SENTIMENT ANALYSIS IN TURKISH

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

UMUT EROĞUL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

JUNE 2009

Approval of the thesis:

SENTIMENT ANALYSIS IN TURKISH

submitted by **UMUT EROĞUL** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen
Dean, Graduate School of **Natural and Applied Sciences** _______________

Prof. Dr. Müslim Bozyiğit
Head of Department, **Computer Engineering** _______________

Dr. Meltem Turhan Yöndem
Supervisor, **Computer Engineering Dept.** _______________

Examining Committee Members:

Prof. Dr. Göktürk Üçoluk
Computer Engineering Dept., METU _______________

Dr. Meltem Turhan Yöndem
Computer Engineering Dept., METU _______________

Asst. Prof. Dr. Pınar Şenkul
Computer Engineering Dept., METU _______________

Dr. Onur Tolga Şehitoğlu
Computer Engineering Dept., METU _______________

Güven Fidan, M.Sc.
AGMLAB _______________

Date: _______________

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: UMUT EROĞUL

Signature :

ABSTRACT

SENTIMENT ANALYSIS IN TURKISH

Eroğul, Umut
M.S., Department of Computer Engineering
Supervisor : Dr. Meltem Turhan Yöndem

June 2009, 55 pages

Sentiment analysis is the automatic classification of a text, trying to determine the attitude of the writer with respect to a specific topic. The attitude may be either their judgment or evaluation, their feelings or the intended emotional communication.

The recent increase in the use of review sites and blogs, has made a great amount of subjective data available. Nowadays, it is nearly impossible to manually process all the relevant data available, and as a consequence, the importance given to the automatic classification of unformatted data, has increased.

Up to date, all of the research carried on sentiment analysis was focused on English language. In this thesis, two Turkish datasets tagged with sentiment information is introduced and existing methods for English are applied on these datasets. This thesis also suggests new methods for Turkish sentiment analysis.

Keywords: Natural Language Processing, Machine Learning, Text Mining, Sentiment Analysis, Turkish

ÖZ

TÜRKÇE METİNLERDE DÜŞÜNCE ÇÖZÜMLEME

Eroğul, Umut
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi : Dr. Meltem Turhan Yöndem

Haziran 2009, 55 sayfa

Sentiment analiz (düşünce çözümleme) bir dökümanın bilgisayar tarafından incelenip, o dökümanın konu hakkında (konudan bağımsız olarak) olumlu veya olumsuz görüş belirttiğini saptamaya çalışır. Çıkartılmaya çalışılan görüş yazarın konu hakkındaki kararı ya da değerlendirmesi, yazarken hissettiği ruh hali, ya da belirtmek istediği etki olabilir.

Günümüzde internet kullanımının artması sonucu daha da yaygınlaşan tartışma grupları ve görüş - eleştiri sitelerinde yer alan fikir belirten yazıların artması, bu verilerin elle işlenmesini imkansıza yakın hale getirmiş ve otomatik sınıflandırmanın önemini daha da arttırmıştır.

Günümüze kadar düşünce çözümleme araştırmaları, genelde İngilizce üzerine yoğunlaşmıştır. Bu tez kapsamında düşünce çözümlemede kullanılabilecek iki yeni veri seti oluşturulmuş, ve daha önceden İngilizce'de yapılmış çalışmalarda denenen yöntemlerin Türkçe'de gösterdikleri başarılar incelenmiş ve Türkçe'ye özel yeni yöntemler önerilmistir.

Anahtar Kelimeler: Doğal Dil İşleme, Makine Öğrenimi, Metin Madenciliği, Düşünce Çözümleme, Türkçe

To my family.

ACKNOWLEDGMENTS

I would like to thank my supervisor Dr. Meltem Turhan Yöndem for her guidance for the last two years.

I would like to thank Dr. Onur Tolga Şehitoğlu and Güven Fidan, for their help and ideas.

I would like to thank my brother Can Eroğul, for his advices and encouragements through my education.

I would like to thank Atıl İşçen for his cooperation for the last four years.

I would like to thank my aunt Dicle Eroğul and Ruken Çakıcı for their corrections and suggestions on this thesis.

Finally, I would like to thank my parents and loved ones for their support.

TABLE OF CONTENTS

CHAPTERS

APPENDICES

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

1.1 Motivation

“What other people think” has always been an important piece of information during decision making, even pre-Web; asking friends who they plan to vote for, requesting letters of recommendation from colleagues, checking Consumer Reports regarding dishwasher brands.

People search for, are affected by, and post online ratings and reviews. Recent studies [8] [16] shows that 60% of US residents have done online research on a product at least once, and 15% do so on a typical day. On the other hand; 73%-87% of US readers of online reviews of services reported that the reviews had a significant influence on their purchase. But, 58% of US internet users report that online information was missing, impossible to find, confusing, and/or overwhelming.

Terabytes of new user-generated content appears on the Web every day. Many of these blog posts, tweets, articles, podcasts and videos presents valuable opinions about a product, a company, a movie, etc.. The quantity of information available is immense, which made it impossible to track manually, thus tools are needed to automatically analyze these comments.

This vast amount of information is searchable with words or word phrases, but in order to perform better results, search engines should understand the meaning and semantics hidden in the text. One of the approaches to this problem is sentiment analysis, whose main concern is the extraction of feelings in a given text.

Sadly the sentiment analysis studies mostly concentrated on English so far, and to the best of our knowledge, there is no study made in Turkish. We wanted to see the effects of the methods applied to English, in a Turkish data. Unfortunately, there was no data available for sentiment analysis in Turkish. So in this thesis, we introduce a new dataset tagged with sentiment information, and performed experiments on this data.

The aim of this thesis is to initiate sentiment analysis study in Turkish; by creating a Turkish dataset, tagged with the sentiment information and conducting the initial experiments on this data, by using the methods that are already applied in English sentiment analysis.

This thesis also aims to find new methods, which would increase the performance using language dependent features, and analyze the effect of linguistic features on Turkish sentiment analysis.

1.2 What is Sentiment?

Sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic. The attitude may be the judgment or evaluation, the emotional state of the author when writing or the emotional effect the author wishes to have on the reader. Computers can perform automated sentiment analysis of digital texts, using elements from machine learning such as latent semantic analysis, support vector machines, ”bag of words” and Semantic Orientation.

There are two main approaches in sentiment analysis, statistical and linguistic. Statistical approach relies heavily on mathematical and statistical comparison of the occurrences and number of negative or positive statements in the text, whereas linguistic approach tries to build a set of rules and compare the analyzed text with them.

As Devitt et Al. (2007) [11] stated , sentiment analysis in computational linguistics has focused on examining what textual features (lexical, syntactic, punctuation, etc) contribute to affective content of text and how these features can be detected automatically to derive a sentiment metric for a word, sentence or whole text.

As pang/Lee (2004) [26] stated, sentiment analysis seeks to identify the viewpoint(s) underlying a text span; an example application is classifying a movie review as ”positive” or ”negative”. It has proven useful for companies, recommender systems, and editorial sites to create summaries of people’s experiences and opinions that consists of subjective expressions extracted from reviews.

Sentiment analysis is tightly coupled with Opinion Mining, which is the term used by commercial applications that apply sentiment analysis methods. Formally Opinion Mining is about automatically getting to know what people think on an item (brand, product, anything) from their explicit opinions in blogs, comments to blog posts and news stories, Social Networks, elsewhere in the Web.

1.3 Importance of Turkish from a Linguistic Perspective

Turkish is highly-inflected and word order free. It is an agglutinating language, which means that words are formed with linear concatenation of suffixes.

