation system has yet to be created because of the incredible amount of knowledge that is needed in order to understand more complicated meanings. (Kumar, 2011)

One of the more famous works in the history of linguistics is *Syntactic Structures* written by Noam Chomsky in 1957. (Liddy, 2001) The book introduced the idea of generative grammar, meaning that there are a set of logical rules that define the combination of words that produce a grammatical sentence. This idea gave the researchers in the field of

machine translation a better view on how linguistics can help them in the creation of a fully working machine translation system.

During the late 1950s and the early 1960s, other areas as, for example, speech recognition emerged. This time period was also influenced heavily by the developments in syntactic theory and language parsing algorithms. This research breakthrough led to an over-enthusiasm in the field and researchers believed that fully automated translation systems that could produce similar results to those that a human translator produces, were going to be developed within a few years. Despite the enthusiasm, limitations in linguistic knowledge and computer systems made an automatic translation system not possible at this point in time. The enthusiasm in the field died in 1966 when ALPAC (Automatic Language Processing Advisory Committee of the National Academy of Science – National Research Council) released a report that suggested that machine translation should not be funded because it is not achievable at the moment. (Liddy, 2001)

Although the interest had declined, research within the field continued in the 1970s with a shift more towards semantics and considerable work was seen on natural language generation. In the late 1980s the statistical approaches got accepted more and more within the research field and were seen as an important compliment to the symbolic approaches that had dominated the field since the 1950s. The field of natural language processing grew very rapidly in the 1990s mostly due to an increase in computer power and an explosion of available text in electronic form. Different tasks of NLP like syntactic parsing and part-of-speech tagging began to incorporate probabilities in their solutions simply because it worked well. The rise of the interactive web in the last 10-15 years has also made the need for language-based information retrieval and information extraction from natural language an important research area within the field of NLP. (Kumar, 2011; Liddy, 2001)

2.1.2.3 NLP tasks

I will use several NLP tasks for the preprocessing of the data in my analysis, so I will now present the most used NLP tasks.

Tokenization

Tokenization is the process of extracting words, also called tokens, from a natural language text. Although this sounds like a trivial task, important to remember is that there

are several language-specific features that might cause problems. The first one is punctuation marks, which naturally do not belong to words but are not separated from words by whitespace. Parentheses, percentage-signs and apostrophes are other types of written language aspects that might cause problems. However, today's tokenizers are very good at handling these problems, if the words are spelled correctly. Tokenization is one of the first preprocessing steps when analyzing natural language. (Bird et al., 2009)

Lemmatization

Lemmatization is the process of extracting a word's *Lemma*, which is the form of a word that is most usually found in dictionaries. Lemmatization is used when there is a need to map a word to an existing resource, for example a lexicon, in order to establish a connection. In languages that are so called *morphologically complex,* meaning that the words are inflected to a higher degree than in other languages, lemmatization is particularly important in order to be able to group words or map them to lexical resources. Lemmatization is also important for other NLP tasks, such as parsing and part-of-speech tagging. (Müller et al., 2015)

Part-of-speech (POS) tagging

Part-of-speech are grammatical descriptors of words in text. The existence of part-of-speech has been recognized for several thousands of years and every language seem to have part of speech (Voutilainen, 2003). Verb, adjective, noun, adverb and so on are some of the different part-of-speech classes that different words in a sentence belong to. A standalone word can belong to several different part-of-speech classes but the context in which the word occur in a sentence decides what type of part-of-speech class the word belong to. The activity of POS-tagging refers to the automatic assignment of part-of-speech tags to words in a text. POS-tagging is used as a preprocessing step in many different NLP applications, including sentiment analysis, since including underlying information about the text has shown to be beneficiary in many types of analyses. (Voutilainen, 2003)

Stemming

Stemming is the automatic process of reducing a word to its stem or root. Contrary to lemmatization, stemming does not always produce a viable word as output, but rather the part of a word, after/before morphological units, such as morphemes, are placed in order to introduce grammatical structure. Stemming algorithms are common in web search

engines, since one want to be able to associate, for example, the word “fish” with “fishing”, which is an example of something that a lemmatization algorithm would not do since the word “fish” is a noun and the word “fishing” is a verb and therefore different words from a semantical point of view. (Bird et al., 2009)

Chunking and chinking

Chunking is the process of grouping a set of tokens together to form individual phrases that do not overlap. An important precondition when we are chunking is that the POS-tags are known, since it is by the structure of the POS-tags that the automated chunking is performed, for example by grouping noun-phrases together or by creating bigrams that consists of an adjective followed by a noun. Chinking on the other hand is the process of removing words from a chunk. Chunking and chinking could be used in a sentiment analysis to find, for example, nouns preceded by adjectives (e.g. “good movie”). (Bird et al., 2009)

2.1.2.4 Applications

According to Chowdhury (2003), making computers understand natural language involves three core problems. The first one is about thought processes, the second one is about the representation and meaning of the linguistic input and the third one is about world knowledge. Liddy (2001) and Feldman (1999) presents seven different levels of natural language processing, also called the synchronic model of language. These levels are phonology, morphology, lexical, syntactic, semantic, discourse and pragmatic.

The phonology level relates to the interpretation of sound level within a word and across many words. The morphology level relates to the smallest units of meaning within words, called morphemes (e.g. the word *preregistration* has a prefix (pre), a root (registra) and a suffix (tion)). Since morphemes have the same meaning across words, computers can use morphemes in order to gain an understanding of the meaning. The lexical level relates to the meaning of individual words. The syntactic level relates to the words within a sentence that form a grammatical structure. The semantic level relates to the possible meanings of a sentence or a word. The discourse level focuses on the properties of a text rather than single sentences. Lastly, the pragmatic level focuses on “extra meaning” beyond the contents of the text. (Liddy, 2001; Feldman, 1999)

These levels can be divided into lower and higher levels of natural language processing. A natural language processing application may incorporate one or many of these levels to perform an analysis. Liddy (2001) writes that current NLP systems do mainly use the lower levels because first of all, the application may not require a high-level interpretation. Secondly, more research has been conducted on the lower levels. Thirdly, the lower levels deal with sentences, words and morphemes that are small units of analysis. The higher levels deal with world knowledge, something that is expensive in terms of both time and money to incorporate into an NLP system.

There are many different applications that uses NLP in order to acquire real-world knowledge about a text or sentence. I will now present the most widely used NLP applications presented by Liddy (2001).

Information Retrieval (IR)

IR is concerned with storing, searching and retrieving information. NLP methods can be used for IR on data in plain text.

Information Extraction (IE)

Recognizing, tagging and extracting key elements from a plain text is defined as IE. Categorizing words of a large text into certain key elements of information, for example persons, companies, animals and so on can be used for question-answering or data mining.

Question-Answering

This application provides answers to a user's queries, unlike IR that provides possible answers.

Summarization

The higher levels of NLP can by understanding the semantics in the text, summarize the text into a summarization of the original text.

Machine Translation

Automatic translation of a language into another.

Dialogue Systems

A complete system able to make a conversation with human beings. This is the ultimate goal of NLP.

2.1.3 Sentiment analysis

In this chapter I will present what sentiment analysis is, different approaches to sentiment analysis and what it can be used to.

2.1.3.1 Definition

Sentiment analysis is quite a new field within NLP research and did not really exist (with a few exceptions) until after the year 2000. It touches upon almost every aspect of NLP and deals with the same issues as presented in the previous sections, i.e. grammatical relations and the ambiguity in words and sentences, but sentiment analysis does not require full understanding of natural language (Liu, 2012). A sentiment analysis system only needs to grasp the positivity or negativity in a sentence or document in order to be successful. Sentiment analysis is therefore a very popular subject within NLP at the moment since it has the capability to make important contributions to real-world problems.

Liu (2012) defines sentiment analysis as "the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes." Pang and Lee (2008) defines sentiment analysis as "the computational treatment of opinion, sentiment, and subjectivity in text." Sentiment analysis is about analyzing huge amounts of textual data in natural language with machine learning techniques in order to find what a large group of people think about something.

Pang and Lee (2008) credits the rise of machine learning methods in natural language processing to the huge increase in textual data in natural language available for free on the internet. Sentiment analysis has many interesting applications that are interesting not only for academia but also for businesses. The opinions of large masses of people are a good indication of, for example, consumer trends and politics and therefore sentiment analysis has grown to become one of the most important tools within social media

research areas. The importance of people's opinions for different areas of our society creates a strong motivation for research within the area.

2.1.3.2 Approaches

Liu (2012) presents different levels on which sentiment analysis is done. These levels are *document level, sentence level,* and *entity and aspect level.* Conducting a sentiment analysis on the document level means trying to classify whether a single document expresses a positive or negative sentiment. (Pang et al., 2002; Turney, 2002) This approach assumes that the document expresses an opinion on a single entity, which means that documents containing an evaluation of multiple entities or a comparison between multiple entities may not be applicable to be analyzed with this technique. A well-studied topic where document level sentiment analysis is used is different kinds of reviews that have a single entity (product or service) that the whole document is focused on. When classifying a document as either positive or negative we also assume that the document contains subjectivity and is not entirely objective. Since a document contains several sentences, it is also a known fact that the document might contain expressions of both positive and negative sentiment. Classifying a document as either positive or negative is therefore only a summarization of the different opinions expressed in the document and it does not take context into account.

Sentence level sentiment analysis is about analyzing a single sentence and classifying it as positive, negative or neutral, where neutral means that the sentence expresses no opinion (Liu, 2012). Sentence level approaches do not differ that much from document level approaches since sentences are basically short documents. Researches often assume that a sentence expresses a single opinion while a document often expresses multiple opinions. This is of not really the case as shown from this sentence: "The car is very good, but it is way too expensive."

Analysis on both the document level and the sentence level do not show any context to the sentiments expressed in the document or sentence, only if the document or sentence is overwhelmingly positive or negative. The entity- and aspect-level approach is finer-grained and can tell us more about what entity in the data that the opinion refers to (Liu, 2012). This is based on the idea that an opinion consists of a sentiment (positive or negative) and a target (for example, a car or the breaks of the car). The goal of this type of analysis is to extract a sentiment—aspect pair from which we can receive a structured

summary of opinions about entities and their aspects. This type of sentiment analysis approach is very usable when an entity contains many different aspects. For example, in reviews about a certain type of car, it is possible to identify specific problem areas of the car instead of only an aggregated opinion if the car itself is overwhelmingly good or bad. Pang and Lee (2008) defines this as a joint topic—sentiment analysis.

The most important property of a text that tells the most about sentiment is so called *opinion words* or *sentiment words* and certain phrases that are used to express a positive attitude or a negative attitude. (Liu, 2012) Although sentiment words and phrases represent a large part of the sentiment in a sentence or text, it is not sufficient for an accurate sentiment analysis to focus only on these words and not care about the context in which the word occurs since many words can have either a positive or a negative meaning in a certain context. Sometimes a sentiment word, like the word "good", does not express any opinion at all, as shown by this example sentence: "If I find a good car, I will buy it."

Sentiment analysis is often considered a classification problem, meaning that we use binary classification, multi-class classification or regression in order to find the knowledge we want. A substantial amount of the research done in sentiment analysis has been made on reviews about products or services. Movie reviews is a well-studied domain, mostly since the amount of available data is so large. Reviews are a good research domain within the field because reviews, more often than not, express a subjective opinion on a single entity (Pang and Lee, 2008). However, certain review domains, like movie reviews, can be somewhat more difficult than others to classify (Turney, 2002). Most of the work done on sentiment analysis assume that the input data always contains a subjective opinion, but for many applications it might be necessary to decide if the document/sentence in question does contain a subjective opinion or not. Mihalcea et al. (2007, cited in Pang and Lee, 2008) explains that "the problem of distinguishing subjective versus objective instances has often proved to be more difficult than polarity classification, so improvements in subjectivity classification promise to positively impact sentiment classification."

There are several ways we can conduct a sentiment analysis. Supervised machine learning methods are the ones that have produced the best result so far, but they need lots of labeled training data to produce good results and the algorithms need to be retrained if we want

to do an analysis on another domain. The reason for this is that people expresses sentiment in a different way when reviewing a movie, then when analyzing what kind of car one is planning on buying. Unsupervised methods have also been developed and the most known is perhaps the PMI-IR algorithm presented by Turney (2002), but they have not received as good results as supervised methods have and generally they have a higher degree of complexity.

2.1.3.3 Applications

As mentioned in the previous section, sentiment analysis has several different applications. Pang and Lee (2008) and Liu (2012) have both devoted a whole chapter each in their respective works to discussing the applications of sentiment analysis. While it is a very new research field, many application areas have already been discovered.

Reviews of products and services is one topic that has been the focus of much research in sentiment analysis. According to Pang and Lee (2008), summarizing user reviews based on the text produced by the reviewer instead of the rating that the user gives it, might solve the problem of reviewers misinterpreting the rating scale or reviewers having different opinions about what kind of quality that a certain rating represents. Liu (2013) also mentions that because of the vast amount of publicly available review material on the web, it is no longer necessary for organizations to conduct surveys and polls about certain areas because a huge number of opinions is already expressed on the web. The difficult part here is naturally the processing of the data that is usually located in different places on the web and in different format and languages. Sentiment analysis could play an important role in summarizing these publicly available reviews. Quickly becoming aware of a product's or service's potential problem areas could be very useful for businesses in understanding how to develop a certain product before its reputation is destroyed by a faulty feature.

Pang and Lee (2008) discusses sentiment analysis as an enabler for other types of technologies. Recommendation systems could use sentiment analysis to filter out products or services that have received many negative reviews. Advertising systems could use sentiment analysis to detect the content of a webpage in order to prohibit inappropriate or embarrassing placement of ads. Email systems could use sentiment analysis to detect inappropriate or overly heated content in an email and warn the user before he or she sends or opens an email.

Sentiment analysis can also be a very powerful tool for governments. Liu (2013) mentions the role that social media had for the political changes that happened in the Arab countries in 2011. Monitoring social opinion is something politicians should be aware of in order to know what the people expect of the government.

Liu (2013) mentions many application-oriented research papers that have been published on sentiment analysis and one of these is a paper by Bollen et al. (2011) where twitter moods where used to predict the stock market.

2.1.4 Sentiment analysis research

According to Liu (2012), the term sentiment analysis was probably coined for the first time in 2003 by Nasukawa and Yi in an article named "Sentiment Analysis: Capturing Favorability using Natural Language Processing". However, the research problem had been discussed before that. Two well-known papers, both released in 2002, were probably the first ones to consider the problem of sentiment classification or sentiment analysis (Turney, 2002; Pang et al., 2002).

In this section, I will present what has already been done in the field of sentiment analysis. I will also present approaches to how researchers have tried to tackle sentiment analysis for other languages than English, and what kinds of challenges exist when doing a sentiment analysis on languages that are morphologically complex.

2.1.4.1 Early work

Reviews and especially movie reviews have been used as data for many different research papers on sentiment analysis. The reason behind this is that reviews are readily available on the internet and they are often labeled with, for example, a thumbs-up or a thumbs-down rating that indicates sentiment, which means that supervised machine learning methods can be used without having to do any manual labeling.

The articles presented in this section are considered to be the first works on sentiment analysis (Liu, 2012) and they all use reviews as subject for their experiments.

Pang et al. (2002)

Pang et al. (2002) tackles the problem of sentiment analysis by using supervised machine learning techniques to show that they outperform human introspection in defining

features to use for classifying a review as positive or negative. The data they used were movie review data where every review was labeled with stars or a number that the reviewer thought the movie deserved as a rating. These stars or numbers were then used to create a sentiment class variable with either *positive, neutral* or *negative* as a value. For the analysis, the authors discarded the reviews considered neutral and only used the reviews that were classified as either positive or negative. They began with explaining the problem of sentiment classification by illustrating how accurate human introspection was in finding features that make a text positive or negative. With the help of two computer scientists they then proceeded to extract words, more specifically adjectives, from the positive and negative reviews. The computer scientists made one list each of positive adjectives and one list each of negative adjectives. Then, simply by counting how many times all of these extracted negative and positive words occur in a review, they then classified the review as positive if the number of positive words was more than the number of negative words, and vice versa. The accuracies they achieved with this approach were 58% and 64%. They then tried another approach by using statistics and some introspection to extract the words instead of only using introspection. This resulted in a different kind of word list and the result they achieved by using the same method as before was 69%.

The machine learning techniques they present are Naïve Bayes, Maximum Entropy and Support Vector Machines. They did multiple analyses with different kinds of features. The features they used were unigrams (one word), bigrams (2 words following each other), unigrams + bigrams, unigrams + part-of-speech-tag, adjectives, top 2633 unigrams and unigrams + position. They found that support vector machines have a slightly better accuracy than the two other machine learning techniques and that unigrams gave the best performance. One could assume that adding more information to the features would make for a better analysis, but for a bag-of-words (bag-of-words is described more in detail in chapter 3.1) analysis like this one, this was not the case. The results for the different techniques and features ranged from 77.3% to 82.9%, so the differences were not big.

The authors sum up their work by stating that the accuracy of their analysis was lower than the accuracy for topic-based categorization (classifying a document as belonging to a certain topic). The authors believe that the reason for this is that the final sentiment of a document is not necessarily represented by the aggregated sentiment within the

document. For example, tragic plots might make the author of a review use more words that are associated with negativity although the movie itself might be good. They suggest that to improve the accuracy of supervised sentiment analysis, it would be necessary to try and identify which sentences that are describing the aspect one is interested in.

Turney (2002)

Peter Turney (2002) takes another approach than Pang et al. (2002) to classifying different reviews from different domains as *recommended* or *not recommended*. Turney (2002) uses a fairly simple unsupervised algorithm called PMI-IR (Pointwise Mutual Information – Information Retrieval). This algorithm was first presented in an earlier paper by the same author (Turney, 2001) where he demonstrated its capability in finding synonyms to words by accessing a search engine. In that paper, he tested the algorithm's capability in finding the correct answer to a part of the TOEFL language test. This test was a synonym test where the user was given a problem word and four alternatives where only one of the alternatives could be considered a synonym to the problem word. PMI-IR scored a 74% accuracy on the test.

Using PMI-IR for sentiment analysis is somewhat different compared to trying to find synonyms for words. In this approach, he wants to find if phrases that includes an adjective are more closely associated with the word *excellent* or the word *poor*. The first step in the analysis that Turney (2002) did was tagging all the words with a Part-of-speech-tag (POS-tag). Then he proceeded to extract two-word phrases that included an adjective or adverb together with a word that provides context. Then he measures the *semantic orientation* of all of the extracted phrases by querying a search engine and measuring if the extracted phrases receive more hits together with the word excellent or with the word poor. If the phrase occurs more often together with the word excellent than the word poor, it is considered to have a positive semantic orientation and vice versa.

Since this is an unsupervised approach to sentiment analysis, the author was not dependent on what kinds of training data that were available. Because of this he tried the PMI-IR algorithm on reviews in different domains and he found the results to differ much between the domains. Reviews on automobiles received an 84.00% accuracy, reviews on banks received 80.00% accuracy, movie reviews received 65.83% accuracy and travel destination reviews received 70.53% accuracy. The average accuracy for all the reviews was 74.39%. As can be seen, movie reviews received the lowest score. The author's

theory why this happened is almost the same as Pang et al. (2002) presented in their article, i.e. that the final sentiment of a movie review is not necessarily the sum of its part. A movie is also a product that has quite intangible aspects, while automobiles, banks and travel destinations have aspects that are more tangible or quantifiable and therefore the final sentiment is perhaps more a reflection of its parts than compared to movies. For example, a movie can have great acting, but the plot can still be quite dull.

Nasukawa and Yi (2003)

Nasukawa and Yi (2003) picks up where Pang et al. (2002) and Turney (2002) left off. A problem that Pang et al. (2003) discovered was that sentiment classification is more difficult to do than topic-based classification. Nasukawa and Yi (2003) explains that sentiment analysis requires high intelligence and deep understanding of the textual context where both domain knowledge and linguistic knowledge is necessary, and the sentiment of a document can be debatable even for humans where the topic of a text is seldom debatable.

Nasukawa and Yi (2003) takes a somewhat more sophisticated approach in trying to classify the sentiment of a text. They try to find sentiments and the subject that a specific sentiment relates to. The authors suggest that by doing this, they can filter out sentiments in a text that are off topic. For example, for movie reviews, it would be possible to discard sentiments that relates to the different aspects of the movie and just consider sentiments that are directed toward the entity in question. The authors motivate this approach by presenting the following example sentence: “XXX beats YYY.” In this sentence, there are no adjectives, which Pang et al. (2002) and Turney (2002) focused on. The sentence only consists of a verb that transfer sentiment between two nouns (subject and object). The sentence expresses positive sentiment towards XXX and negative sentiment towards YYY. The algorithms used by Pang et al. (2002) and Turney (2002) would not factor in this type of sentiments since they are focused on the polarity of single words and phrases, and not the knowledge created by a syntactic structure.

In order to include the type of sentiment presented above, Nasukawa and Yi (2003) manually created a sentiment lexicon of 3513 entries consisting of both verbs, adjectives and nouns. The verbs were tagged with their part-of-speech-tag, a sentiment tag expressing good, bad or neutral and an argument such as a subject or object that receive the sentiment associated with the verb. With other words than verbs, they simply tagged

them with their polarity and their part-of-speech tag. After the tagging is done, they use a syntax parser (shallow parser) to find phrases where subject, predicate and object were bound together. When they had found the syntactic dependencies, they then extracted the phrases that contained words that were also in their manually constructed dictionary and assigned the sentiment that existed in the dictionary. For this experiment they used recall and accuracy to interpret their results. Recall refers to the number of phrases that their system interpreted as containing a sentiment out of all the phrases containing a sentiment. Accuracy measures how many of those phrases were labeled with the correct sentiment. They conducted two experiments and received an accuracy well over 90% with this approach but the recall was below 30% for both cases.

It is important to note that this 90% accuracy cannot be compared with the work of Pang et al. (2002) and Turney (2002) for several reasons. Nasukawa and Yi (2003) did not perform a sentiment analysis on the document level. The purpose of their work was defined by themselves as to try and "find sentiment expressions for a given subject and determine the polarity of the sentiments." They also had a very low recall, meaning that their system did not find more than two thirds of all the sentiment expressions in their data. The authors explained that this was due to the fact that some of the sentiment expressions were located in different sentences meaning that they could not be chunked together by the shallow parser they used, and a large part of the sentiment expressions were in quite long sentences that contained nested sub-clauses, and this caused the shallow parser they used to miss these sentiment expressions. They motivated their choice of a shallow parser and not a full parser with that there were many typographic errors and ill-formed sentences that would have made a full parser produce a non-reliable output. But they suggest that it might make sense to use a full parser instead in order to find more sentiment expressions.

2.1.4.2 Recent research

Since the articles presented in the previous section were published 14-15 years ago, there has been much research done on sentiment analysis, but the supervised approach presented by Pang et al. (2002) and the unsupervised approach presented by Turney (2002) still represents the most mainstream approaches of performing a sentiment analysis. The models Pang et al. (2002) and Turney (2002) presented have been extended by utilizing more sophisticated NLP-methods in order to eliminate noise from the data

but the accuracy of document level sentiment analysis has not improved much and lies slightly over 80% for the supervised methods and around 70% for the unsupervised methods. (Li and Liu, 2013)

Li and Liu (2013)

Li and Liu (2013) presented a clustering-based approach to document level sentiment analysis first in 2010. The approach was further developed in 2013. Clustering is an unsupervised machine learning technique where the goal is to group the data into clusters where the data objects have similar characteristics. Li and Liu (2013) explains that using an unsupervised technique has many advantages and also some disadvantages compared to using a supervised machine learning technique. The advantages with unsupervised techniques are that it is much less expensive since it does not require any annotated training data, which reduces the amount of required human labor quite substantially. No required training data makes an unsupervised approach domain-independent and language-independent. The largest drawback with unsupervised approaches is that they generally have lower accuracy than supervised techniques. (Li and Liu, 2013).

To perform clustering on data in natural language, the data need to be converted into a structured format. Li and Liu (2013) uses the same data as Pang et al. (2002), i.e. movie reviews where the reviews are labeled as positive and negative. Note that the sentiment class label is not used in creating the clusters, only for validation of the model. Li and Liu (2013) creates two sparse matrices where adjectives and adverbs from the text are represented as the features. The values in the first matrix is presence (true or false) and the values in the second matrix is frequency (the number of times a word occurs in the text).

The clustering method they use is the k-means algorithm. K-means clustering is a technique that divides the data into different partitions without any internal hierarchy. When applying the K-means algorithm on these two matrices, they received an accuracy of 55% with a standard deviation of 2.6% for the data containing adjective and adverb frequencies and a standard deviation of 4.9% for the data containing adjective and adverb presence. Since 55% is a value way too low for considering this a useful model, they attempted to improve the accuracy of the model by adding a weighting method called TF-IDF (Term Frequency-Inverse Document Frequency). This is a method used to “evaluate how relevant a word in a corpus is to a document.” (Li and Liu, 2013) By applying this

weighting model, the accuracy increased to 72.2% and 73.1% respectively but the standard deviation also increased to 4.0% and 6.7%. To solve this problem with a relatively high standard deviation, the authors set up a voting mechanism by running the clustering algorithm multiple times and assigning a document to the class to which it is assigned to the most number of times.

To include linguistic knowledge into the model, the authors added two values that represent the distances (WordNet Synonym Distance) to the words *good* and *bad*. The words that they were not able to link to neither the word good nor bad, were discarded from the model, which greatly reduced the dimension of the matrices. This further improved the accuracy to 77.88% and 77.25% with 0.4% and 0.9% as the standard deviations. By adding the WordNet Synonym Distance, it is also possible to tell which cluster that is more associated with positive words and which cluster that is more associated with negative words. (Li and Liu, 2013)

Lin and He (2009)

Lin and He (2009) presents an unsupervised alternative to sentiment analysis, but they do use prior information in order to improve their results, which makes it not completely unsupervised. It is called the joint topic—sentiment model, and the purpose is to find the topics in a document and their related sentiments. The authors mean that sentiment polarities are dependent to topics or domains, which means that by using this technique the user can focus on the aspects in a document that are relative to what the user is trying to do. As stated by Turney (2002) about movie reviews: the final opinion about a movie is necessarily not a sum of its parts, meaning that a movie review can be very critical to some of the aspects of a movie (for example special effects) but the overall sentiment on that movie can still be very positive.

The joint sentiment/topic model extends the Latent Dirichlet Allocation (LDA) model that is maybe the one that is most used when trying to find the topics of a document. (Blei et al., 2003) By using the joint sentiment/topic models, we can find both the document level sentiment and a set of topics in that document. Since the model is quite mathematically complex, I will only present the results that Lin and He (2009) achieved by using this model. They did experiments with different kinds of prior information incorporated into the model and the highest accuracy they achieved on a movie review dataset was 84.6%. This accuracy level was achieved by incorporating a subjectivity list,

which included positive and negative words that occurred more than 50 times in the whole dataset of reviews.

2.1.4.3 Sentiment analysis across languages

Since sentiment analysis is done on natural language, the differences between different human languages needs to be considered. A complete automatic sentiment analysis system needs to be able to use different languages as input data. Since most of the resources within natural language processing and sentiment analysis exists in English, and the costs for creating labeled training data or sentiment lexicons are quite high, many have explored the opportunity to use machine translation to make use of annotated training data that exists in a different language. While this approach has shown to work well in some settings, machine translation does not perform equally well for all languages and especially morphologically complex languages might require more extensive preprocessing techniques to produce an acceptable accuracy. (Abdul-Mageed et al., 2011)

Boiy and Moens (2009) conducted a sentiment analysis on English, French and Dutch texts. They found a need to preprocess the data in the different languages because of differences between the languages. For example, in the Dutch language, adjectives can be glued together with another adjective, noun or verb to form a new word which incorporates the meaning of both those words. For example, “topfilm” would be the Dutch word of the English expression “top movie”. This may cause potential problems when conducting a sentiment analysis using the bag-of-words model, because the English expression “top movie” would be two different words for the classifier while “topfilm” would be one word. This language feature does not exist only in the Dutch language but also in, for example, Swedish (“jättestor”) and Finnish (“hyväkuntoinen”). They also found that the English data they used had the simplest vocabulary while the French had the richest vocabulary, which also plays a part when using the bag-of-words model. Further, they also pointed out that there are notable differences between how the writers in the different languages express sentiment. The French and the English writers prefer to express a feeling towards an entity while Dutch writers prefer to describe the entity with an adjective. These language-specific features play an important role especially when translating the text from another language into English and conducting a sentiment analysis using a model trained on data in English. A translation system will not translate

the Dutch sentence “the movie is good” into the English sentence “I like the movie”, which means that there will be semantical differences. Boiy and Moens (2009) received much better accuracy on the English data than on the French and the Dutch data. They attribute this difference to the sparsity of the training data and the non-formal language that was used especially within the French data.

Denecke (2008) compared the results of a sentiment analysis conducted on data in English and on data translated from German into English. The polarity classifier in SentiWordNet was used to calculate the sentiment of the documents and no difference was found between the two approaches. The accuracy Denecke (2008) received was under 66% for both the German movie reviews and the data in English, which is anyhow not very reliable and suggests that there were issues with the training data since many others have managed to achieve a much higher accuracy. Many have been sceptic to the machine translation approach since they believe that there are language-specific aspects to sentiment that we completely lose when we are translating a text into another language. Hogenboom et al. (2013) found through their research that in lexicon-based sentiment analysis, translating a text into another language in order to use the available sentiment lexicon is less fruitful than creating a specific sentiment lexicon for the language in question.

Another study investigating the effect of translation was done by Dadoun and Olsson (2016) who investigated the effect of translation on the sentiment analysis. They used tweets in Swedish as their data but excluded hashtags, emoticons and user IDs. They then constructed a sentiment lexicon of 30 positive and negative words in Swedish with which they then conducted a sentiment analysis simply by counting the occurrences of the sentiment words in the dataset. This result was then compared with a sentiment analysis done via a supervised machine learning approach on same data but translated into English with the help of Google Translate. They used a Naïve Bayes Classifier that was trained on a movie review data set containing 1000 positive and 1000 negative tweets. This is a quite popular dataset used by Pang and Lee (2008). They found that 5.2% of the tweets were classified differently by the two approaches. In this approach, it is very difficult to say if the difference can be attributed to the language itself since different methods of analysis was used. The dataset was also quite small, only consisting of about roughly 300 tweets.

Duh et al. (2011) further investigates the problem regarding machine translation and sentiment. They argue that even though we would have a perfect machine translation system that would not only syntactically perfectly but also semantically perfectly translate a text into a different language, the translation process will still introduce important differences in the data. They argue that the word distribution, which is what classifiers use as input, will change when translating a text. For example, the word "awesome" might be translated into the equivalent of "excellent", which is semantically correct but for a classifier it is a totally different word. If a classifier is trained on data where the word "awesome" is heavily linked to a positive text and then the test data uses the word "excellent" instead to express the same sentiment, it will introduce problems for the classifier. According to their experiments, where they experimented with test data in different languages and different domains, they found that translating training data and training a classifier on the translated data, resulted in the same kind of accuracy decrease as training a classifier on data from a different domain. They also found that the differences in cross-lingual settings cannot be eliminated by using domain adaptation algorithms. For this they did not find a reason, but cross-lingual differences seem to be of a different nature than monolingual domain-differences.

Another aspect of language that can have an impact on sentiment analysis is the demographic differences. Volkova et al. (2013) showed through their research that there exists a significant difference among male and female writers. This difference also differs across languages. For example, the difference in usage of words between English males and females are higher than the differences in Spanish and Russian. Volkova et al. (2013) where able to improve the subjectivity and polarity accuracies by incorporating the gender of the author as a feature.

According to the research presented above, machine translation might or might not be a valid tool for overcoming the problem with lack of training data for a specific domain. Successful studies exists, which suggests machine translation is a valid approach for sentiment analysis (Denecke, 2008) while others mean that there is a possibility of a language-specific aspect of sentiment that will impact the results for machine translated texts negatively (Boiy and Moens, 2009; Duh et al., 2011).