In an agglutinating language like Turkish, a single word can be a sentence with tense, agreement, polarity, and voice as in Table 1.1 and translate into a relatively longer English sentence. Morphological structure of the words bears clues about part-of-speech tags, modality, tense, person and number agreement, case, voice and so on [6].

Two example sentences in Turkish are given with their translations in Table 1.1. We think it is a good example of the importance and use of suffixes in Turkish. We can see that a lot of meaning can be given to a word by extending with appropriate suffixes.

Table 1.1: Two Example Sentences in Turkish with Translations

Gidemeyebilirdim go-Abil-Neg-Possib-Aor-Past-P1sg I might not have been able to go
Oturtmalısın sit-Caus-Oblig-P2sg (You) must make (someone) sit.

1.4 Outline of the Thesis

The organization of the thesis is as follows:

In Chapter 2, a survey of related work on sentiment analysis is given. Chapter 3 gives detailed background information about the fields and methods, used in this thesis. In Chapter 4, the data used for the experiments and the methods used are described in detail. Chapter 5 gives the experimental results and the discussions about the results. Finally, Chapter 6 summarizes the thesis and gives the conclusions. Possible improvement ideas to the applied methods are given in the future work section.

CHAPTER 2

LITERATURE SURVEY

Polarity classification, also known as opinion mining, is a subfield of sentiment analysis, where the task is to classify an opinionated document as either positive or negative according to the sentiment expressed by the author, which has gained a lot of interest in the past years.

Early studies on sentiment-based categorization started with the use of models inspired by cognitive linguistics [15] [30], or the manual construction of discriminant-word lexicons [9]. Das et Al. introduced real-time sentiment extraction, trying to automatically label each finance message in online stock message boards as a “buy”, “sell” or “neutral” recommendation. But their method requires creating a discriminant-word lexicon by manual selection and tagging.

One of the first major research in sentiment analysis with supervised learning approach is the ”Thumbs up? Sentiment Classification using Machine Learning Techniques” by Pang et Al. [28] conducted in 2002; where they have established a dataset from the movie reviews collected from the well-known internet movie database (www.imdb.com); . They have automatically classified 1400 reviews (700 positive, 700 negative) where the author clearly expressed his opinions with such phrases like (9 out of 10, 4 out of 5, etc..). They cleaned the reviews from such remarks, and performed various tests on them. Their main objective was to automatically determine the sentiment. They had tried several features (part of speech information, position information) for the bag-of-words approach, where the document is represented as an unordered collection of words that serves as an input to machine learning methods, which will be explained in section 3.3. They tried three machine learning methods (Naive Bayes, Maximum Entropy, Support Vector Machines). Their results show that using only uni-grams and bi-grams in the bag-of-words approach, which will be explained in section 3.3.1 generates the best results (82.9%) with the support vector machines, which will be explained in section 3.1.1.

Dave et Al. [10] also took the supervised learning approach and experimented with multiple datasets of product reviews on amazon.com. They examined an extensive array of more sophisticated features based on linguistic knowledge or heuristics, coupled with various feature-selection and smoothing techniques. In conclusion they state the general problems of taking online data, some of which can be states as, users not understanding the rating system clearly and giving low ratings for the things they like, or giving only the negative comments and conclude with a final sentence stating that the overall performance is satisfactory. They also state that the average length of product reviews are too short and the positive reviews are dominant in online sites.

Subjective - Objective Classification

After analyzing the results, researchers concluded that classifying between objective and subjective sentences could improve the performance on sentiment analysis. These studies [26] [32], trained another classifier for the classification of the subjective-objective sentences which filters subjective sentences from the sentiment classifier. One of the first datasets created for the classification of subjectivity is “subjectivity dataset v1.0 [5]” created by the Bo Pang and Lillian Lee, introduced in Pang/Lee (2004)[26] which includes 5000 subjective movie review snippets, and 5000 movie plot sentences which are assumed to be objective. In the experiments, they used a two layer classifying algorithm, the first layer of classifiers classified between the subjective and objective sentences, the sentences which are classified as subjective are processed in the sentiment classifier, which increased the overall result by 4% (86.9%). In the subjectivity classifying, they used the subjectivity scores of the neighboring sentences, and graph algorithms which is outside the scope of this thesis.

Pang et Al. (2005) [27] aimed to extend the current problem by trying to find the correct rate given to a movie (1-5) from the reviews made, rather than just assigning positive or negative. This research field has focused on improving multi-class labeling techniques such as one-vs-all, one-vs-one and regression. Since there is no current data for the sentiment scale classification, they have created a new dataset for this purpose. They achieve 70% performance on average with 3-class classification, and 50% performance on 4-class classification. The best results are achieved in one-vs-all by support vector machine models. Also in this study a positive-negative classifier is trained and the positive-sentence-percentages of each review in a rating group is analyzed, it can be said that the positive sentence percentage in a review is directly proportional to the rate it received.

Automatic Data Generation

Some researchers tackle the need for annotated data in supervised learning by automatically assigning the effects of user comments on the sale price in amazon marketplace [14] or by analyzing the stock prices of a given company with the news published about the company [11]. They trained on the sales made on the marketplace within the last year, and tried to guess which seller will sell the same item first. They determined a set of key phrases important for the overall rating like 'delivery', 'packaging', 'service', etc; for each merchant, and created the feature vector under the assumption that each of these attributes are expressed by a noun, noun phrase, verb, or a verb phrase in the feedback postings. They assigned a dollar value to each noun phrase representing the effect of the phrase on the sale price, which performed 87% on guessing which merchant will sell the item.

Using the natural language processing tools available for English on other languages is also an interesting research area for the researchers [23] in different language domains. A previously translated corpus, and a lexicon is used for sentiment classification in Romanian. This shows that the cross-lingual transfer performance is mainly based on the amout of multilingual data available for training.

Domain Transfer

One interesting problem arising from the domain-specific nature of sentiment analysis in the context of supervised learning is how to adapt the classifier to a new domain. In domain transfer, the classifier is trained in a domain, and tested on another domain [31]. The main purpose is to find the properties of domain independent classifiers [2]. Blitzer et Al. (2007) [4] showed that domain transfer in the product reviews gathered from Amazon.com performed succesfully. Models trained with the reviews about the dvd products, had performed (79.5%) as good as models trained with reviews about books (80.5%) when tested with the book reviews.

Feature Selection in Sentiment Analysis

The selection of the feature set is a major part of the machine learning process. Minimizing the feature set without losing critical information is another problem, which the machine learning and the sentiment analysis community tackles. Eliminating features that are subsumed by other features [29] could minimize the feature space by 98%, while increasing the performance by two percent. The importance of the feature set for machine learning is described in Section 3.3.2. Various other types of features like pattern extraction for value phrases, Turney values, Osgood semantic differentiation with WordNet values and feature selection methods like subset selection and feature subsumption have been analyzed [24] [3] [21] [13].

The information loss created by the bag-of-words approach is also a problem that the sentiment analysis community tackles. Keeping the document in a tree structure and applying the sentiment results to neighboring sentences is shown [22] to have a small positive effect on the performance, but even using N-grams with bag-of-words in sentence level is not able to prevent information loss. After the features are known, Airoldi et Al.(2006) [1] converted the feature set to a graph, and used markov blanket method to merge features, in order ro refine the dataset. Wilson et Al. (2004) represented the data in a tree structure, and used features representing the relations in the tree, with boosting and rule based methods.