2.1.4.4 Impact of morphological complexity

In a bag-of-words model, a supervised classification algorithm uses the different words in a corpus as features and uses either the presence, frequency count or some other measure as feature vectors to create a classification model. Words are most often distinguished from each other by how they are spelled, and sometimes also spelling together with the words part-of-speech is used. Because of this, inflected words will automatically be treated as a different word. In morphologically simple languages, such as English, this is not a problem since, for example, adjectives are never inflected. If we change the phrase "the good car" into plural ("the good cars"), nothing happens to the adjective "good" and hence, a classifier will treat the adjectives in these two examples as the same word. However, in morphologically complex languages, such as Finnish, changing the same phrase "hyvä auto" into plural will not only inflect the substantive but also the adjective, and there are also two kinds of plural forms in Finnish, so depending on the situation the plural form of the phrase "hyvä auto" can be "hyvät autot" or "hyviä autoja". In this case, it would be beneficial if the classifier could treat the words "hyvä", "hyvät" and "hyviä" as the same word, since it expresses the same type of emotion in all cases. Lemmatization is one preprocessing step that would solve this problem.

Abdul-Mageed et al. (2011) states that morphologically rich languages create significant challenges for Natural Language Processing tasks in general and that sentiment analysis is no exception. Abdul-Mageed et al. (2011) conducted a subjectivity analysis and a sentiment analysis on sentences written in Modern Standard Arabic and analyzed how different preprocessing settings affect the result. They used a manually annotated dataset in the newswire domain together with a polarity lexicon consisting of 3982 adjectives that were marked as either positive, neutral or negative. They conducted the experiment with three different preprocessing settings, Surface (the word as it occurs in the text), Lemma (a word's lexical form) and Stem (the Surface excluding morphemes). The results of the analysis were that the Stem-setting was the one that performed the best, the Lemma-setting performed the second best and the Surface-setting performed the worst. In the sentiment analysis, 1-grams performed the best for all settings, which is in line with the findings of Pang et al. (2002). Adding the features *has_pos_adjective* and *has_neg_adjective,* calculated based on the polarity lexicon, improved the result with over 20%, which is a very large increase. (Abdul-Mageed et al., 2011)

Another experiment performed on French texts also show that lemmatized unigrams give better result compared to surface unigrams in a sentiment analysis (Ghorbel and Jacot, 2011). However, the analysis performed by Ghorbel and Jacot (2011) on the French data received a precision of over 90% even without the lemmatization, while the analysis performed by Abdul-Mageed et al. (2011) on Arabic texts only received an F-score around 55% (F-score is calculated based on the precision and recall).

Thus, the findings of Abdul-Mageed et al. (2002) and Ghorbel and Jacot (2011) show that lemmatization and stemming are both beneficial when performing a sentiment analysis on morphologically rich languages.

2.2 Linguistic aspects

Since the purpose of this thesis is to explore how different preprocessing settings affect the result of a sentiment analysis performed on Finnish text compared to English texts, I will now also present the most important linguistical differences between the languages. In this chapter, I will pinpoint the differences that might have an impact in a sentiment analysis.

2.2.1 The Finnish language

2.2.1.1 History

The Finnish language is a language spoken mainly in Finland, but also in small minorities in Sweden and Norway. The Finnish language belongs to the Finno-Ugric language family, which is quite different from the Indo-European language family to which a broad range of languages belong, including English, Swedish, Norwegian, German, Russian, Persian and Hindi. Other languages belonging to the Finno-Ugric language family are Estonian, Hungarian and Sámi as well as several smaller languages spoken in Russia in the areas around the Gulf of Finland. Finnish and Hungarian are often mentioned as the two languages that are the most prominent in the Finno-Ugric language family, but they are in fact not very much related to each other and the similarities can only be observed through linguistical study. Estonian on the other hand, is more similar to Finnish and a Finnish speaker and an Estonian speaker can quite quickly start understanding each other. (Karlsson, 2008)

Because Finland was a part of Sweden for so long, Finnish has over the centuries been heavily influenced by Swedish. The only official administrative language in Finland was Swedish for a very long time. However, during the nationalist movement during the 19th century, many loan words and Swedish influences in the Finnish language were forced out, but many still remains. The first book in Finnish was written in the 16th century by a bishop named Mikael Agricola, who also began the translation of the Bible into Finnish during the reformation. Over 5000 words of Agricola's vocabulary are still used today. Other written works that have had a large impact on the Finnish language are the national epic Kalevala compiled by Elias Lönnroth and others. (Karlsson, 2008)

Spoken Finnish can differ much from Standard Finnish, which is considered the normal written form. However, the standard language spoken in formal situations is quite close to the written norm, but colloquial spoken Finnish can be quite different, both grammatically and how certain words and phrases are pronounced. (Karlsson, 2008)

2.2.1.2 Characteristics

The basis of forming words in Finnish is the addition of suffixes and bound morphemes to a word stem. This is also the case for many Indo-European languages, but Finnish differ from them on a couple of points. First of all, the total number of case endings is significantly higher in Finnish than in Indo-European languages. Finnish has in total around 15 cases that are used to inflect words to create meaning in a sentence, which is usually handled by prepositions or postpositions in Indo-European languages. Another significant difference is that Finnish sometimes uses word endings where Indo-European languages uses words, for example possessive suffixes are used instead of possessive pronouns in Finnish (e.g. "autoni" is the equal of "my car"). Finnish also has a set of endings which meaning can only in other languages be translated into intonation (e.g. "-han"). (Karlsson, 2008)

The endings in Finnish can be mechanically piled up after each other and they have in several cases been used to create new independent words. Learning the endings is often thought to be difficult, but written Finnish is quite easy to analyze and understand what a word means if one knows what the different endings mean. Finnish does also not have articles or any grammatical gender like most Indo-European languages do. In Finnish, the semantic functions of articles are often created by changing the word order. For example,

“auto on kadulla” and “kadulla on auto” would in English translate into “the car is in the street” and “a car is in the street” respectively. (Karlsson, 2008)

Many people believe Finnish is very difficult to learn and Karlsson (2008) has listed some of the reason why. First of all, the words differ much from the words used in Indo-European languages. However, there exists quite many loan words from Swedish that are just spelled and pronounced slightly different, which then somewhat eases the burden for someone who is a native speaker of a Germanic language. An example is the word “coffee”, which is in Swedish and Finnish “kaffe” and “kahvi” respectively. Another difference that might seem difficult is that the word stem to which the suffixes are added, does not always stay the same. For example, for the word “käsi” the correct translation for the English phrase “in my hand” would be “kädessäni”, and the word stem is therefore “käde“. If we change this to plural; “in my hands” would then translate into “käsissäni” for which the word stem is “käsi-“. Thus, the basic stem can quite often change when certain endings are added to it. This means that if we are doing stemming and lemmatization on data in Finnish, stemming will produce more unique words than lemmatization will.

Maybe one of the most difficult parts of Finnish is that the case endings are not only applied to substantives, but also to verbs and adjectives. This language property also gives the language much flexibility in terms of word order and how a sentence of the same meaning can be constructed. (Karlsson, 2008) As for substantives, the case endings can often be applied according to place and direction. For example, “in the box” translates into “laatikossa”, but for the sentence “I am walking in the park”, the translation “Olen kävelemässä puistossa” is as valid as “Kävelen puistossa”. Both ways are widely used but the first one does indeed lack a valid literal translation into English (literally it would translate into “I am in my walking in the park”). In this case, lemmatization would return the word “kävellä” for both “kävelemässä” and “kävelen”, which might be beneficial in a sentiment analysis.

2.2.1.3 Spoken Finnish

When a non-Finnish person is learning Finnish, it is often the standard written Finnish that is taught in schools and language courses. While this is important and lays the foundation for understanding the words, morphology and syntax, there are quite many differences between spoken Finnish and standard written Finnish. While some say that

"Spoken Finnish is the same as written Finnish", they then mean that the phonemes (units of sound) always correspond to a certain letter. However, both the pronunciation, the morphology and the syntax can differ in spoken Finnish compared to the standard Finnish.

There are some specific letters of words that are very often left out in spoken Finnish. One of these letters is the final vocal of a word that has a long consonant before it. For example, the final letter is often not pronounced for the case endings -ssa, -ssä, -sta, -stä, -lla, -llä, -lta, -ltä, -ksi, -si and -isi. For example, the word "talossa" is often pronounced "talos". This can create confusion for non-native speakers since the word "talosi" can be shortened in the same way so that it becomes the same word "talos". The spoken word "talos" can in this case have two different meanings that can only be determined based on the context. (Karlsson, 2008)

Also letters in the middle of a word are in many cases left out when there is a diphthong. For example, "punainen" often becomes "punanen" and "tuommoinen" often becomes "tommonen". Words with diphthongs in the end of the word are pronounced in a slightly different way in the sense that the first vocal replaces the second vocal so that it becomes one long vocal. For example, "kauhea" becomes "kauhee" and "tärkeä" becomes "tärkee". In addition, almost all of the numerals are shortened and perhaps the most extreme cases are the way magnitudes of tenths are pronounced. For example, "kaksikymmentäseitsemän" (eng. twenty-seven) is very often shortened to "kakskytseittemän". Another aspect of the spoken language that can be confusing is that the passive forms are often used instead of the first person plural ending. For example, "sanomme" becomes "sanotaan", "tulimme" becomes "tultiin" and so on. (Karlsson, 2008)

These examples above are just some of the differences between the Finnish spoken language and the written language. Karlsson (2008) lists more omissions and assimilations as well as changes of form that happen in the Finnish spoken language. These differences might make the language somewhat more difficult to learn for people who are learning Finnish since the language they are taught in school or language courses are different than the language that people use when they are engaging in conversations. On discussion forums and social media, it is very common to write many words in the way they are pronounced in spoken Finnish. This is possible since phonemes almost all the time match a specific letter. This is also possible in English, but to a much lesser

extent since, as we will see below, the connection between spelling and pronunciation is more irregular.

2.2.2 The English language

2.2.2.1 History

English is one of the world's most spoken languages. Around 800 million people speak English word wide and about 350 million people have English as their mother tongue. English is an official language, or has special status, in over 60 countries. There are several versions of English and the three most dominant are British English, American English and Australian English. American English is today the most dominant version if English, for which the United States' mass media and entertainment industry is responsible to a large extent. (Nelson, 2002)

The history of English is very different compared to the history of Finnish. The English language as of today is a story of cultures in contact for around 1500 years. Although English has its origin in the British Isles, the development of the language has been greatly influenced by many different cultures and languages. The language has also been subject to political, economic and social forces, and such events over the years have influenced the language's vocabulary, accents and even the structure and grammar. The Roman Christianizing of Britain, the Scandinavian invasions and the Norman conquest were all events which pushed the language in different directions. Later, the expansion of the British empire and the rise of science and literature gave birth to different versions of the English language and made it move in different directions at different geographic locations, hence the differences between British English, American English and other versions of English that exists today.

2.2.2.2 Characteristics

One of the most prominent characteristics of English is the large and diverse vocabulary. While English is classified as a Germanic language, more than half of its vocabulary comes from Latin. No matter whether a person speaks a Germanic language, or a language derived from Latin as their first language, English will have characteristics in both cases that seem familiar to the person. When studying the English vocabulary, one can notice that the English language even has words that are borrowed from languages such as

Persian, Hindi, Arabic and Bengali, to name only a few. Therefore, the vocabulary is perhaps one of English greatest assets since it makes the language more familiar to speakers of different languages. (Baugh and Cable, 1993)

Since English is spoken in so many different countries and the language has changed much over the years, the grammatical rules about how to use English are not as clear as in Finnish. Phonetics and spelling are the two areas where most of the differences exists between the most used variations of English; American English, British English and Australian English. Since English is also considered a Lingua Franca, there are countless of different accents. People coming from different language backgrounds tend to do different errors in terms of spelling, word sequence and pronunciation. (Baugh and Cable, 1993)

English has both *prescriptive* grammar rules and *descriptive* grammar rules. Prescriptive rules are more general guidelines about how to use the language. To never end a sentence with a preposition is an example of a prescriptive rule. However, the prescriptive rules are not always followed, and we can create a perfectly valid English sentence even though a prescriptive rule is broken (i.e. “He is the man I will vote for.”). Descriptive rules are more concerned about how a language is used and consists of rules that cannot be broken, for example that the subject should be before the verb. (Aarts and McMahon, 2008)

Just like in Finnish, words are made up of prefix, base, and a suffix. The base is most often a word that can stand alone on itself, and by adding either a prefix and/or a suffix, the meaning of the word can be changed. However, there is a major difference between English and Finnish when it comes to inflections. As mentioned before, all the case endings in Finnish can be applied to both substantives, adjectives and verbs, which makes Finnish quite an inflectionally complex language. English however is inflectionally very simple. Also compared to other Germanic languages, English lacks almost all of the complicated agreements that might make, for example German, difficult to learn for a non-native speaker. The only inflections of the substantive that English has is the plural form and a form of the possessive case. This results in that words in English will occur more often in their lexical form than they will in Finnish. (Aarts and McMahon, 2008)

Let’s take as an example the sentence “in the good car”. Here, the adjective “good” and the substantive “car” is in its lexical form, and there also exist a preposition (“in”) and a

determiner (“the”), which gives information about locality and if we are referring to a definite or an indefinite element of the class “car”. The Finnish translation would be “hyvässä autossa”, where both the adjective and the substantive is inflected and replaces the preposition. However, no determiner exists nor does the case ending give any information which would replace the determiner, so here we have no way of knowing if we are referring to a definite or an indefinite element of the class “auto”. These differences do not give a native speaker of respective languages any troubles nor has the level of inflectional complexity shown to have any impact whatsoever on a child’s ability to learn the language as their first (Baugh and Cable, 1993). However, the example above shows that preserving the semantics when translating English into Finnish, or vice versa, is not an easy task.

Another aspect that is lacking from the English language but is a very important part in German and other European languages, is the gender of substantives. English has a neutral gender, while German, for example, has feminine, neutral and masculine. This means that in the German language, not only the word itself must be learned but also the gender, since it affects how the pronouns are inflected and also how adjectives are inflected in connection with the substantive. (Baugh and Cable, 1993)

Based on what is presented so far in this chapter, English might seem like an easy language to learn, but English has its fair share of complexities as well. Idioms might differ some between English and other languages, but on the other hand every language has its own way of expressing different things. The most difficult thing to learn for a non-native speaker is probably the connection between spelling and pronunciation, which can be quite unregular. Just think of the numbers, one (1) and two (2), and how they are pronounced. It is totally different than how the same combination of letters is pronounced in the words “alone” or “twilight”. In this sense, Finnish is simpler since every letter basically has its own sound which is always the same.

2.3 Establishing hypothesis

Sentiment analysis can be performed by analyzing the content (words and/or word order) of a sentence or document that expresses a subjective opinion. We also know that the means used to express the same opinion can be quite different depending on factors such as the writer, the target domain and the language in which the sentence or document is

written. The purpose of this thesis is to find out if the Finnish language require a different approach than English for performing a sentiment analysis. To concretize the main question in this thesis, I have formed the following hypothesis based on the theory presented in the theoretical framework:

> H1: Data in Finnish require different preprocessing steps than data in English in order to find the most sentiment-loaded features to create a supervised sentiment analysis classification model.

First, a baseline will be established by performing a sentiment analysis on both languages on tokenized data with stop words removed. Then, different preprocessing techniques will be applied to the data for both languages, and the performance will be calculated for each preprocessing setting and compared to the performance of the baseline. If the accuracy change is significantly different for the Finnish data than the accuracy changes for the English data, my hypothesis will prove to be true. Since there might be differences in quality between the English data and the Finnish data, I will not compare the results of the two languages to each other, only how the preprocessing settings affect the results for both data sets.

3 DATA AND RESEARCH METHODS

3.1 Model

I will use a model called the *bag-of-words* model. Bag-of-words is a supervised statistical learning model, which has been successfully applied to a large extent in text classification but has also proved to be successful in sentiment analysis. In the bag-of-words model we assume that it is the words used in a text that contain the most important information about the polarity of that text. Bag-of-words is not as concerned about word order and grammar. The bag-of-words model first creates a vocabulary from all the documents in the data. The words that go into this vocabulary can be chosen based on different factors. For example, words that occur in less than 1% of the documents might be ignored, since they might cause overfitting. After the vocabulary is chosen, a sparse matrix is created with the words as features and, for example, the presence or the frequency of those words in each document as feature vectors. A classifier is then trained on the sparse matrix and learns the differences in word distributions between the different classes. No agreement exists yet in the field about which features that should be used in a sentiment analysis, since previous research has received different results with different features. The same goes for classifiers, even though Support Vector Machines and Naïve Bayes Classifiers are those that have proved to work best and are simple enough to be able to handle large datasets. (Scott and Matwin, 1998)

3.1.1 Features

The feature selection process is naturally a very important step in every machine learning application. There is much research done in this area but here I will only discuss what has proved to be relevant in the field of sentiment analysis. Many have experimented, through trial and error, with feature selection in a sentiment analysis context and some differing results exist regarding which features work best. Pang and Lee (2008) provide an extensive overview over which features that seem to work better than others.

In topic-based text categorization, term frequency has shown to be more informative than term presence. However, Pang et al. (2002) found that, in the movie review domain, term presence worked better than term frequency when conducting a sentiment analysis. They

explain that this might be one of the key differences between topic-based text categorization and sentiment analysis.

Word positioning is another debated aspect where research has shown conflicting results. Some argue that in a specific context, the sentiment that represents the whole text is more likely to be expressed in a certain part of the text, for example the last quarter. Using features consisting of only one word (unigram) versus using features consisting of more words (bigrams, trigrams, n-grams) is another debated aspect of features selection. Pang et al. (2002) showed that unigrams outperform n-grams, while Dave et al. (2003) show that n-grams work better in some contexts.

Part-of-speech is another widely used preprocessing step in sentiment analysis. POS-tagging can help with word sense disambiguation, meaning that certain words that are used as, for example, both nouns and verbs can be distinguished from each other by tagging them with their part of speech. Some have also experimented with using only adjectives as features, since research on subjectivity analysis has shown that the presence of adjectives correlates positively with sentence subjectivity. Some have also used certain part-of-speech patterns as features, for example an adjective followed by a noun (Turney, 2002). However, Pang et al. (2002) found that the most used unigrams outperform adjectives as features in a movie review context. Their theory to this is that sentiment is not only expressed with adjectives but can also be expressed by using verbs and nouns.

Incorporating information about syntax into the classification model, has shown to be useful when studying shorter pieces of text. In longer texts, the importance of syntax has shown to decrease and using only n-gram based features seem to perform as good (Pang and Lee, 2008). Syntactic parsing can be used to incorporate other types of important information, such as negation, into the training data. Negation has shown to be a useful feature and can raise the accuracy of the sentiment analysis (Na et al., 2004).

Another type of feature that has shown to be informative in sentiment analysis as well as in text categorization in general is something called TF-IDF (term frequency-inverse document frequency). As term frequency shows how prevalent a word is in a specific document, the inverse document frequency shows how important a specific word is to the whole corpus. For example, a word occurring frequently in only a few of the documents is likely to have quite a high importance for those documents, while a word occurring

frequently in all documents is not very likely to be important for any document. The formula for inverse document frequency is the following:

$$IDF(t) = 1 + \log\left(\frac{Total\ number\ of\ documents}{Number\ of\ documents\ containing\ t}\right).$$

Thus, for a rare term, the value for its inverse document frequency will be high. In TF-IDF, the term frequency is multiplied with the inverse document frequency in order to smooth out the differences and to make it a document-specific measure, since IDF is a corpus-specific measure. The formula for TF-IDF can be seen below:

$$TFIDF(t, d) = TF(t, d) \times IDF(t).$$

Terms that are too rare are usually not relevant, since they are unlikely to represent a feature that describes the class to which it belongs. Rare terms can be filtered out by using a threshold for how low a portion of the total number of documents that a term can exist in. Terms that are too common are also a problem for the same reason and they can be filtered out by either using a threshold or by using a stop-words lexicon to remove words that are considered common in a specific language and are not used to express sentiment. TF-IDF is usually calculated as one of the last steps in the preprocessing phase before training the classifier (Provost and Fawcett, 2013).

Consequently, in a text categorization problem, both too common words and too rare words are filtered out in the preprocessing step. Filtering out terms that are too rare be done by using a stop-words lexicon or thresholds for how large a portion of documents a term is allowed to exist in versus how few documents a term is allowed to exist in.

Prior research in the field seems to be somewhat contradictory regarding which features work best. There are several possible explanations for this. Different classifier algorithms might work better with different features. Another explanation might lie in the data, since sentiment can be expressed differently by different people. Depending on the forum in which the texts are written, a certain type of dialect could be used if the group of people producing the texts is a homogeneous enough group. This means that using different datasets might be a reason why different research projects show different results, but this is quite difficult to prove.

3.1.2 Classifier

Many different classifiers have been used in the sentiment analysis domain. Pang et al. (2002) compared the performance of a Naïve Bayes Classifier, a Support Vector Machine and a Maximum Entropy classifier. In their experiment, the SVM classifier reached the best accuracy 82.9% when using unigram term-presence as features. However, the accuracies for the Naïve Bayes Classifier and the Maximum Entropy classifier where not very far off with 81.0% and 80.4% respectively. Since the objective of this study is to compare the differences between doing a sentiment analysis in English with doing one in Finnish, I will use the same classifier for both analyses. I have decided to use the Naïve Bayes Classifier since it is simple, fast and has proven to be a good baseline classifier for sentiment analysis.

The Naïve Bayes Classifier is derived from the Bayes Theorem, which is a statistical formula used to compute the probability of an event based on prior knowledge. The Bayes Theorem is a widely used formula and many of the existing data science methods use Bayesian methods. However, explaining Bayesian methods more broadly is out of scope for this study. Below can be seem the formula for the original Bayes Theorem.

$$p(A|B) = \frac{p(A)p(B|A)}{p(B)}$$

In the formula, we want to compute the probability of A given the condition B in a setting where A is dependent on B. In a sentiment analysis context, A would here represent the class, positive or negative, and B would represent the feature vector (b_1, b_2, b_3, … b_n). However, when calculating the probability of A given B in a setting where B represents lots and lots of features, the model would become very complex if we would treat all features as dependent on each other. Most of the Bayesian methods deal with this problem and the solution is to assume a probabilistic independence among the features, meaning that we assume that the probability of b_1 is not dependent on the probability of b_2 and so on. The assumption that the probability of b_1 is independent of b_2 will not be true in many cases and that is the reason it is called the *Naïve* Bayes Classifier. The base formula for the Naïve Bayes Classifier is the following:

$$p(A|B) = \frac{p(b_1|A) * p(b_2|A) \dots p(b_n|A) * p(A)}{p(B)}$$

In a classification problem, p(B) will never even have to be calculated, since it will be the same no matter the value of A, so we can simply look at which nominator is the bigger one when we change the value of A. Thus, the value of A for which the nominator is largest, is the class that the classifier chooses for the feature vector B. If we would want to find a probability value that indicated how probable it is that a certain feature vector belongs to a certain class, we would need to calculate p(B). However, this is where the Naïve Bayes Classifier has one of its largest drawbacks and will always overestimate the likelihood of it belonging to a specific class. As mentioned earlier, in the Naïve Bayes classifier we assume that the features in B (b_1, b_2, … b_n) are independent of each other in order to simplify the calculation of p(b_1 and b_2 … and b_n | A). This means that the occurrence of one feature is assumed to not impact the probability of the occurrence of another feature. When dealing with textual data, this is simply not true because of how natural language works. If we have two features that are heavily dependent on each other it means that if one is present then the other one is also almost always present, this leads to the classifier counting the same evidence twice and thus overestimating the likelihood of it belonging to a specific class. Given that the data points the classifier in the right direction, it will still classify it correctly. In this study, this drawback does not cause problems but in a professional setting where the output of a sentiment analysis will be used as input for decision making, we might also want to know the accurate confidence of the classification decisions. Naïve Bayes is not the correct classifier to use if there is a need to include confidence thresholds into the application. (Provost and Fawcett, 2013).

An advantage of the Naïve Bayes Classifier is that it is an incremental learner and can update itself as more data becomes available and thus increase its accuracy on the go. It is, despite its simplicity, used for instance in spam detection systems, which is a typical text classification problem. Another advantage of the Naïve Bayes Classifier is that it does not require much computing power nor disk space in order to train or classify. The Naïve Bayes Classifier will serve as a good classifier for this study.

3.2 Data

In order to perform this analysis, I have used two separate datasets that both consists of movie reviews. One of the datasets is in English and is the same dataset that Pang and Lee (2002) used in their analysis. It consists of 2000 movie reviews, that are pre-annotated

as either positive or negative. This dataset will be used to represent the baseline for my analysis. The dataset is available through the Natural Language Toolkit package for the Python programming language.

The second dataset is in Finnish and also consists of movie reviews. The data is gathered in a html format from http://www.leffatykki.com on the 29.11.2017. To obtain the data, the programming language Python was used together with various packets to navigate and scrape the site for movie reviews. All in all, 14332 reviews in Finnish were obtained together with a rating between 1 and 10 which the reviewer assigned to the movie with 1 being the worst and 10 being the best. This rating is used to determine which reviews are positive and which are negative.

3.2.1 Data preprocessing

To perform a sentiment analysis, meaningful features need to be extracted from the datasets so that the sentiment classifier can be trained as accurately as possible. The different features that are usually used in sentiment analysis are described in section 3.1.1. In this chapter, I will describe the steps taken in order to create meaningful features. For the dataset in English, the freely available Natural Language Toolkit (NLTK) and Sci-Kit Learn packages in Python are used to perform the different operations we need.

For the baseline, only tokenization and stop words removal will be used. Then I will apply lemmatization, part-of-speech tagging and TF-IDF score in order to create the different preprocessing settings.

Tokenization

Tokenization, as described in chapter 2.1.2, is the process of creating a list of words (or tokens) from of a raw text and is usually the first step in the preprocessing phase of a sentiment analysis. The English dataset I will use in this study is already tokenized while I will use NLTK to tokenize the raw data in Finnish.

Stop-words removal

As mentioned in chapter 3.1.1 regarding features, text classification problems benefit from not having corpus-wide high-frequency words or punctuation marks included as features. NLTK provides lexicons of stop-words both in English and Finnish and I will map these lexicons against my data in order to remove the stop-words.

Lemmatization and part-of-speech tagging

After calculating the baselines, I will apply more sophisticated Natural Language Processing techniques and compare the differences in the classification result on the English data and the Finnish data. NLTK does not provide any lemmatization or part-of-speech tagging methods for Finnish, only for English. This is a result of what was mentioned before; that the research within NLP has been focused mostly on the English language. Since Finnish is quite a different language than English on many aspects, it also requires a different approach when constructing POS-taggers and lemmatization toolkits.

For the lemmatization and part-of-speech tagging on the Finnish data, I will use a package named FinnPos, which is especially designed for POS-tagging and lemmatization on natural language in Finnish. Apart from the part-of-speech tagging and lemmatization, FinnPos also handles negation tagging of negated words. (Silfverberg et al., 2016)

3.2.2 Descriptive statistics

I will now present some descriptive statistics about the data I will use and observe potential differences between the two datasets. As mentioned before, the English dataset is already labeled with positive and negative while the Finnish dataset is only labeled with a rating ranging from 1 to 10. This means that we need to draw a line between positive and negative and probably also discard the ratings that represents reviews that are neither positive nor negative.

3.2.2.1 Finnish dataset

The data obtained consists of more than enough data to be able to make a meaningful sentiment analysis as we have a total of 14332 reviews. As we can see from table 1 and table 2, the data is not equally distributed across the ratings scale and there are less than 100 reviewers who have assigned the rating 1 to a movie. The reason for this might be that the reviewers assigned the ratings in the form of stars, with one rating accounting for half a star and thus, people may rather assign one whole star for a bad movie instead of just half of a star. In total, a movie can receive everything between half a star as the lowest rating up to 5 stars as the highest rating. I will not refer to the stars in this study, but to the ratings scale 1-10.

	review
Rating	
1	93
2	560
3	569
4	1072
5	1047
6	2114
7	2187
8	3163
9	1686
10	1841

Table 1 – Ratings frequency for the Finnish dataset

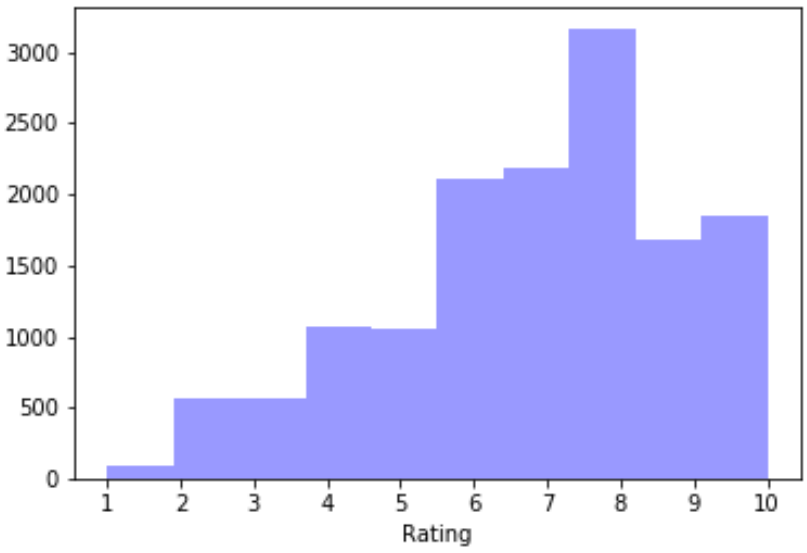

Figure 1 - Graphical ratings frequency for the Finnish dataset

There is no clear explanation to why the data is heavily skewed towards more positive ratings. The total number of reviews that are given a rating of 6 or higher are 10991, or 76.7% of the total number, which means that less than a quarter of the reviews are assigned a rating of 5 or less. One reason for the ratings leaning more towards the positive side of the scale might be that movies in general entertain people and therefore

the average movie might be enjoyable to watch. Another reason might be individual factors such as how one interprets the rating scale. Perhaps, if a movie has a rating between 1 and 5, it might in general be interpreted as bad and not worth watching while everything from rating 6 and up might range from okay to masterpiece, but this is very difficult to prove.

	Rating
count	14332.000000
mean	6.930226
std	2.191200
min	1.000000
25%	6.000000
50%	7.000000
75%	8.000000
max	10.000000

Table 2 - Summary statistics for ratings distribution in the Finnish dataset

As we can see from table 3, the mean is 6.93 and the median is 7.00, so quite close to each other. What we can ask here is if the rating 7 is the average movie and the sentiment poles can be found under seven and above seven. Here are some examples of sentiment-loaded sentences from some of the reviews:

Rating 1:

"En voi suositella tätä Ö-luokan teosta kenellekään. Tunsin tyhmentyväni (jos se edes on mahdollista) entisestään tätä katsellessani. American Battleship on elokuvaa huonoimmillaan. Typerää, mielenkiinnotonta, eikä lainkaan viihdyttävää. Toivon sydämeni pohjasta, ettei tällaista verkkokalvot saastuttavaa roskaa enää tarvitse katsella."

Rating 3:

"Jackson on kyllä ollut mukana melko kyseenalaisissa pläjäyksissä, mutta olisi jo luullut käsikirjoituksen perusteella ymmärtävän, ettei tästä leffasta ole mihinkään. Eli elokuva

on jopa toimintaviihteenä huono. Tämä ei viihdytä ollenkaan, kuten yleensä huonotkin toimintaelokuvat tekevät. Lähes kaikki tässä pläjäyksessä on pielessä. En tajua lainkaan, miksi tällainen roskaläjä on tehty. Ja saipa tämä vielä vuonna 2005 jatko-osankin, jota en kyllä ainakaan tämän perusteella halua nähdä."

Rating 5:

"Vaikka elokuvassa onkin ongelmia, ja varsinkin Mannajan alku on hieman väsähtänyt, lähtee se loppuaan kohden lopulta hienosti käyntiin. Väkivallan kuvaus on synkkää, rujoa ja parhaimmillaan välittää tuskan katsojalle erittäin hyvin. Loppu on töksähtävyydessään toimiva ja sulkee kokonaisuuden hyvin. Jopa "pahis" McGowanin hahmo saa lopussa kiitettävällä tavalla syvyyttä. Ja vaikkei Maurizio Merli olekaan päärolissa erityisen hyvä, ja hänellä on kovin kiiltävä hymy villin lännen desperadoksi, seuraa Bladen ennalta-arvattavia seikkailuja ihan mieluusti koko elokuvan ajan."