The use of Neutral Data

Most of the research in the sentiment analysis domain is focused on positive-negative classification because the neutral documents decrease the performance drastically. Koppel et Al. (2006) [19] shows that using neutral data in training improves the performance. In this study, rather than training one classifier for the positive-negative classification, the authors trained two classifiers for positive-neutral and neutral-negative classification. The interesting result is that, if the results of these two classifiers are analyzed correctly; the overall classification performance on positive-negative classification is increased by using neutral data in the training.

CHAPTER 3

BACKGROUND

In this chapter, the background information is given for the topics relevant to the thesis study. The major related topics are listed as Natural Language Processing (NLP) and Machine Learning (ML). NLP is used for finding the roots of the words via morphological analysis, and determining the part of speech. The main machine learning tool used in this work is Support Vector Machines (SVM).

3.1 Machine Learning

Machine learning (ML) is the subfield of artificial intelligence that is concerned with the design and development of algorithms that allow computers to improve their performance over time, based on data. In addition to using data mining and statistics, ML also uses computational and statistical methods in order to collect a variety of information automatically from raw data. The use of machine learning in natural language processing has peaked in recent years, due to its high performance on various studies.

In this thesis, we use Support Vector Machines as a machine learning technique, which has been shown to be effective in previous text mining studies.

3.1.1 Support Vector Machines

In general Support Vector Machines (SVMs) get a feature vector as input, which expresses the relevant data with a set of real values. Representing the data with a set of real values can be a hard task depending on the data and the domain. A common approach in text mining is the bag-of-words approach which will be explained in section 3.3.

Given that, data points each belong to one of two classes, the goal of SVMs are to decide which class a new data point will be in. A data point is viewed as a p-dimensional vector (feature vector (a list of p numbers)), and a SVM tries to find whether it can separate such points with a p - 1-dimensional hyperplane, which is called a linear classifier. There may be many hyperplanes that might classify the data, however, SVMs try to find the hyperplane with the maximum separation between the two classes. That is to say that the nearest distance between a point in one separated hyperplane and a point in the other separated hyperplane is maximized.

Formally a data point can be represented with Equation 3.1a, and a hyperplane can be written as the set of points x satisfying Equation 3.1b, where $\cdot$ denotes the dot product. The vector w is a normal vector, and it is perpendicular to the hyperplane. SVM selects the w and b to maximize the margin, or distance between the parallel hyperplanes that are as far apart as possible, while still separating the data. This is illustrated in Figure 3.1.

$$D = \{(x_i, c_i) | x_i \in R^p, c_i \in \{-1, 1\}\}_{i=1}^{n} \tag{3.1a}$$

$$\mathbf{w} \cdot \mathbf{x} - b = 0, \tag{3.1b}$$

After receiving the input, machine learning techniques try to find a relation between given features and the outcome. The complexity of these relations can vary from linear (feature#n has a 0.15 positive effect on the outcome, etc..), to quadratic or more.

We used Joachim's SVMperf package [18] for training and testing because of its speed and performance on large datasets. SVMperf is an implementation of the Support Vector Machines formulation for optimizing multivariate performance measures described in [18].

3.1.2 Automated Data Gathering

The Machine Learning methods have a strong dependency on the quality and quantity of the data. Supervised learning methods' need for large ammount of data is usually answered by automatic collection of tagged data from the Internet. There a lot of web sites that have ratings associated with the polarity of a given text. For instance, stars associated with "Internet Movie

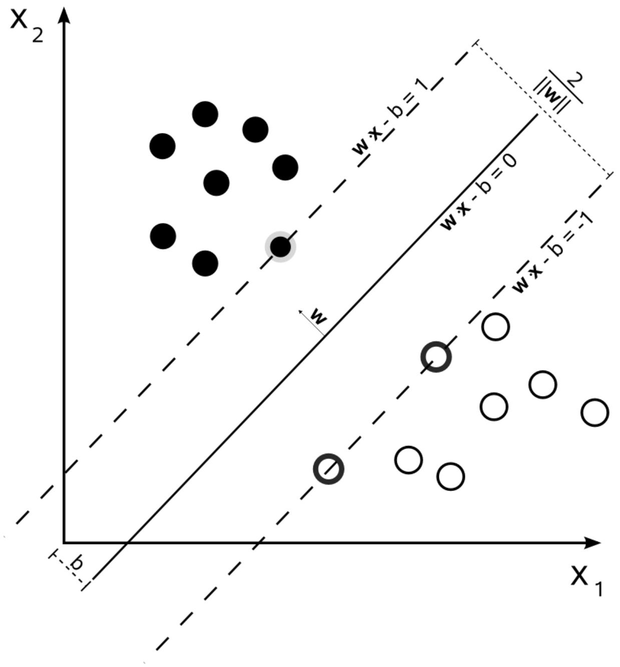

X_2
$\mathbf{w} \cdot \mathbf{x} - b = 1$
$\frac{2}{\|\mathbf{w}\|}$
$\mathbf{w} \cdot \mathbf{x} - b = 0$
$\mathbf{w} \cdot \mathbf{x} - b = -1$
$\mathbf{w}$
b
X_1

Figure 3.1: SVM Hyperplane

DB comments" hint at polarity. For a user that gave the movie 8 stars out of 10: it is likely that his sentiments captured in the text were generally positive and moderately forcefully held. It's not surprising that a 5/10 review has the title, "A huge disappointment for fans of this memorable series" and 10/10 is coupled with "I just LOVED IT!". Similarly, a hotel guest who chose a Fair rating in a satisfaction survey is likely to have posted more complaints than praise in free-text response fields.

Most of the datasets used in this field are created by gathering all the reviews and their ratings from web sites. If the site does not give rating information for a review, automated programs try to extract the rating information, from phrases in the review like "* OUT OF 10" or "*/10".

3.2 Natural Language Processing

Natural Language Processing (NLP) is an area of computer science where the computer tries to process naturally occurring human language, therefore NLP researchers try to understand the underlying patterns used in human language and extract these patterns for the use of computers. NLP tackles a wide range of problems from speech segmentation, to syntactic ambiguity, from part-of-speech tagging to word sense disambiguation. Sentiment analysis (opinion mining) is a sub task of information extraction, whose goal is to automatically extract structured information from unstructured text. In sentiment analysis, the structured information extracted is the general sentiment (opinion) the author tries to give in the document. This field has gained more importance over the years because of the growing amount of unstructured text available on the Internet. With the increasing use of the Internet, the number of persons expressing their opinions in writing, has increased drastically. The amount of data available made it nearly impossible for manual information extraction to process this data, thus increased the importance of automatic information extraction.

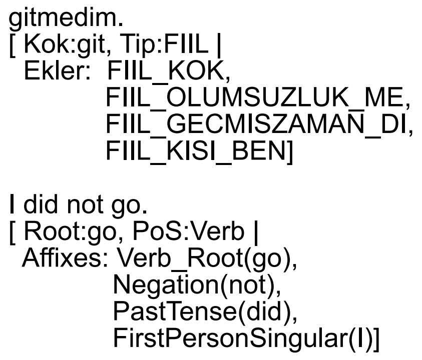
Chowdhury [7] claimed that every NLP task is built according to the natural language understanding issue. There are three main problems for understanding the natural language in computer programs. The first problem is about the thought process, the second one is the representation and meaning of the linguistic input, and the third one relates to the word knowledge. Therefore, from the beginning to the end, every NLP system should start studying at the word level, then move on to sentence, and finally conclude to the context level on the whole domain. The reason is, first the morphological structure, nature (such as part-of-speech, meaning etc.) of the word has to be determined and then the word order, grammar, meaning of the entire sentence and in conclusion the domain knowledge, should be determined. There can be a specific meaning or connotation for each word and/or sentence in a given context or domain, and may have relations with many other words and/or sentences in the given context.

Statistical NLP, which coupled with machine learning and data mining; uses stochastic, probabilistic and statistical methods to resolve some difficulties faced using standard approaches.

3.2.1 Part of Speech

Part-of-Speech (POS) is the linguistic category, which a word belongs to. Common linguistic categories include “noun”, “verb” and “adjective” among others. Part-of-speech tagging is defined as the process of parsing each word as a token in each sentence and labeling each token with the most probable category label.

We would like to increase the performance of our system by integrating the POS knowledge



```
gitmedim.
[ Kok:git, Tip:FIIL |
  Ekler:  FIIL_KOK,
          FIIL_OLUMSUZLUK_ME,
          FIIL_GECMISZAMAN_DI,
          FIIL_KISI_BEN]

I did not go.
[ Root:go, PoS:Verb |
  Affixes: Verb_Root(go),
           Negation(not),
           PastTense(did),
           FirstPersonSingular(I)]
```

Figure 3.2: An example POS tagging by Zemberek

into the system.

There are also some complications in switching languages here. POS tagging is easier in English where more complicated NLP tools exists, whereas in Turkish, the current state of NLP tools are not as advanced as its English counterparts and are limited to morphological analysis in word level for the time being.

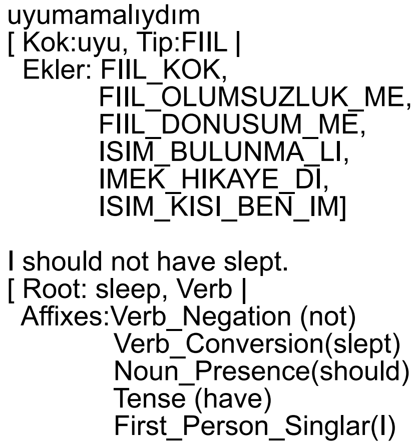
We would like to see the effect of each part-of-speech in the total sentiment of the review. So we performed experiments using only features that are classified as “verb”. We discussed those experiments in the relevant section. In this thesis, Zemberek library, a natural language library for Turkish, is used to gather the part of speech information about a word. An example for POS tagging process is given in Figure 3.2.

3.2.2 Natural Language Processing Tools

For Natural Language processing, we used two major libraries, one for each language. Selecting a library for English is an easier job due to the availability, but in Turkish we had only one option, namely Zemberek library.

3.2.2.1 Natural Language Toolkit

The Natural Language Toolkit (NLTK) [20] is a suite of open source program modules and tutorials providing ready-to-use computational linguistics written in Python. NLTK provides



```
uyumamalıydım
[ Kok:uyu, Tip:FIIL |
  Ekler: FIIL_KOK,
         FIIL_OLUMSUZLUK_ME,
         FIIL_DONUSUM_ME,
         ISIM_BULUNMA_LI,
         IMEK_HIKAYE_DI,
         ISIM_KISI_BEN_IM]

I should not have slept.
[ Root: sleep, Verb |
  Affixes:Verb_Negation (not)
          Verb_Conversion(slept)
          Noun_Presence(should)
          Tense (have)
          First_Person_Singlar(I)
```

Figure 3.3: An example morphological analysis by Zemberek

tools necessary for symbolic and statistical natural language processing, and is interfaced to annotated corpora. The abilities of this toolkit includes: corpus readers, tokenizers, stemmers, taggers, chunkers, parsers, classifiers, clusterers, estimation tools and interfaces to common databases like WordNet. WordNet is a large lexical database of English, developed under the direction of George A. Miller. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations.

The NLTK library is mainly used in part of speech taggers and stemmers in English. The PoS tagger is automatically generated from the tagged Brown corpus with the help of standart learning agents both of which is available in the toolkit. It is also used for its functions in tokenizing, which is basically the separation of the words within the input.

3.2.2.2 Zemberek

The main NLP tool we used in Turkish is the Zemberek library [17] which is used for morphological analysis, and spellchecking purposes. With this library we are able to analyse each word in Turkish morphologically. An example is given in Figure 3.3.

3.3 Combining ML and NLP

In text mining, a major problem is that machine learning methods work with real numbers, while our data is in text format. To convert the text into real values, most common approach taken in this field is deciding on a set of features, and giving real values for those features according to the input. Main challenges in this approach is selecting the features. Since our data is represented with these features, each feature should contain significant information. These features should cover the domain, to decrease the data loss during the transformation. On the other side, the number of features should be small to ease the learning process, by decreasing noise.

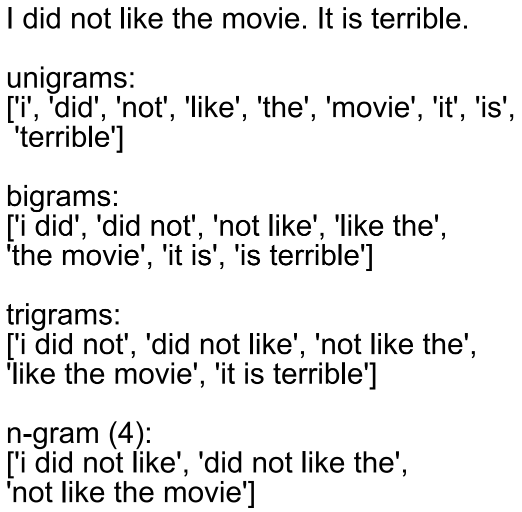
The methods used to tackle these problems are described in this section.

3.3.1 N-gram model

One of the methods of converting text into features is the bag-of-words approach, where each word is an element of the feature set, and each document is represented with a set of real numbers where each element of the set represents the frequency of that word in the document.

N-gram model can be formally defined as;
Let $f1, ..., fm$ be a predefined set of m features that can appear in a document; examples include the word “still” or the bigram “really stinks”. Let $ni(d)$ be the number of times fi occurs in document d. Then, each document d is represented by the document vector $d := (n1(d), n2(d), ..., nm(d))$ [28].

In bag of words model, a text is represented as an unordered collection of words, disregarding grammar and even word order, which causes significant data to be lost in the processing of the input. To decrease the data loss, N-gram model is used, where rather than taking each word as a feature, each sub-sequence of length n is set as a feature. This model, while increasing the complexity of the process, cannot account for long dependencies but preserves word order within N words. The N-gram sets of an example sentence is given in Figure 3.4. An N-gram of size 1 is referred to as a ”unigram”; size 2 is a ”bigram”; size 3 is a ”trigram”; and size 4 or more is simply called an ”N-gram”.



```
I did not like the movie. It is terrible.

unigrams:
['i', 'did', 'not', 'like', 'the', 'movie', 'it', 'is',
 'terrible']

bigrams:
['i did', 'did not', 'not like', 'like the',
'the movie', 'it is', 'is terrible']

trigrams:
['i did not', 'did not like', 'not like the',
'like the movie', 'it is terrible']

n-gram (4):
['i did not like', 'did not like the',
'not like the movie']
```

Figure 3.4: N-gram sets of an example sentence

3.3.2 Feature Selection

Creating the feature set for the representation of input is a hard task, because the ideal feature set should be small in size but rich in information. One of the advantages of feature selection is the prevention overfitting, which is fitting a statistical model that has too many parameters.

Feature selection algorithms can be divided into two categories; Feature Ranking and Subset Selection. Feature Ranking ranks the features by a metric and eliminates all features that do not achieve an adequate score. Subset Selection searches the set of possible features for the optimal subset.

Most common techniques to refine the feature set are described below.