Rating 7:

"Uuno Turhapuron saaga loppuu arvokkaasti, kauniisti ja sujuvasti puutaheinää selittäen, vaikka sisältö ei todellakaan loista. Sama saattaa päteä koko leffasarjaan: sujuvaa ja kiistatta miellyttävää puutaheinää, mutta varsinaista sisältöä on erittäin vähän. Muistakaamme Uunoa siis viihdyttävänä, värikkäänä tyhjänpuhujana."

Rating 9:

"Elokuvan ainoana pikkuriikkisenä miinuksena voisi sanoa sen, että Banen ääntä on hiukan muutettu tietokoneella, jotta siitä saisi paremmin selvää(tai ainakin näin Internet väittää). Yön Ritarin Paluu on luultavasti paras ikinä tehty elokuva ja se antoi minulle sellaisen olon, että minun ei tarvitse katsoa enää yhtään muuta elokuvaa sen jälkeen. Voin ylpeästi sanoa, että kauan odotettu Batman -trilogia sai sen arvolle sopivan eeppisen päätöksen."

Rating 10:

"Kokonaisuutena Hullu Pierrot on hämäävä, katkera ja lumoava kokonaisuus, joka pitää katsojan vahvassa, absurdissa otteessaan ja tarjoaa paljon piilotettua sanomaa upeassa ja täydellisessä paketissa. Lopetus on tyly, makaaberi ja naulitseva - hulluus on todellakin

suhteellista. Katso tarkkaavaisesti, äläkä tallaa vaikkei se ensimmäisellä katselukerralla aukeaisikaan."

An interesting and important aspect that we can observe from the examples above is that a positive sentiment can be expressed without using adjectives. For example, the sentence "*…se antoi minulle sellaisen olon, että minun ei tarvitse katsoa enää yhtään muuta elokuvaa sen jälkeen…*" does not contain a single adjective but still expresses an overwhelmingly positive opinion. On the other side, the last example with rating 10 is filled with adjectives that describe the movie. This means that sentiment can be expressed both by using lots of adjectives but also by not using a single adjective. Boiy and Moens (2009) concluded in their paper that there are differences across languages how sentiment is expressed, so knowing how sentiment is generally expressed in the target language might be of importance.

Based on the random examples above, it suggests that a movie review with rating 5 or below expresses much negative sentiment. At rating 7, the author seems to express both positive and negative sentiment while in the reviews with rating 9 and 10, the sentiments are overwhelmingly positive. Since these are just random examples of a few sentences in the reviews, further analysis could be performed on the data to illustrate where the line can be drawn between negative and positive reviews.

Ratings and part-of-speech

To be able to identify some possible differences in the usage of part-of-speech classes across different ratings, I calculated the frequency of different part of speech classes in the reviews. From table 4, we can see the differences in the average frequencies of the most used part-of-speech classes, plotted by rating for the whole dataset. At this point, no stop-words are yet removed, but I chose to leave out the part-of-speech classes "punctuation", "numbers" and "unknown" from the plots to reduce potential noise. As can be observed from the charts, the frequencies of different part-of-speech classes seem to be slightly different across the ratings scale. However, important to keep in mind when

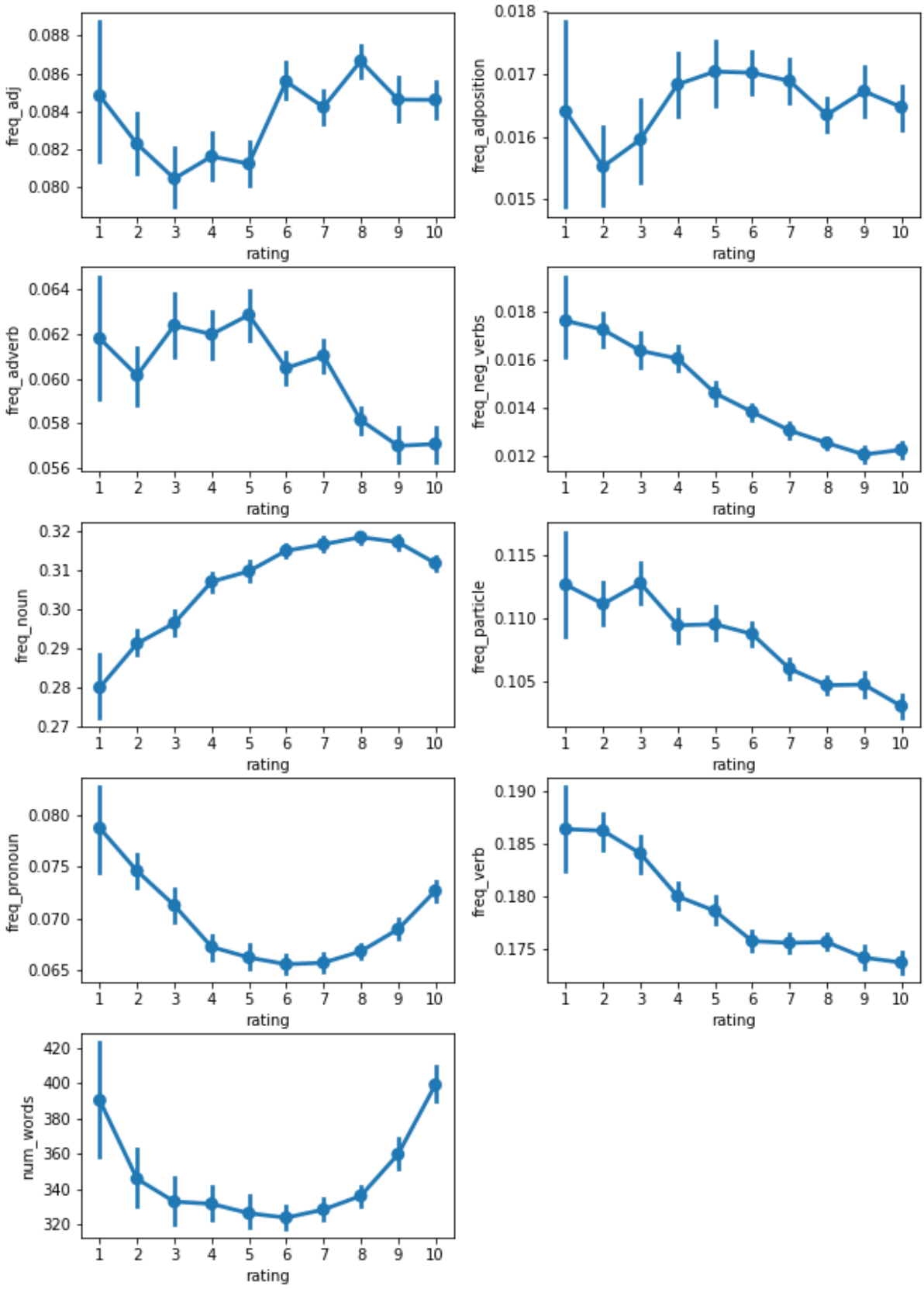

Figure 2 - Part-of-speech class frequency across ratings in Finnish dataset

interpreting these plots is the unequal distributions of data points across the ratings scale, which we can see from the high spread of data points for the lower ratings in every chart.

The chart suggests that the reviews tend to become longer as the ratings approach the minimum and the maximum values. Pronouns and nouns seem to change in opposite directions as the rating changes, which is logical since pronouns are replacement words for nouns. However, the usage of pronouns and nouns seem to respectively increase and decrease when the ratings approach the minimum and maximum values. The usage of non-negated verbs is higher in negative reviews than in positive reviews and the same is true for particles. The negated verbs interestingly follow the same pattern as non-negated verbs, which suggest that verbs altogether are used less in positive reviews. The usage of adjectives, on the other hand, does not follow a similar pattern as verbs and nouns. However, the differences here are indeed very small so it is not possible to draw any conclusions from it.

In table 5 we can see the Spearman's correlation coefficients between the different part-of-speech classes including the rating, which give a better view on which part-of-speech classes correlates better with the rating. None of the part-of-speech classes have a strong correlation with rating as we can see from table 5. Negated verbs has the strongest correlation with a value of -0.17. Nouns seem to have the strongest correlation values with the rest of the classes in general with all of them being negative. As table 5 suggests, pronouns and nouns have the strongest correlation with a value of -0.49, but none of them seem to correlate very well with rating.

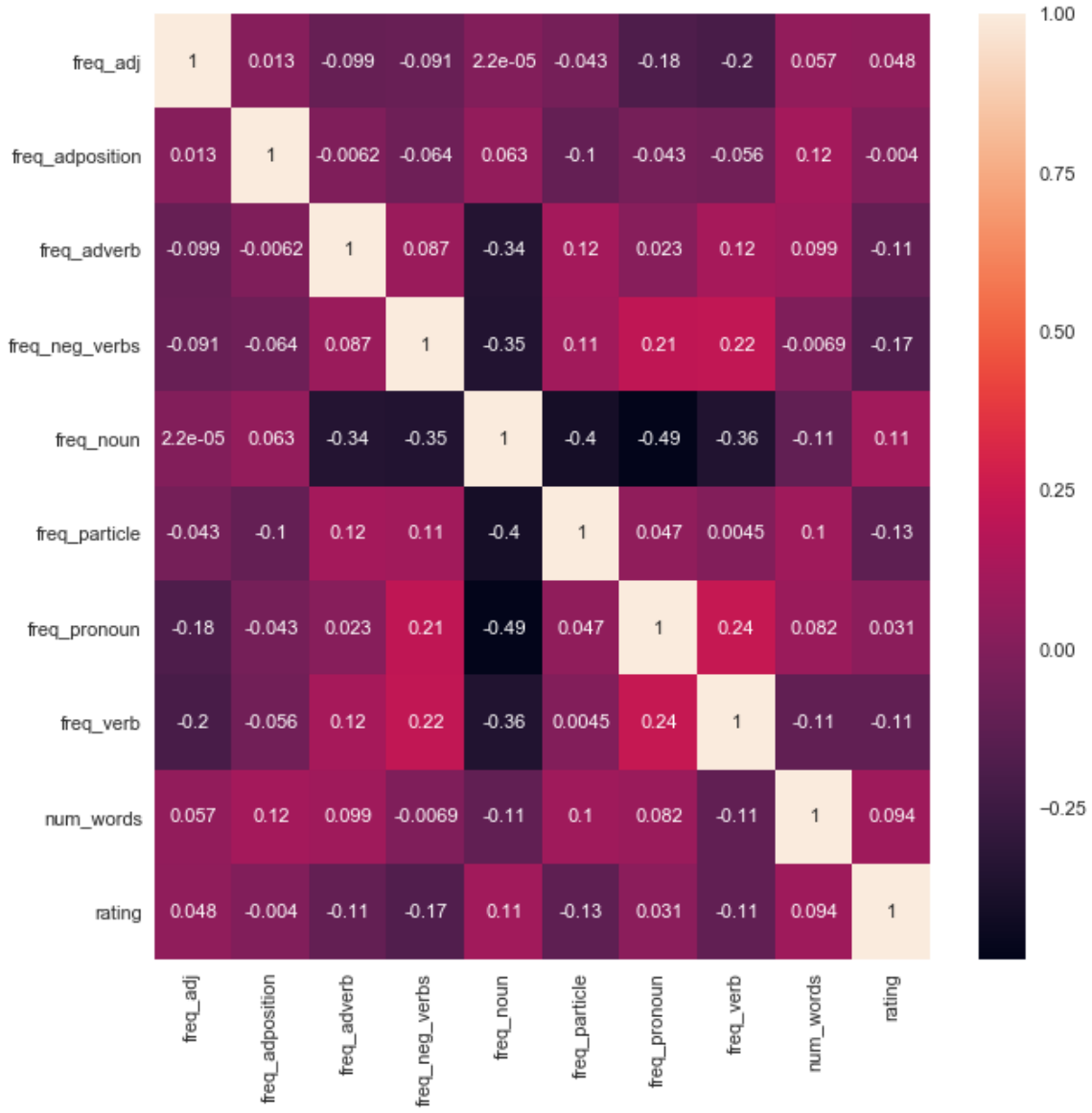

Figure 3 - Correlation matrix for part-of-speech classes in Finnish dataset

3.2.2.2 English dataset

The data in the English dataset, that will be used to identify what features that are more informative for data in Finnish, is the same dataset as the one used by Pang and Lee (2002). It is freely available with the python package Natural Language Toolkit and contains 2000 reviews. The data is already split into positive and negative reviews and as we can see from table 6. The data is also equally distributed between the two classes.

sentiment	review
neg	1000
pos	1000

Table 3 – Sentiment class frequencies for the English dataset

Here are some examples of sentences from the negative and positive reviews in the dataset:

Negative:

The actors are pretty good for the most part, although Wes Bentley just seemed to be playing the exact same character that he did in American Beauty, only in a new neighborhood. But my biggest kudos go out to Sagemiller, who holds her own throughout the entire film, and actually has you feeling her character's unraveling. Overall, the film doesn't stick because it doesn't entertain, it's confusing, it rarely excites, and it feels pretty redundant for most of its runtime, despite a pretty cool ending and explanation to all of the craziness that came before it.

Positive:

The print I saw wasn't finished (both color and music had not been finalized, so no comments about Marilyn Manson), but cinematographer Peter Deming (Don't say a word) ably captures the dreariness of Victorian-era London and helped make the flashy killing scenes remind me of the crazy flashbacks in twin peaks, even though the violence in the film pales in comparison to that in the black-and-white comic. Oscar winner Martin Childs' (Shakespeare in love) production design turns the original Prague surroundings into one creepy place. Even the acting in From Hell is solid, with the dreamy Depp turning in a typically strong performance and deftly handling a British accent.

As we can see from the examples, the line is not necessarily crystal clear here since some aspects can be positive and some negative even though the reviews themselves respectively leans towards the positive or the negative side. However, Pang and Lee (2002) achieved an 82.9% accuracy on this same dataset, which suggests that the data is good enough also for this analysis.

Part-of-speech

The part-of-speech tagger for the English dataset included two additional part-of-speech classes compared to the Finnish part-of-speech tagger. These two classes are determiner and conjunction. However, almost all the words that belong to these two classes are stop words and will be discarded from the dataset prior to analysis. These two classes have therefore been discarded from the charts in this chapter. From table 7 we can see how the different part-of-speech classes are distributed between the two sentiment classes where 1.0 means positive and 0.0 means negative.

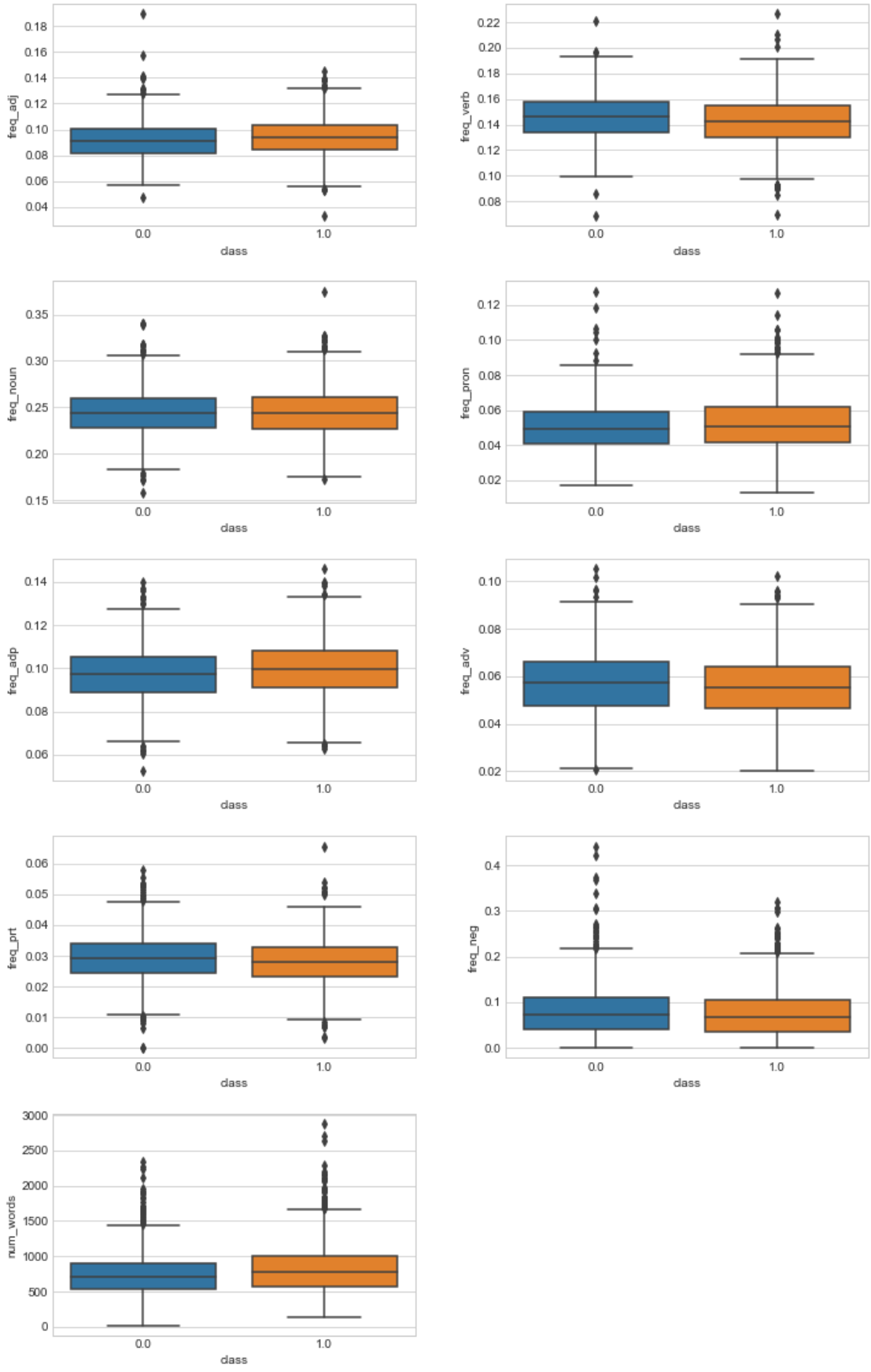

Table 4 - Part-of-speech class frequency across sentiment classes in English dataset

As we can see from the table 7, the differences are very small but notable. As with the Finnish data, negation seems to be the factor where the largest differences between positive and negative can be observed. So even though sentences like “the movie was not bad” is a sentence fitting a positive opinion, it seems that negation is still more widely used in the dataset when expressing a negative opinion, for example “the movie was not good”. In table 8 we can see the Spearman’s correlation of the different part of speech classes and the rating.

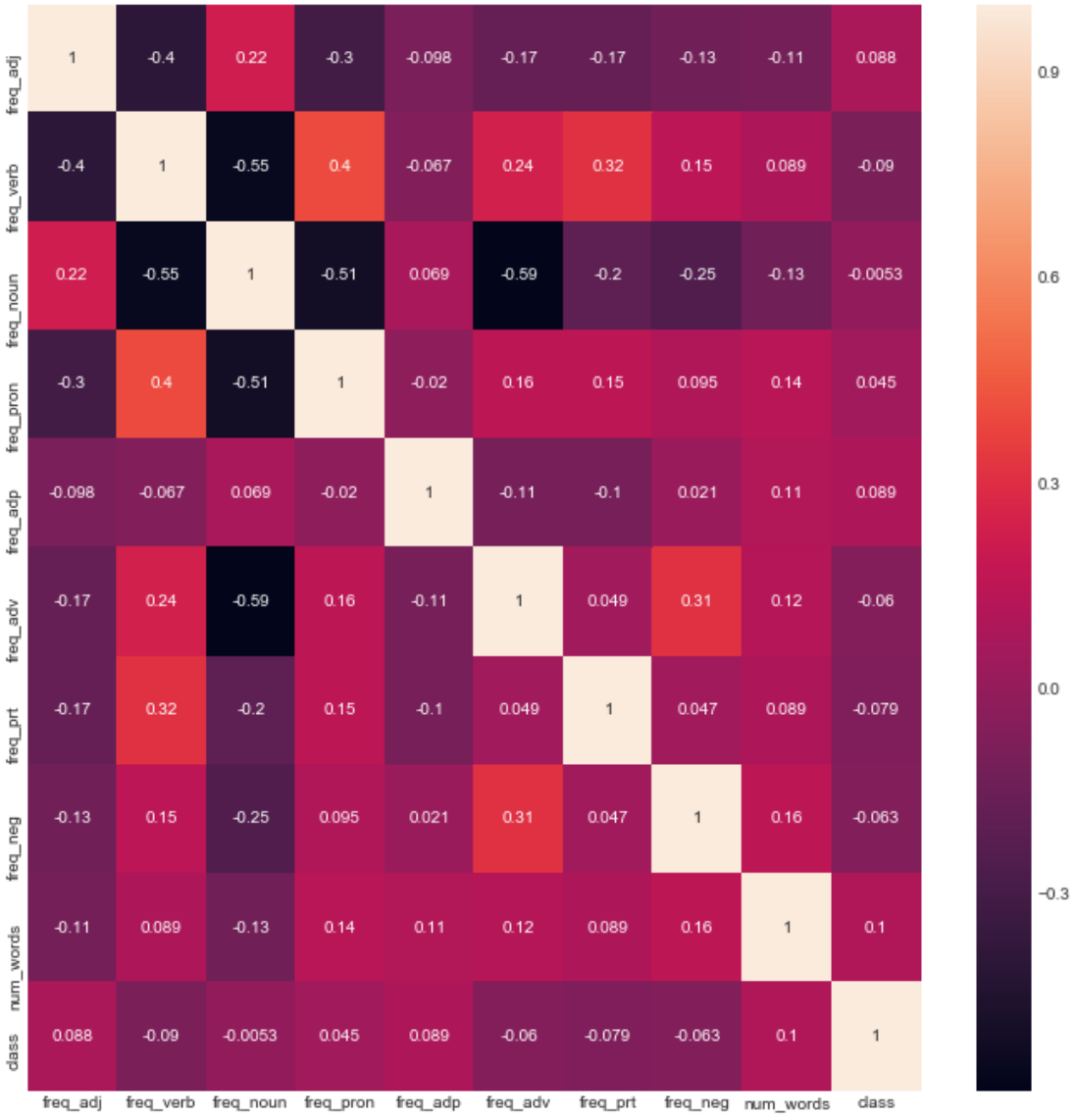

Figure 4 - Correlation matrix for part-of-speech classes in the English dataset

Some of the part-of-speech classes in the English dataset seem to have quite strong either positive or negative correlation, but none of the classes correlate particularly well with the sentiment class.

3.2.2.3 Lemmatization and stop words

Since Finnish is a morphologically complex language and English is not, I expect that we will see differences in classification performance between the two datasets when I run the analysis on lemmatized features. As we can see from table 9, the number of unique words decrease more for the Finnish dataset than for the English dataset and there are more words not originally in their lemmatized form in the Finnish dataset. In table 9 and table 10, we can see the statistics calculated based on the lemmatization for the two datasets, with punctuation, numerals, unknown words and truncated words not included in the calculations, since they cannot be inflected.

	changed_words	total_words	freq_changed
count	14332.000000	14332.000000	14332.000000
mean	160.140385	287.689018	0.555374
std	80.532718	140.574493	0.042901
min	35.000000	64.000000	0.291971
25%	106.000000	193.000000	0.527673
50%	141.000000	255.000000	0.556338
75%	193.000000	345.000000	0.583333
max	1155.000000	2026.000000	0.748252

Table 5 - Lemmatization summary statistics for the Finnish dataset

	changed_words	total_words	freq_changed
count	2000.000000	2000.000000	2000.000000
mean	51.975500	673.173000	0.077159
std	24.654367	297.401516	0.013837
min	1.000000	17.000000	0.035019
25%	35.000000	472.000000	0.067990
50%	48.000000	631.000000	0.076483
75%	64.000000	815.250000	0.085776
max	172.000000	2482.000000	0.147239

Table 6 - Lemmatization summary statistics for the English dataset

As expected, we can observe from the tables that the Finnish dataset has more words that differ from their base form than the English dataset has. The average ratio of inflicted words per review in the Finnish dataset is 55.5% with a max value of 74.8% and minimum value of 29.2%. In the English dataset the average is 7.7% with a max value of 14.7% and minimum value of 3.5%. However, we can also see that the reviews in the English dataset has on average more words than the reviews in the Finnish dataset. The reason for this is that prepositions are used in the English language to express syntax, while prepositions are not used almost at all in Finnish and instead morphemes are used to express syntax. This means that the lemmatization process is eliminating many of the syntactical tools used in the Finnish language and the number of unique words become less when lemmatizing. The change in the number of unique words can be seen from the tables 11 and 12.

	unique_non_lemmatized	unique_lemmatized	unique_removed_perc
count	14332.000000	14332.000000	14332.000000
mean	227.757256	195.600684	0.133273
std	98.877387	79.679841	0.039062
min	59.000000	56.000000	0.000000
25%	160.000000	140.000000	0.105528
50%	206.000000	179.000000	0.132353
75%	273.000000	233.000000	0.159722
max	1149.000000	832.000000	0.300885

Table 7 – Lemmatization's effect on unique words in the Finnish dataset

	unique_non_lemmatized	unique_lemmatized	unique_removed_perc
count	2000.000000	2000.00000	2000.000000
mean	335.837000	327.18400	0.024656
std	111.761607	107.72734	0.009143
min	17.000000	17.00000	0.000000
25%	258.750000	253.00000	0.018405
50%	322.000000	315.00000	0.024194
75%	399.000000	388.00000	0.030712
max	1049.000000	1020.00000	0.060172

Table 8 - Lemmatization's effect on unique words in the English dataset

As we can see from table 12, the English data has more unique words than the Finnish dataset in table 11, and the effect of the Lemmatization is much smaller. For the English dataset, the average decrease of unique words because of the lemmatization is only 2.5% while it is 13.3% for the Finnish dataset. Thus, at this stage we have more words in the English dataset than in the Finnish dataset.

In order to actually see the differences in word counts between the two datasets, stop words need to be removed. The stop-words removed from the data can be seen from

Appendix A. In table 13 we can see how many stop-words the reviews contain on an average for the two datasets.

	eng_words	eng_stop	fin_words	fin_stop
count	2000.000000	2000.000000	14332.000000	14332.000000
mean	791.910000	308.650500	343.399177	62.458973
std	347.338096	142.591681	168.437989	32.920875
min	19.000000	11.000000	78.000000	4.000000
25%	560.000000	214.000000	230.000000	41.000000
50%	745.000000	287.000000	304.000000	55.000000
75%	957.250000	371.250000	411.250000	76.000000
max	2879.000000	1096.000000	2485.000000	584.000000

Table 9 - Summary statistics about stop-words removal

The English dataset contains more stop-words per review, which was expected. But even after removing the stop-words, the reviews in English contains more words on average. The possible reasons for this can be that the English dataset simply contains longer reviews or that the stop-word lists are not really matching each other.

3.2.2.4 Summary

The tables below show the average distribution of words between the word classes for each dataset.

POS-class	**English**	**Finnish**
Adposition	9.87%	1.66%
Adjective	9.28%	8.44%
Punctuation	13.90%	13.89%
Pronoun	5.11%	6.81%
Noun	24.49%	31.30%
Particle	2.87%	10.66%

Unknown	0.17%	1.23%
Verb	14.44%	19.01%
Adverb	5.64%	5.96%
Numeral	1.00%	1.04%
Determiner	10.18%	-
Conjunction	3.05%	-

Table 10 - Average distribution of part-of-speech classes for both datasets

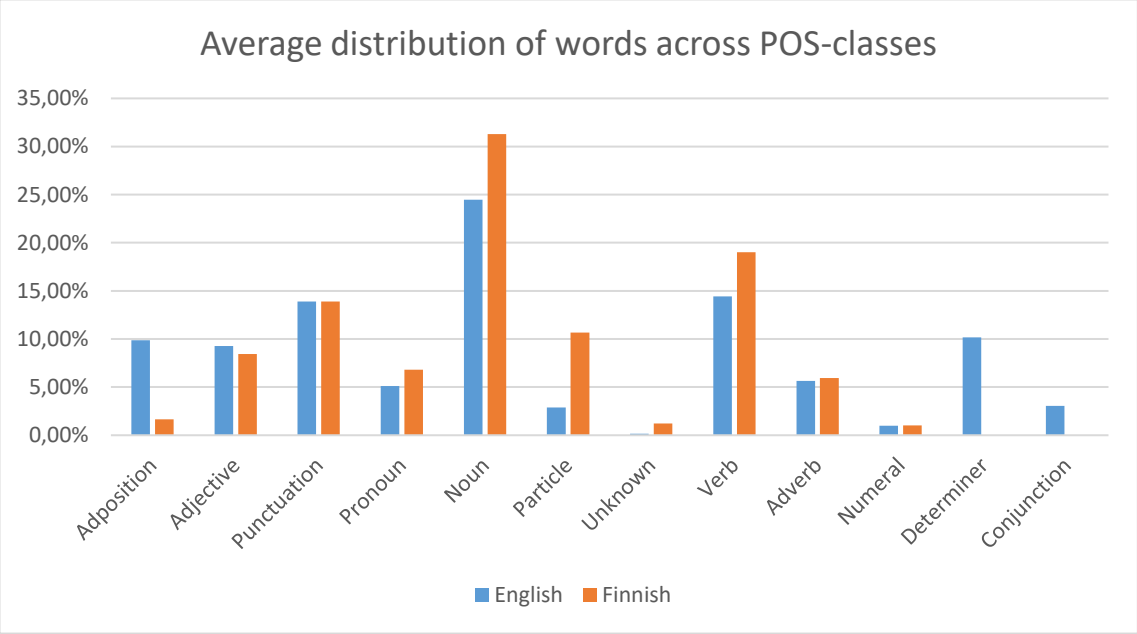

Figure 5 - Graphical distribution of part-of-speech classes for both datasets

If we look at the distributions of word classes between the two languages in table 14 and table 15, there are a few part-of-speech classes where there is quite a large difference in the average rate that they occur in the text. For example, determiners are completely missing in the Finnish language and adpositions are used at a much lower rate than in English. However, as I have mentioned earlier in this thesis, the lack of adpositions (prepositions and postpositions) in the Finnish language is compensated for by inflecting verbs and nouns, so therefore the verbs and the nouns make up a larger part of the Finnish texts than the English texts. One thing to be noted here is that there is a risk that the part-of-speech taggers have tagged the words differently, since the same part-of-speech tagger could not be used for both languages.

As the descriptive analysis shows, none of the part-of-speech classes seem to be very indicative of if a review is negative or positive. The part-of-speech class that has a small but observable difference in occurrence rate between negative and positive reviews for both data sets are verbs. Verbs occur slightly less in positive reviews than in negative reviews, but the differences are very small.

Negation is the feature that has the largest differences between positive and negative reviews for both data sets. This indicates that negated words are a useful feature when doing a sentiment analysis. However, the differences in average occurrence between negative and positive reviews are probably too small for negation to be used as a meta data feature. Since the descriptive analysis of the part-of-speech classes and negation shows that the differences in averages are very small, I will not experiment with using meta data features for the sentiment analysis. Instead, I will use unigram features, since it has shown to be effective in previous studies.

4 EXPERIMENT SETUP AND EMPIRICAL RESULTS

In this chapter, I will present how the analysis was conducted and what kinds of results were obtained by the different preprocessing settings. Both the preprocessing of the data and the results were produced with the Python programming language and Jupyter Notebook development environment. In the preprocessing phase, the pandas library and the Natural Language Toolkit library were used. For training the classification model and validating its performance, the Scikit-learn package was used. All software and data used in this thesis are open source and freely available.

4.1 Experiment setup

I have chosen 5 different test scenarios that are run for both the English data and the Finnish data. These scenarios are based on results from earlier research presented in chapter 2 and 3, and are meant to illustrate how the precision of the classification algorithm differs between the two data sets when different preprocessing settings and features are used.