3.3.2.1 Threshold

Thresholding is the simplest feature elimination method used in statistical natural language processing. It is the most common way of combining “feature ranking” with the “bag-of-words” approach. It rates each feature by the number of times that word is used in the training data, and eliminates all features that do not meet a certain score. In our case, it prevents an uncommon word (if a word is not used t times in the training set) from being a feature.

3.3.2.2 Negation

Since the bag-of-words approach disregards the word order, there is a significant data loss, in the use of negating words like “not”, especially when we do not know which word or word phrase it refers to.

When considering only the words in a sentence, processing the sentence “I do not like this movie,” could result with a positive classification meaning that this author loved the movie, when the word “like” is processed. A solution for this in English is to tag each word after the negation until the first punctuation. The previous sentence will then become: “I don’t NOT_like NOT_this NOT_movie”.

This approach in English was used by many studies with an increase in performance [28] [26] [12] , whereas Dave et Al. [10] showed that the negation tagging gives a slight decrease in performance, and note that simple substrings (N-grams) work better at capturing negation phrases.

Table 3.1: An Example Sentence Translation

Gitmedim	Git	-me	-di	-m
I did not go	go	not	did	I
Sevmedim	Sev	-me	-di	-m
I did not like	like	not	did	I

In Turkish the negation is usually handled morphologically. A negation suffix (-me, -ma) is attached to the verbs as shown in Table 3.1. Another option is to use the word “değil” (which can be translated to “not” in English) which effects the words prior to its use in a sentence. These different forms are implemented and tested, the results are discussed in Chapter 5.

CHAPTER 4

METHODS

In section 4.1 we will give information about the data used. In section 4.2 we will describe the methods we used to increase the performance over the data. We will give detailed results in Chapter 5.

4.1 Data Description

The work described in this research is based on three datasets, two of them is widely used in English, and the third is the one we created for Turkish. All these datasets are gathered from the movie review domain. This domain is selected because of the large online communities built around certain sites, creating a valuable information for researchers. One of the benefits of such sites is that, users choose certain icons (emoticons) for the review they are writing, which can be seen as manual tagging from the owner of the review.

It should be stated that the methods described below are not domain specific and can be tested on other domains easily given enough tagged data in the relevant domain. A set of positive and negative examples for each dataset are given in the Appendix A.

4.1.1 English Subjectivity Data

The widely used polarity datasets in English for sentiment analysis are described in this section.

4.1.1.1 Polarity 2.0

Polarity 2.0 is introduced by Pang/Lee(2004) [26]. This dataset is created by using the June 2004 dated movie reviews extracted from “rec.arts.movies.reviews” newsgroup in Internet Movie Database (IMDb), which can be accessed from [5] under polarity v2.0. The data consists of 1000 positive and 1000 negative movie reviews. The baseline for this dateset is 85% as described in [27]

The data is automatically gathered from the site and tagged if there is a definite description of rating indicators in the data like; ”8/10”, ”four out of five”, ”OUT OF ****: ***” and similar rating indications. After collecting the data, they have set thresholds for a document to be positively tagged, like “three-and-a-half stars and up in a 5 star rating are considered positive” and automatically tag the content.

One positive and one negative movie review example from Polarity 2.0 can be found in Appendix A.1.

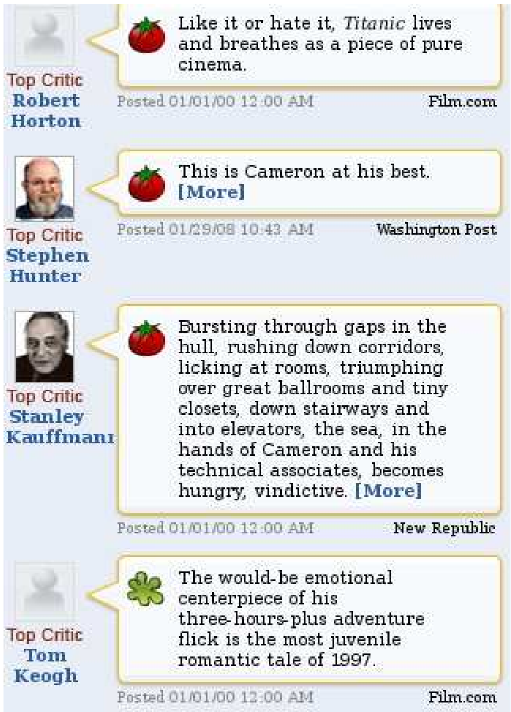
Table 4.1: General Statistics on Polarity Data 2.0

Rating	Number of Reviews	Number of Words	Words/Review
Positive	1000	787050	787
Negative	1000	705613	705

4.1.1.2 Sentence Polarity v1.0

In 2005, Pang et Al. [27] introduced a new dataset, created with the data gathered from the reviews taking place at “rottentomatoes.com”. It consists of 5331 positive and 5331 negative sentences, reviews tagged with a “fresh tomato” icons are considered positive, and a “rotten tomato” icon is considered a negative review.

A screenshot from “rottentomatoes.com” is given in 4.1, the website is famous for the reviews made by the top critics, who usually works for newspapers or magazines. Since most of the reviews are published magazines, there is nearly no typos; in other words, the site has reviews of high editorial quality.



Top Critic
Robert Horton

Like it or hate it, *Titanic* lives and breathes as a piece of pure cinema.

Posted 01/01/00 12:00 AM Film.com

Top Critic
Stephen Hunter

This is Cameron at his best. **[More]**

Posted 01/29/08 10:43 AM Washington Post

Top Critic
Stanley Kauffmann

Bursting through gaps in the hull, rushing down corridors, licking at rooms, triumphing over great ballrooms and tiny closets, down stairways and into elevators, the sea, in the hands of Cameron and his technical associates, becomes hungry, vindictive. **[More]**

Posted 01/01/00 12:00 AM New Republic

Top Critic
Tom Keogh

The would-be emotional centerpiece of his three-hours-plus adventure flick is the most juvenile romantic tale of 1997.

Posted 01/01/00 12:00 AM Film.com

Figure 4.1: A screenshot from rottentomatoes.com

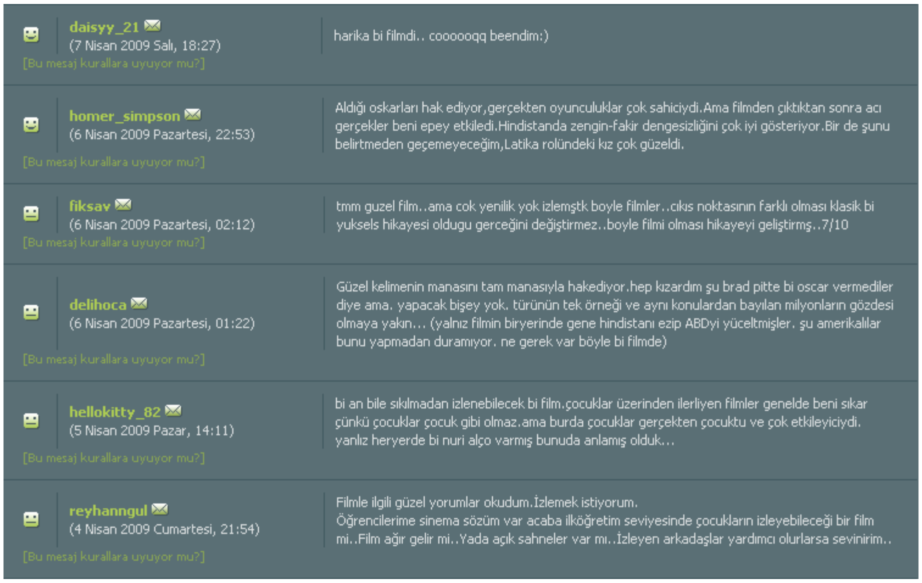
One positive and one negative movie review example from Sentence Polarity 1.0 can be found in Appendix A.2.