4.1.1 Classes

The two datasets to be used were presented earlier in chapter 3.2. The English dataset is already divided into two classes, positive and negative. The Finnish dataset, however, is divided into ten classes where each class represents a rating that was given by the writer of the review. In order to perform the analysis in a way so that the results can be comparable to each other, the Finnish dataset needs to be divided into two classes where one represents negative reviews and the other represents positive reviews.

It is not clear where the line should be drawn between positive and negative based on the ratings scale and going through all of the 14,332 reviews and marking them as either positive or negative would be an enormous amount of work. Therefore, I decided to make the split between positive and negative based on the descriptive statistics calculated in chapter 3.2.2.1. It can be observed that the distribution of the ratings is skewed towards higher ratings and the mean rating is 6.93, so not exactly in the middle. For this analysis, I have decided to assume that the ratings 9 and 10 represent positive reviews and the ratings 1, 2, 3, 4 and 5 represent negative reviews. The reviews with ratings 6, 7 or 8 have

been discarded from this analysis, since they are assumed to represent a neutral opinion towards the movie. As can be seen from table 11, this gives me a dataset with 3341 negative reviews and 3527 positive reviews, which is a fairly equal distribution between the two classes.

	review
sent	
neg	3341
pos	3527

Table 11 - Sentiment class division for Finnish dataset

4.1.2 Preprocessing and features settings

The experiment is set up so that we have five different preprocessing settings.

In the first setting, unigrams are extracted and stop words, punctuation and numerals are removed from the set of words. The stop word lists used are included in the NLTK Python package for both the English and the Finnish language. The punctuation and numerals removed are tokens that are marked with the part-of-speech classes *punctuation* and *numeral* by the part-of-speech taggers. To avoid unnecessary features and overfitting, words that occur in less than 0.5% of the documents and more than 90% of the documents are ignored. Word count is used as feature vector. These preprocessing steps result in a total of 5757 features for the Finnish dataset and 7647 features for the English dataset.

In the second setting, TF-IDF values are used as feature vectors instead of word counts. Otherwise, the same preprocessing steps are used as in setting 1.

In the third setting, lemmatization is added as a preprocessing step. This results in different features than in the two first settings and a smaller number of features. The total number of features for the Finnish dataset is in this setting 4876 and for the English dataset 7145.

In the fourth setting, negated words are added to the features. Words that are marked as negated by the part-of-speech taggers are marked with the prefix “neg_”. Because of this,

the number of features become higher. Here we have 7758 features for the English dataset and 4928 features for the Finnish dataset.

In the fifth and last setting, we only use lemmatized adjectives. This results in a much smaller number of features, with only 787 features for the Finnish dataset and 1545 features for the English dataset.

Preprocessing settings		
Setting #	Feature vectors	Preprocessing steps
Setting 1	Word count	• Stop words removed • The POS classes Punctuation and Numeral removed
Setting 2	TF-IDF	• Stop words removed • The POS classes Punctuation and Numeral removed
Setting 3	TF-IDF	• Stop words removed • The POS classes Punctuation and Numeral removed • Lemmatization
Setting 4	TF-IDF	• Stop words removed • The POS classes Punctuation and Numeral removed • Lemmatization • Negation
Setting 5	TF-IDF	• All POS-classes removed except adjectives • Lemmatization

Table 12 - Experiment overview

4.1.3 Validation

To validate the results that the model produces, I use five-fold cross validation and calculate the average of the produced F1-scores. F1-score is a measure that includes both precision and recall and is widely used when rating the performance of a classifier. Cross

validation is a popular accuracy estimation method that is commonly used to validate how well statistical models generalize to an independent, real-world data set. A good model-validation method should be able to reduce bias and provide low variance, which k-fold cross validation handles well (Kohavi, 1995).

4.1.4 Results and discussion

The performance of the model created with the baseline setting shows that it was possible to obtain approximately a 10% better result on the sentiment analysis done on the Finnish data than on the English data. The reasons for this can probably be found by analyzing the differences between the datasets more in depth. Possible factors might be the size difference between the datasets, the differences in review lengths and the quality of the language used in the reviews. Comparing the baseline result obtained for the English dataset with the result that Pang et al. (2002) received show that it is quite similar. Pang et al. (2002) managed to obtain an accuracy between 77.3% and 82.9% for their experiments. Table 18 shows that the best F1-score my analysis managed to produce was 83.3%, which is better than what Pang et al. (2002) managed to produce. Pang et al. (2002) used a different classifier, different preprocessing steps and different feature vectors, which are probably the causes of the small differences.

F1-scores		
Setting	Finnish dataset	English dataset
1	90.4%	83.3%
2	90.3%	82.2%
3	88.8%	81.8%
4	89.0%	81.9%
5	85.1%	81.2%

Table 13 - Experiment results

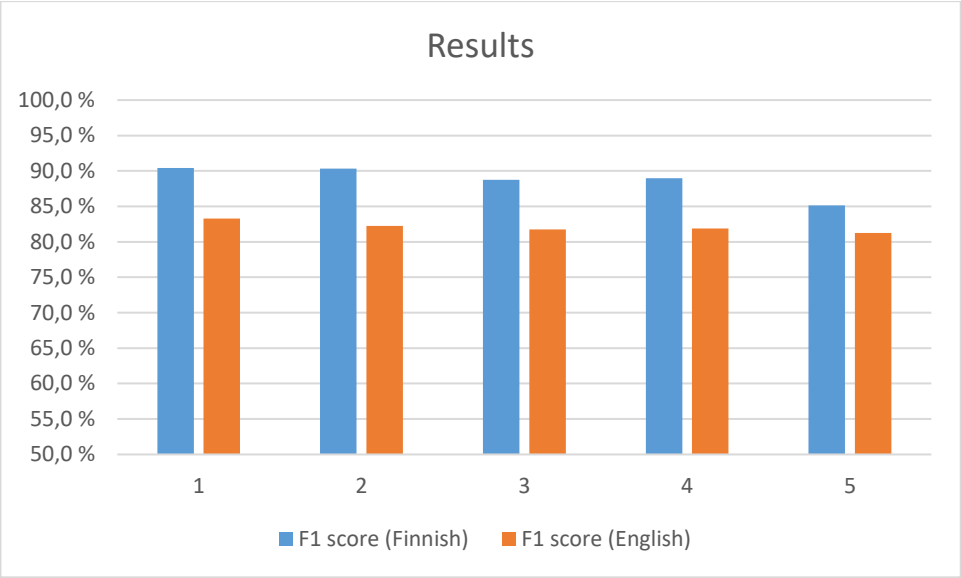

Table 14 - Experiment results bar chart

The most interesting thing to point out about these results is that neither lemmatization nor TF-IDF vectors did improve the score. The baseline gave the best performance for both datasets with an F1-score of 90.4% for the Finnish dataset and 83.3% for the English dataset. Using TF-IDF vectors compared to word count produced results that were 0.1% worse for the Finnish dataset and 1.1% worse for the English dataset. Adding lemmatization as a preprocessing step decreased the performance of the classifier further, with a decrease of 1.5% for the Finnish dataset and a decrease of 0.4% for the English dataset from setting 2 to setting 3.

The only preprocessing step that seemed to add performance to the model is the addition of negated verbs, but the difference in F1-score between 3 and 4 was very small with 0.2% for the Finnish dataset and 0.1% for the English dataset. The usage of negated verbs between the datasets also shows that negated verbs exist with a higher frequency in the Finnish dataset than the English dataset, which indicates that negation is more widely used in Finnish than in English to express sentiment.

Using lemmatized adjectives as features resulted in the worst classification performance for both models with a larger decrease for the Finnish dataset than for the English dataset compared to the baselines. Using lemmatized adjectives in the Finnish dataset produced a result that was 5.3% worse than the baseline, while using lemmatized adjectives in the English dataset only produced a score that was 2.1% worse than the baseline.

The expectation based on the theory presented in chapter 2 was that lemmatization would increase the F1-score for the Finnish dataset, since it would normalize the features used and remove syntactic noise that introduces differences between words that semantically are the same. However, one possible reason for the baseline producing such a good result for the Finnish dataset might be that it is subject to overfitting because of domain-specific morphology. This can, however, only be proven if the classifiers were tested on independent datasets in a different domain. However, if the purpose of the classifier is to classify movie reviews, then it is not a problem if the classifier is fitted to that domain only, but if the classifier is to be used for classifying different kinds of texts, then lemmatization might still be an option to consider.

One reason why setting 5 received a so much worse performance compared to the baseline for the Finnish dataset than for the English dataset might be that adjectives in the Finnish language are less indicative of the polarity of a text than adjectives in the English language. The average number of adjectives in the Finnish texts is 8.44% while the average number of adjectives in English texts was over 9.28%, so adjectives are slightly more frequently used in the English dataset.

The Finnish dataset used in this thesis seems to be better suited for sentiment analysis than the English dataset, since it produced better results for all the experiments. This can possibly be attributed to the overall quality of the data; however, no tests were performed in this thesis to test the grammatical correctness of the language used in the texts in the two datasets.

5 SUMMARY AND CONCLUSION

The goal of this thesis was to investigate whether linguistic differences between languages play a role in a sentiment analysis or not. As mentioned in chapter 2, sentiment analysis has many areas where it could be useful and because most of the research within the field of sentiment analysis is done on data in English, the best practices for how to design a sentiment analysis system are formed to suit data in the English language. Technological advancements should not be restricted only to a certain part of the world and, therefore, it is important to understand if different languages require different sets of practices for designing sentiment analysis systems.

Morphologically rich languages differ from morphologically simple languages in many ways. The descriptive statistics calculated in chapter 3.2.2 show a couple of notable differences between Finnish and English. First, the number of unique words per text is higher, while the total number of words is lower in Finnish. The reason for this is that morphemes in Finnish replaces the usage of prepositions in English, and many two-word expressions in English exists as a single word in Finnish. Second, the lemmatization process showed that there are more words in Finnish than in English, which do not appear in its lexical form. Thus, since classifiers are trained on data consisting of feature vectors created based on the words used in a text, the performance of a classifier could change considerably if more than 13% of the words are merged into existing words through lemmatization. In addition, removing prepositions from a text can in English be done by using word lists of stop words while in Finnish the morphemes handle the grammatical function of prepositions, which makes it somewhat more difficult to achieve the same type of preprocessing.

To answer the research question of this thesis, an empirical experiment was conducted on two separate datasets, one in Finnish and one in English, that consisted of movie reviews. The hypothesis was that data in morphologically rich languages require different preprocessing steps in order to find the most sentiment-loaded features. Especially lemmatization, which transforms a word into its basic lexical form (and hence ignores the morphemes), was expected to be a key step in the sentiment analysis of a morphologically complex language. However, the experiments undertaken to answer that question show

that the morphology of a language does not impact the result of a sentiment analysis significantly.

The best performance was received from using the baseline setting (setting 1) and the performance of the model did in fact decline as more preprocessing were performed on the data before training the classifier. The only step that improved the score was adding negated verbs and the effect was similar for both languages. The only notable difference between the two datasets was the usage of adjectives. Using adjectives as features for English data seem to be a viable option while the performance declined quite substantially for the Finnish dataset when using only the adjectives as features. This indicates that adjectives are used in a more extensive way in the English language to express sentiment and hence have a larger predictive value in a sentiment analysis done on English texts than what adjectives in the Finnish language have.

Since the different preprocessing settings did affect the performance of the two experiments in a very similar way, the hypothesis of this thesis cannot be accepted. To further investigate this subject, more sophisticated sentiment analysis techniques could be used. Bag-of-words is a fairly simple way of conducting a sentiment analysis and, for example, the joint topic—sentiment model could be used to understand more of the challenges that morphologically complex languages pose on the field of sentiment analysis. Using a different and more accurate classifier could be one way to understand more about the results, since a Naïve Bayes classifier cannot give an accurate estimate of the confidence of its classification decisions.

6 SWEDISH SUMMARY

6.1 Introduktion

Attitydanalys (eng: sentiment analysis) är ett område inom datorlingvistik (eng: natural language processing) som fått stor uppmärksamhet under de senaste åren. Attitydanalys innbär att automatiskt analysera en text och klassificera den som, till exempel, antingen positiv eller negativ. I en attitydanalys används oftast maskininlärning för att skapa en modell som lär sig känna igen de mest attitydbärande särdragen i en stor mängd texter och på basis av dessa särdrag räkna ut en sannolikhet för att texten är positiv eller negativ. Även symboliska metoder, det vill säga metoder som inte använder sig av statistisk slutledning, har använts inom forskningsområdet men metoder baserade på maskininlärning har visats ha en högre nogrannhet. Den teknologiska utvecklingen har under de senaste 10 åren möjliggjort användningen av attitydanalys genom att minska två av de hinder som tidigare gjort det problematisk att bygga ett attitydanalyssystem för en inte allt för dyr prislapp. Det ena är datamängderna som krävs för att träna en maskininlärningsmodell och den andra är datorprestanda. Tack vare sociala medier finns det idag näst intill oändliga mängder av fritt tillgängliga data i textform som kan användas som träningsmaterial för en maskininlärningsmodell. Datorer har även blivit snabbare och kan hantera större datamängder än tidigare och processera dem snabbare samtidigt som molntjänster har gjort det möjligt att för en begränsad tid köpa enormt kraftfulla datorer till ett pris som är mycket lägre än vad en dator med motsvarande prestanda kostar i sig själv.

Attitydanalys och datorlingvistik överlag skiljer sig från andra maskininlärningsproblem eftersom indatan som används är i formen av naturligt språk. Naturligt språk kan vara tvetydigt och samma ord och mening behöver inte alltid betyda samma sak. Dessutom blir maskininlärningsmodellen beroende av språket i sig. En modell tränad med data skrivna på engelska förstår förstås inte data skrivna på andra språk. Majoriteten av forskningen gjord inom attitydanalys är gjord med engelska data och därför håller praxis inom området på att formas kring data på engelska. Jag vill därför undersöka ifall det finns skillnader orsakade av språk i hur data i textformat språk bör processeras för att hitta de mest sentimentbärande aspekterna.

6.2 Forskningsfråga

I denna avhandling undersöker jag ifall skillnaderna mellan ett morfologiskt komplext språk och engelska har en påverkan på hur man hittar de mest sentimentbärande särdragen i texter. Två separata dataset bestående av filmrecensioner används. Det ena datasetet består av recensioner på finska och det andra består av recensioner på engelska. För att besvara forskningsfrågan undersöker jag hur klassificeringsresultaten för de båda dataseten ändrar när olika nivåer av förhandsprocessering används.

6.3 Teori

Attitydanalys hör till området datorlingvistik som i sin tur hör till området artificiell intelligens. Artificiell intelligens är ett forskningsområde som existerat redan en ganska lång tid. Många filosofer och vetenskapsmän har länge försökt förstå vad det mänskliga sinnet riktigt är och ifall det är möjligt att skapa en kopia som kan tänka och agera intelligent i en given omgivning. Det som vi idag förknippar med artificiell intelligens är ett forskningsområde som uppstod under 1940- och 1950-talet. 1950 publicerade Alan Turing sin idag berömda artikel ”Computing machinery and intelligence” i vilken han föreslog ett sätt att testa ifall en dator eller någon annan form av maskin, är intelligent eller ej. Redan på 1950-talet byggdes program som klarade av att spela schack med en människa. Neurala nätverk kom också att utvecklas under denna tidsperiod, men ett problem som bestod var hur man skulle få tillgång till den enorma datamängd som krävdes för att göra dessa applikationer ”intelligenta”. (Poole and Mackworth, 2010)

Att skapa en maskin som förstår naturligt språk har länge varit ett av de viktigaste underområden till artificiell intelligens. I forskningsområdets början troddes det att automatisk översättning var enkelt att lösa med en dator på grund av en dators förmåga att lagra stora mängder data i minnet och enkelt kunna komma åt vad som finns lagrat. De första översättningsapplikationerna var baserade på ordlistor men det visade sig vara ett otillräckligt tillvägagångssätt. Det kom att dröja en ganska lång tid innan det gjordes några större genombrott inom maskinöversättning men idag finns det maskinöversättningsapplikationer som de facto är mycket avancerade och fungerar bra. Idag är artificiell intelligens ett mycket omtalat ämne och det existerar höga förväntningar på dess potential. (Kumar, 2011)

Attitydanalys är ett relativt nytt område inom datorlingvistik som har blivit ett populärt forskningsområde under 2000-talet. Attitydanalys innebär att automatiskt med hjälp av en dator, analysera en text och klassificera den som, till exempel, antingen positiv eller negativ. Av de tillvägagångssätt som gett de bästa resultaten har de flesta använt sig av maskininlärningsmodeller av typen övervakad inlärning, som innebär att man använder träningsdata som färdigt innehåller en klassvariabel som beskriver vilka texter är positiva och vilka som är negativa. Med hjälp av dessa träningsdata kan man då träna en modell som då kan automatiskt klassificera nya texter på basis av mönstrena i träningsdatat. Orsaken till det ökade intresset för attitydanalys under 2000-talet är som tidigare nämnt, minskade kostnader för att skapa träningsdata och bättre tillgång till snabbare datorer. Dessutom finns det ett stort antal användingsområden för attitydanalys såsom politik, markandsföring och sökmotorer. (Liu, 2012; Pang and Lee, 2008)

Som tidigare nämnts, så har majoriteten av forskningen inom attitydanalys gjorts på Engelska data. Det finns dock en del forskning som undersökt effekterna av att använda maskinöversättning för att ta del av resurserna som finns på Engelska. Resultaten visar dock att maskinöversättning inverkar negativt på resultaten av en attitydanalys eftersom det finns språkspecifika aspekter som försvinner när man använder sig av maskinöversättning. (Boiy and Moens, 2009)

En språklig aspekt som tenderar att skapa lite större problem för metoder inom datorlingvistik överlag är morfologisk komplexitet. Morfologisk komplexitet innebär att ord i en mening tenderar att böjas väldigt ofta. Finska är ett morfologiskt komplext språk eftersom till exempel ändelser används istället för prepositioner. Forskning inom detta område har visat att det är en fördel att använda sig av lemmatiserade former av ord istället för den form som ordet förekommer i texten i. (Abdul-Mageed et al., 2011) Eftersom finska och engelska är två väldigt olika språk och hör till olika språkfamiljer, vill jag därför jämföra ifall textdata på finska kräver andra metoder för att processera data för en attitydanalys än vad textdata på engelska kräver.

6.4 Data och forskningsmetoder

Metoden som används för att utföra attitydanalysen är en metod som kallas *bag-of-words* (direkt svensk översättning: säck-med-ord) och innebär att analysera distributionen av ord som existerar i texten eller texterna. Bag-of-words antar att den viktigaste information

kan fås genom att analysera en texts orddistribution och metoden beaktar därför inte ordföljd alls. I bag-of-words använder man en given orddistribution som variabler för en observation. Variablerna kan mätas på olika vis, några av dessa är förekomst, antal eller TF-IDF (Term-Frequency Inverse-Document-Frequency). I denna avhandling används både antal och TF-IDF i de fem experimenten som utförs.

Klassificeringsalgoritmen som används är av typen naiv bayesiansk klassificerare. Algoritmen är matematiskt ganska simpel men har visat sig duga bra för attitydanalys. I ett professionellt attitydanalyssystem vill man dock undvika att använda naiv bayesiansk klassificerare på grund av att algoritmen har en tendens att överestimera sannolikheten för att en viss observation hör till en viss klass. Dock är kan algoritmen lära sig i flera inkrementella steg, det vill säga att man kan vartefter träna den med mera data för att få bättre klassificeringsresultat. (Provost and Fawcett, 2013)

Datan som används är som tidigare nämnt två stycken olika dataset som båda består av filmrecensioner. Det ena datasetet innehåller filmrecensioner på engelska och det andra datasetet innehåller filmrecensioner på finska. Datan på finska är samlad från hemsidan http://www.leffatykki.com och datan på engelska är hämtad från ett Python-paket som heter NLTK (Natural Language Toolkit). Den finska datan består av filmrecensioner och varje recension har ett betyg mellan 1 och 10, där 10 representerar det bästa betyget och 1 det sämsta betyget. Den engelska datan innehåller recensioner som redan är märkta som antingen positiva eller negativa. För att kunna göra en jämförbar analys, krävs att de finska recensionerna också märks som antingen positiva eller negativa. Från distributionen av den finska datan (Table 1 och Figure 1) kan vi se att den är inte jämnt fördelad över de olika betygsklasserna, utan en övervägande del har ett betyg på 6 eller högre. För att dela upp datan i två lika stora klasser, antogs att recensioner med betyget 9 eller högre är positiva och recensioner med 5 eller sämre är negativa recensioner.

För att processera datan används olika metoder, bland andra lexikalanalys, lemmatisering och ordklasstaggning. Utöver dessa borttas också så kallade stoppord från datat för att ta bort ord som inte kan anses innehålla subjektivitet. Dessutom märks också negerade ord, det vill säga ord som i texten har motsatt betydelse till vad de betyder ifall de står ensamma (t.ex. ordet ”bra” i frasen ”inte bra”).

Den deskriptiva statistiken som räknats ut på basis av datan och resultatet av ordklasstaggningen visar att distributionen av ordklasser inte är nämnvärt olika mellan de båda attitydklasserna ”positiv” och ”negativ”. Detta innebär att den språkliga syntaxen som används inte skiljer sig mellan att uttrycka positiva eller negativa åsikter. Dock används negering aningen mera när negativa åsikter uttrycks i text både inom engelska och finska.

6.5 Experiment

Experimentet är indelat i fem delar. De olika delarna består av olika nivåer av förhandsprocessering av datan. En klassificerare tränas i varje experiment både på den finska och den engelska datan. Resultaten av de fem experimenten jämförs sen för att se ifall något av dataseten svarar annorlunda på annan form av förhandsprocessering av datan. Resultaten av de fem experimenten valideras med hjälp av korsvalidering och resultaten mäts i F1-resultat som är ett statistiskt jämförelsetal som ofta används för att mäta noggrannheten av en klassificeringsalgoritm.

Resultaten av experimenten visar dock att de båda dataseten svarar likadant på de olika nivåerna av förhandsprocessering. Den ända märkbara skillnaden är att använda sig av adjektiv verkar fungera bättre för engelskspråkig data eftersom noggrannheten föll märkbart mera för det finska datasetet än för det engelska datasetet när endast adjektiv användes som variabler.

6.6 Sammandrag

Avhandlingens syfte var att undersöka ifall textdata på finska kräver andra metoder än textdata på engelska för att få fram de mest attitydbärande aspekterna i datasetet. Finska är ett morfologiskt komplext språk som är mycket annorlunda jämfört med engelska. För att besvara forskningsfrågan ufördes ett experiment i fem delar, där det undersöktes ifall data på de båda språken svarar annorlunda på olika nivåer av förhandsprocessering. Även om engelska och finska är två väldigt olika språk så visar resultaten i denna avhandling att dessa skillnader inte har desto större betydelse i en attitydanalys. Endast adjektiv verkar ha en större attitydbärande kraft i det engelska språket än i det finska språket, men

annars visar resultaten av denna avhandling att data på finska inte behöver processeras på ett annat vis än engelska data.

7 APPENDICES

7.1 Appendix A – Stop words

English stop words	Finnish stop words
i	olla
me	olen
my	olet
myself	on
we	olemme
our	olette
ours	ovat
ourselves	ole
you	oli
your	olisi
yours	olisit
yourself	olisin
yourselves	olisimme
he	olisitte
him	olisivat
his	olit
himself	olin
she	olimme
her	olitte
hers	olivat
herself	ollut
it	olleet
its	en
itself	et
they	ei
them	emme
their	ette
theirs	eivät
themselves	minä
what	minun
which	minut
who	minua
whom	minussa
this	minusta
that	minuun
these	minulla
those	minulta
am	minulle

is	sinä
are	sinun
was	sinut
were	sinua
be	sinussa
been	sinusta
being	sinuun
have	sinulla
has	sinulta
had	sinulle
having	hän
do	hänen
does	hänet
did	häntä
doing	hänessä
a	hänestä
an	häneen
the	hänellä
and	häneltä
but	hänelle
if	me
or	meidän
because	meidät
as	meitä
until	meissä
while	meistä
of	meihin
at	meillä
by	meiltä
for	meille
with	te
about	teidän
against	teidät
between	teitä
into	teissä
through	teistä
during	teihin
before	teillä
after	teiltä
above	teille
below	he
to	heidän
from	heidät
up	heitä
down	heissä

in	heistä
out	heihin
on	heillä
off	heiltä
over	heille
under	tämä
again	tämän
further	tätä
then	tässä
once	tästä
here	tähän
there	tallä
when	tältä
where	tälle
why	tänä
how	täksi
all	tuo
any	tuon
both	tuotä
each	tuossa
few	tuosta
more	tuohon
most	tuolla
other	tuolta
some	tuolle
such	tuona
no	tuoksi
nor	se
not	sen
only	sitä
own	siinä
same	siitä
so	siihen
than	sillä
too	siltä
very	sille
s	sinä
t	siksi
can	nämä
will	näiden
just	näitä
don	näissä
should	näistä
now	näihin
	näillä

	näiltä
	näille
	näinä
	näiksi
	nuo
	noiden
	noita
	noissa
	noista
	noihin
	noilla
	noilta
	noille
	noina
	noiksi
	ne
	niiden
	niitä
	niissä
	niistä
	niihin
	niillä
	niiltä
	niille
	niinä
	niiksi
	kuka
	kenen
	kenet
	ketä
	kenessä
	kenestä
	keneen
	kenellä
	keneltä
	kenelle
	kenenä
	keneksi
	ketkä
	keiden
	ketkä
	keitä
	keissä
	keistä
	keihin

	keillä
	keiltä
	keille
	keinä
	keiksi
	mikä
	minkä
	minkä
	mitä
	missä
	mistä
	mihin
	millä
	miltä
	mille
	minä
	miksi
	mitkä
	joka
	jonka
	jota
	jossa
	josta
	johon
	jolla
	jolta
	jolle
	jona
	joksi
	jotka
	joiden
	joita
	joissa
	joista
	joihin
	joilla
	joilta
	joille
	joina
	joiksi
	että
	ja
	jos
	koska
	kuin

	mutta
	niin
	sekä
	sillä
	tai
	vaan
	vai
	vaikka
	kanssa
	mukaan
	noin
	poikki
	yli
	kun
	niin
	nyt
	itse

8 REFERENCES

AARTS, B. and MCMAHON, A., 2008. *The handbook of English linguistics.* John Wiley & Sons.

ABDUL-MAGEED, M., DIAB, M.T. and KORAYEM, M., 2011. Subjectivity and sentiment analysis of modern standard Arabic, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2* 2011, Association for Computational Linguistics, pp. 587-591.

BAUGH, A.C. and CABLE, T., 1993. *A history of the English language.* Routledge.

BIRD, S., KLEIN, E. and LOPER, E., 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

BLEI, D.M., NG, A.Y. and JORDAN, M.I., 2003. Latent dirichlet allocation. *Journal of machine Learning research,* **3**(Jan), pp. 993-1022.

BOIY, E. and MOENS, M., 2009. A machine learning approach to sentiment analysis in multilingual Web texts. *Information Retrieval,* **12**(5), pp. 526-558.

BUCHANAN, B.G., 2005. A (Very) Brief History of Artificial Intelligence. *AI Magazine,* **26**(25),.

CHOWDHURY, G.G., 2003. Natural language processing. *Annual review of information science and technology,* **37**(1), pp. 51-89.

CROWL, J., 22.8.2018, 2018-last update, How Sentiment Analysis is Changing SEO and Improving User Experience. Available: https://www.skyword.com/contentstandard/marketing/how-sentiment-analysis-is-changing-seo-and-improving-user-experience/ [8.9.2018, 2018].

DADOUN, M. and OLSSSON, D., 2016. Sentiment Classification Techniques Applied to Swedish Tweets Investigating the Effects of translation on Sentiments from Swedish into English.

DENECKE, K., 2008. Using sentiwordnet for multilingual sentiment analysis, *Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on* 2008, IEEE, pp. 507-512.

DUH, K., FUJINO, A. and NAGATA, M., 2011. Is machine translation ripe for cross-lingual sentiment classification? **Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers**(Volume 2), pp. 429-433.

FELDMAN, S., 1999. NLP Meets the Jabberwocky: Natural Language Processing in Information Retrieval. *ONLINE-WESTON THEN WILTON-,* **23**, pp. 62-73.

GHORBEL, H. and JACOT, D., 2011. Sentiment analysis of French movie reviews. *Advances in Distributed Agent-Based Retrieval Tools.* Springer, pp. 97-108.

GOLDBERG, A.B. and ZHU, X., 2006. Seeing stars where there aren't many stars: Graph-based semi-supervised learning for sentiment categorization. *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing,* , pp. 45-52.

JAMES, G., WITTEN, D., HASTIE, T. and TIBSHIRANI, R., 2013. *An introduction to statistical learning.* Springer.

JORDAN, M.I. and MITCHELL, T.M., 2015. Machine learning: Trends, perspectives, and prospects. *Science (New York, N.Y.),* **349**(6245), pp. 255-260.

KARLSSON, F., 2008. *Finnish: An essential grammar.* Routledge.

KOHAVI, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection, *Ijcai* 1995, Montreal, Canada, pp. 1137-1145.

KRAUTHAMMER, C., 1997-last update, Be Afraid: The Meaning of Deep Blue's Victory [Homepage of The Weekly Standard], [Online]. Available: www.weeklystandard.com/be-afraid/article/9802 [2/7, 2016].

KUMAR, E., 2011. *Natural language processing.* IK International Pvt Ltd.

LANGLEY, P., 2011. The changing science of machine learning. *Machine Learning,* **82**(3), pp. 275-279.

LI, G. and LIU, F., 2013. Sentiment analysis based on clustering: a framework in improving accuracy and recognizing neutral opinions. *Applied Intelligence,* **40**(3), pp. 441-452.

LIDDY, E.D., 2001. Natural language processing.

LIU, B., 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies,* **5**(1), pp. 1-167.

MCCORDUCK, P., 2004. *Machines Who Think.* 2 edn. Natick, MA, USA: A K Peters Ltd.

MÜLLER, T., COTTERELL, R., FRASER, A. and SCHÜTZE, H., 2015. Joint Lemmatization and Morphological Tagging with LEMMING, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* 2015, pp. 2268-2274.

NA, J., SUI, H., KHOO, C.S., CHAN, S. and ZHOU, Y., 2004. Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews.

NELSON, G., 2002. *English: an essential grammar.* Routledge.

PANG, B. and LEE, L., 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval,* **2**(1–2), pp. 1-135.

PANG, B., LEE, L. and VAITHYANATHAN, S., 2002. Thumbs up? Sentiment classification using machine learning techniques, *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10* 2002, Association for Computational Linguistics, pp. 79-86.

POOLE, D.K. and MACKWORTH, A.K., 2010. *Artificial Intelligence: Foundations of Computational Agents.* New York: Cambridge University Press.

PROVOST, F. and FAWCETT, T., 2013. *Data Science for Business: What you need to know about data mining and data-analytic thinking.* " O'Reilly Media, Inc.".

SCOTT, S. and MATWIN, S., 1998. Text classification using WordNet hypernyms. *Usage of WordNet in Natural Language Processing Systems, .*

SILFVERBERG, M., RUOKOLAINEN, T., LINDÉN, K. and KURIMO, M., 2016. FinnPos: an open-source morphological tagging and lemmatization toolkit for Finnish. *Language Resources and Evaluation,* **50**(4), pp. 863-878.

SMOLA, A.J. and VISHWANATHAN, S., 2008. *An Introduction to Machine Learning.* Cambridge: Cambridge University Press.

TURING, A.M., 1950. Computing machinery and intelligence. *Mind,* **59**, pp. 433-460.

TURNEY, P.D., 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews, *Proceedings of the 40th annual meeting on association for computational linguistics* 2002, Association for Computational Linguistics, pp. 417-424.

TURNEY, P.D., 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL, *European Conference on Machine Learning* 2001, Springer, pp. 491-502.

VOUTILAINEN, A., 2003. Part-of-speech tagging. *The Oxford handbook of computational linguistics,* , pp. 219-232.

YANG, M., TU, W., LU, Z., YIN, W. and CHOW, K., 2015. LCCT: a semisupervised model for sentiment classification, *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL* 2015, Association for Computational Linguistics (ACL).