Table 4.2: General Statistics on Sentence Polarity Dataset v1.0

Rating	Number of Reviews	Number of Words	Words/Review
Positive	5331	112407	21
Negative	5331	111596	20

4.1.2 Turkish Polarity Data

We used Turkish Polarity data in our experiments. Due to the unavailability of tagged reviews in Turkish, we had to gather this data from the Internet. The website, namely beyazperde.com [25], is a major movie site in Turkey, which appears under "mynet.com", and is currently the



☺ **daisyy_21** ✉
(7 Nisan 2009 Salı, 18:27)
[Bu mesaj kurallara uyuyor mu?]

harika bi filmdi.. coooooqq beendim:)

☺ **homer_simpson** ✉
(6 Nisan 2009 Pazartesi, 22:53)
[Bu mesaj kurallara uyuyor mu?]

Aldığı oskarları hak ediyor,gerçekten oyunculuklar çok sahiciydi.Ama filmden çıktıktan sonra acı gerçekler beni epey etkiledi.Hindistanda zengin-fakir dengesizliğini çok iyi gösteriyor.Bir de şunu belirtmeden geçemeyeceğim,Latika rolündeki kız çok güzeldi.

😐 **fiksav** ✉
(6 Nisan 2009 Pazartesi, 02:12)
[Bu mesaj kurallara uyuyor mu?]

tmm guzel film..ama cok yenilik yok izlemştk boyle filmler..cıkıs noktasının farklı olması klasik bi yuksels hikayesi oldugu gerceğini değiştirmez..boyle filmi olması hikayeyi geliştirmş..7/10

😐 **delihoca** ✉
(6 Nisan 2009 Pazartesi, 01:22)
[Bu mesaj kurallara uyuyor mu?]

Güzel kelimenin manasını tam manasıyla hakediyor.hep kızardım şu brad pitte bi oscar vermediler diye ama. yapacak bişey yok. türünün tek örneği ve aynı konulardan bayılan milyonların gözdesi olmaya yakın... (yalnız filmin biryerinde gene hindistanı ezip ABDyi yüceltmişler. şu amerikalılar bunu yapmadan duramıyor. ne gerek var böyle bi filmde)

😐 **hellokitty_82** ✉
(5 Nisan 2009 Pazar, 14:11)
[Bu mesaj kurallara uyuyor mu?]

bi an bile sıkılmadan izlenebilecek bi film.çocuklar üzerinden ilerliyen filmler genelde beni sıkar çünkü çocuklar çocuk gibi olmaz.ama burda çocuklar gerçekten çocuktu ve çok etkileyiciydi. yanlız heryerde bi nuri alço varmış bunuda anlamış olduk...

😐 **reyhanngul** ✉
(4 Nisan 2009 Cumartesi, 21:54)
[Bu mesaj kurallara uyuyor mu?]

Filmle ilgili güzel yorumlar okudum.İzlemek istiyorum.
Öğrencilerime sinema sözüm var acaba ilköğretim seviyesinde çocukların izleyebileceği bir film mi..Film ağır gelir mi..Yada açık sahneler var mı..İzleyen arkadaşlar yardımcı olurlarsa sevinirim..

Figure 4.2: A screenshot from beyazperde.mynet.com

eighth most visited site in Turkey, according to "www.alexa.com". In this site, a user can enter a comment for each movie, stating his/her general opinion about the movie with a related icon (positive, neutral, negative). We gathered all the data from the comments section of the site and used the icons to tag the related comments. The distribution of data is given in Table 4.3.

An example movie review can be seen in Figure 4.2 and Appendix A.3. The users are asked to select an icon representing their general opinion about the movie, which can be seen at the left part of the reviews in Figure 4.2. This icon is used as the indicator of positive, negative and neutral reviews.

Table 4.3: Distribution of Reviews from "beyazperde" on Rating

Rating	Number of Reviews	Number of Words	Words/Review
Positive	117712	4233124	35
Neutral	40360	1426651	35
Negative	22102	824781	37

4.1.2.1 Filtering Process

We had to clean the data because we found out that a lot of the reviews have sentences like “my rating is 10/10”, and our support vector machine realized that 10/10 is a very good feature. We had some encoding problems due to the nature of Turkish data on the Internet, and changed the encoding to UTF-8 for all the data. We cleared rating identifiers like “%d/10” and “10/%d”. We cleared external links which generate only noise “http://www.%s” and “www.%s”. Finally we removed reviews which are smaller than 5 characters after cleaning.

For the experiments, we had randomly selected 22100 reviews from each class (positive, negative, neutral) and splitted the data in to ten parts. We used 9 parts (19890 reviews) from each class in the training process, and used the unused part (2210 reviews) in the test process. We applied ten-fold cross validation techniques in our experiments.

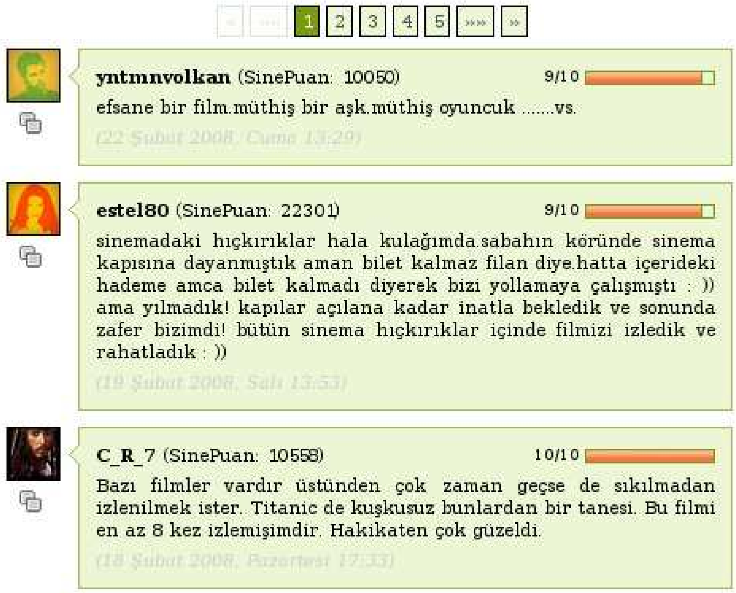
Since there are a lot of unprocessed named entities (movie names, etc), we also created a database of movie names, actors and directors; and replaced these names in the reviews with appropriate tags.

4.1.3 Turkish Polarity Scale Data

For creating a dataset that contains reviews rated from 1 to 10, we have gathered movie comments from another movie site “www.sinemalar.com”, where the users rating for a movie, together with their comments about the movie can be gathered. The distribution of data on ratings is given in Table 4.4. This dataset is gathered for the tests on multiclass labeling techniques such as one-vs-all and regression. A screenshot from the review page at “www.sinema.com” is given in Figure 4.3. The rating descriptor of each review is given in the upper right part of each review.

4.1.4 Major Differences Between Turkish and English Data

While the English polarity data has 746 words per review on average; the Turkish polarity data has only 32 words per review on average. This is a natural consequence of the different behaviour of critics writing in English and Turkish. In the English data, the people write a full movie critic besides a lot of other arguments before giving their main opinion about the



yntmnvolkan (SinePuan: 10050) 9/10

efsane bir film.müthiş bir aşk.müthiş oyuncuk ....vs.

(22 Şubat 2008, Cuma 13:29)

estel80 (SinePuan: 22301) 9/10

sinemadaki hıçkırıklar hala kulağımda.sabahın köründe sinema kapısına dayanmıştık aman bilet kalmaz filan diye.hatta içerideki hademe amca bilet kalmadı diyerek bizi yollamaya çalışmıştı :)) ama yılmadık! kapılar açılana kadar inatla bekledik ve sonunda zafer bizimdi! bütün sinema hıçkırıklar içinde filmizi izledik ve rahatladık :))

(19 Şubat 2008, Salı 13:53)

C_R_7 (SinePuan: 10558) 10/10

Bazı filmler vardır üstünden çok zaman geçse de sıkılmadan izlenilmek ister. Titanic de kuşkusuz bunlardan bir tanesi. Bu filmi en az 8 kez izlemişimdir. Hakikaten çok güzeldi.

(18 Şubat 2008, Pazartesi 17:33)

Figure 4.3: A screenshot from “www.sinema.com”

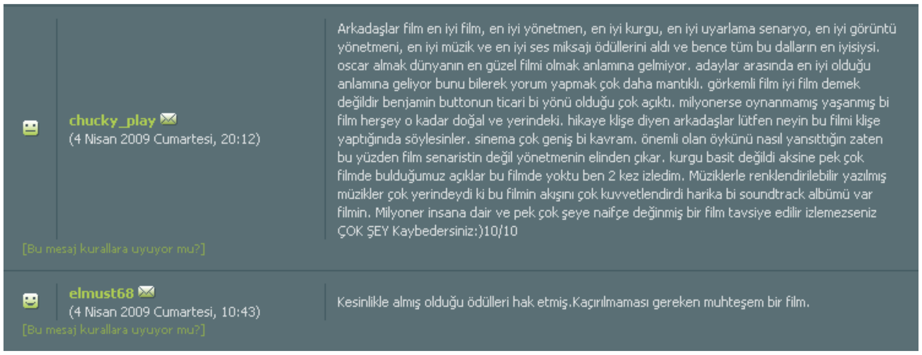
movie, making the sentiment analysis difficult. While in the Turkish data, users only express their opinions in one or two sentences on average, which makes the sentiment analysis easier.

On the other side, the English data is edited and there is nearly no syntactic or grammar errors in the reviews; while the Turkish data has a lot of Internet slang, typos, and errors. One of the main challenges we have faced, is that a lot of people do not use the complete Turkish alphabet for writing on the web sites. They substitute some Turkish letters for similar ASCII letters (ğ - g, ş - s). Moreover there are lots of reviews where the writer drops all the vowels in a word (“magnificent” becomes “mgnfcnt”) which is an ongoing trend in (SMS) Short Message Services in Turkey. These factors have drastically reduced the effect of using NLP tools.

A example of great diversity in collected reviews can be seen in Figure 4.4. The second review consists of 11 words, whereas the first review is around 130 words. Another problem in the first review, is that the writer clearly expresses his positive opinions throughout the review and summarized his main opinion with a “10/10” (which will be eliminated in the filtering process) for his review indicating his rating, but the writer selects a neutral icon indicating his opinions are neither positive or negative for his review.

Table 4.4: Distribution of Reviews at Turkish Polarity Scale Data on Rating

Rating	Number of Reviews	Number of Words	Words/Review
0	454	15020	33
1	6796	200416	29
2	2221	72491	32
3	316	10487	33
4	3036	97719	32
5	1630	55005	33
6	5268	164364	31
7	1383	44424	32
8	9923	296986	29
9	2375	72416	30
10	12599	356788	28



chucky_play
(4 Nisan 2009 Cumartesi, 20:12)

Arkadaşlar film en iyi film, en iyi yönetmen, en iyi kurgu, en iyi uyarlama senaryo, en iyi görüntü yönetmeni, en iyi müzik ve en iyi ses miksajı ödüllerini aldı ve bence tüm bu dalların en iyisiysi. oscar almak dünyanın en güzel filmi olmak anlamına gelmiyor. adaylar arasında en iyi olduğu anlamına geliyor bunu bilerek yorum yapmak çok daha mantıklı. görkemli film iyi film demek değildir benjamin buttonun ticari bi yönü olduğu çok açıktı. milyonerse oynanmamış yaşanmış bi film herşey o kadar doğal ve yerindeki. hikaye klişe diyen arkadaşlar lütfen neyin bu filmi klişe yaptığınıda söylesinler. sinema çok geniş bi kavram. önemli olan öykünü nasıl yansıttığın zaten bu yüzden film senaristin değil yönetmenin elinden çıkar. kurgu basit değildi aksine pek çok filmde bulduğumuz açıklar bu filmde yoktu ben 2 kez izledim. Müziklerle renklendirilebilir yazılmış müzikler çok yerindeydi ki bu filmin akışını çok kuvvetlendirdi harika bi soundtrack albümü var filmin. Milyoner insana dair ve pek çok şeye naifçe değinmiş bir film tavsiye edilir izlemezseniz ÇOK ŞEY Kaybedersiniz:)10/10

[Bu mesaj kurallara uyuyor mu?]

elmust68
(4 Nisan 2009 Cumartesi, 10:43)

Kesinlikle almış olduğu ödülleri hak etmiş.Kaçırılmaması gereken muhteşem bir film.

[Bu mesaj kurallara uyuyor mu?]

Figure 4.4: A screenshot from beyazperde.mynet.com

4.2 Methods Used

During each experiment the dataset is divided into two sets; the training and the testing sets. The sets are randomly divided into ten equal sized sets. Each set is selected for the testing set once, and the remaining nine sets are used for training set, therefore for each experiment result, the experiment is run ten times and the results are the averages of those runs. This process is called ten-fold cross validation testing.

At each experiment, the following steps are performed. For each review in the training dataset, a preprocessing stage is applied. The preprocessing stage starts with the spellchecking method, in which the non-recognized words are dropped from the review. The spellchecking methods are described in detail in Section 4.2.1.

The second step in the preprocessing stage is the polarity spanning process, where the negation words and suffixes are spread through the sentence, if specified by experiment specifications. It is described in detail in Section 4.2.2. The next step is, if specified by experiment specifications, stemming process of each word. The final step in preprocessing is the part-of-speech testing. If a certain part-of-speech is selected, words that are not a member of that POS are discarded.

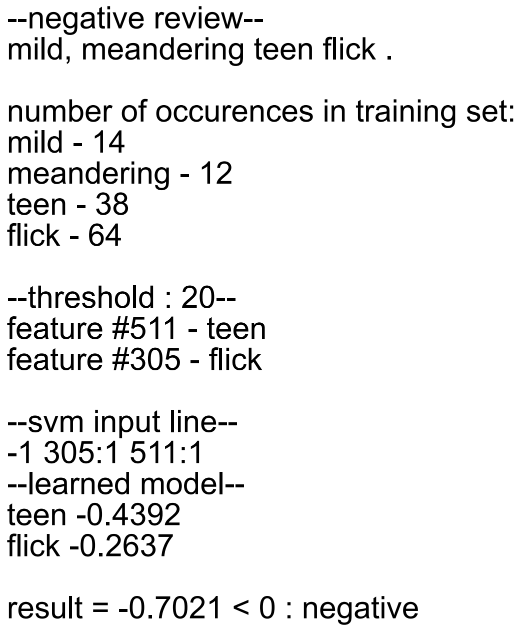
After the preprocessing stage, the feature selection stage starts. The feature selection has two parts; the creation of features and the feature elimination. For the creation of features the N-gram method is used, and each N-gram within the experiment specifications are considered a feature. The number of occurences of each feature is analyzed, and the features that do not meet a certain threshold is discarded.

After the features are selected, each review is converted to a real valued array, representing the features and its values. The presence versus frequency problem, which is described in Section 4.2.5, is addressed here. Applying these steps on an example sentence can be seen in Figure 4.5. Using a linear kernel, an SVM would classify this example correctly.

4.2.1 Spellcheck Methods

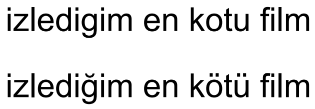
The Zemberek library has two kinds of spellchecking algorithms available to us. The first method (from ASCII to Turkish) is intended to correct the misuse of Turkish characters (üğışçö), and tries to find a Turkish word by testing each combination of replaced characters (ugisco) with their Turkish counterparts. An example spellchecking is given in Figure 4.6.

This library also allows us to perform extended spellchecking algorithms on a given string. This algorithm handles up to 3 misplaced or wrong characters in the root of the word, and 2 misplaced and wrong characters in the suffixes. An example spellchecking is given for the word “uyumamalıydım” which can be translated as “I should not have slept”, in Figure 4.7.



```
--negative review--
mild, meandering teen flick .

number of occurences in training set:
mild - 14
meandering - 12
teen - 38
flick - 64

--threshold : 20--
feature #511 - teen
feature #305 - flick

--svm input line--
-1 305:1 511:1
--learned model--
teen -0.4392
flick -0.2637

result = -0.7021 < 0 : negative
```

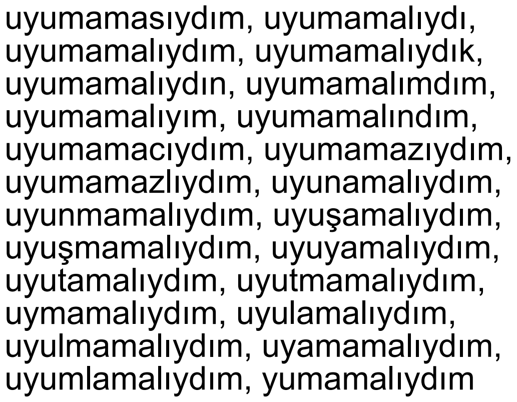
Figure 4.5: The steps taken on an example sentence



izledigim en kotu film

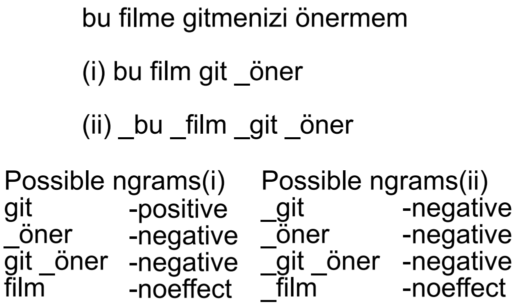
izlediğim en kötü film

Figure 4.6: An example of the first spellchecking method with Zemberek



uyumamasıydım, uyumamalıydı,
uyumamalıydım, uyumamalıydık,
uyumamalıydın, uyumamalımdım,
uyumamalıyım, uyumamalındım,
uyumamacıydım, uyumamazıydım,
uyumamazlıydım, uyunamalıydım,
uyunmamalıydım, uyuşamalıydım,
uyuşmamalıydım, uyuyamalıydım,
uyutamalıydım, uyutmamalıydım,
uymamalıydım, uyulamalıydım,
uyulmamalıydım, uyamamalıydım,
uyumlamalıydım, yumamalıydım

Figure 4.7: The suggestions of the Zemberek library for the word “uyumamalıydım”



bu filme gitmenizi önermem

(i) bu film git _öner

(ii) _bu _film _git _öner

Possible ngrams(i)		Possible ngrams(ii)	
git	-positive	_git	-negative
_öner	-negative	_öner	-negative
git _öner	-negative	_git _öner	-negative
film	-noeffect	_film	-noeffect

Figure 4.8: An example of the polarity spanning

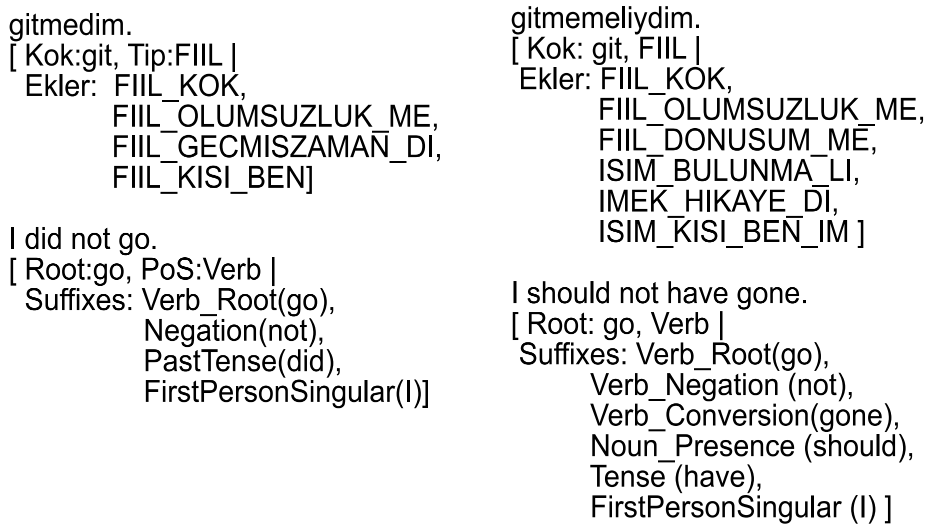
4.2.2 Polarity Spanning

In English, negation is handled with the adverb ”not”, and this is handled in previous works[28] by negating the following word (”*notbad*”− > ”*_bad*”). In Turkish, most of the negation is done by the negation suffix in the predicate, which usually affects the whole sentence. So we analyzed each sentence separately and changed the negation variable for the whole sentence, if there is a negation suffix in the predicate.

An example sentence is given in Figure 4.8. The sentence can be translated as “I do not advise you to go to this movie”. If we apply the current methods applied to English, the sentence will be processed as in (i), where the “go” verb is not affected by the preceding negation. In our approach, the sentence is processed as in (ii) where all of the words before the negation is affected, which we think will improve our performance in Turkish.

4.2.3 Using Roots of Words

Since there are a lot of combinations using the same root with different suffixes in Turkish, there are a lot of words that have the same root (Figure 4.9), but treated completely independent in our research. To create a relation between these words in our experiment framework, we applied a stemming process and discard the suffixes except the negation suffix which is explained in the previous sections. The discarding of the suffixes has created an information loss, but we hoped it would refine the feature set so that the performance will increase.



```
gitmedim.
[ Kok:git, Tip:FIIL |
  Ekler:  FIIL_KOK,
          FIIL_OLUMSUZLUK_ME,
          FIIL_GECMISZAMAN_DI,
          FIIL_KISI_BEN]

I did not go.
[ Root:go, PoS:Verb |
  Suffixes: Verb_Root(go),
            Negation(not),
            PastTense(did),
            FirstPersonSingular(I)]
```

```
gitmemeliydim.
[ Kok: git, FIIL |
 Ekler: FIIL_KOK,
        FIIL_OLUMSUZLUK_ME,
        FIIL_DONUSUM_ME,
        ISIM_BULUNMA_LI,
        IMEK_HIKAYE_DI,
        ISIM_KISI_BEN_IM ]

I should not have gone.
[ Root: go, Verb |
 Suffixes: Verb_Root(go),
        Verb_Negation (not),
        Verb_Conversion(gone),
        Noun_Presence (should),
        Tense (have),
        FirstPersonSingular (I) ]
```

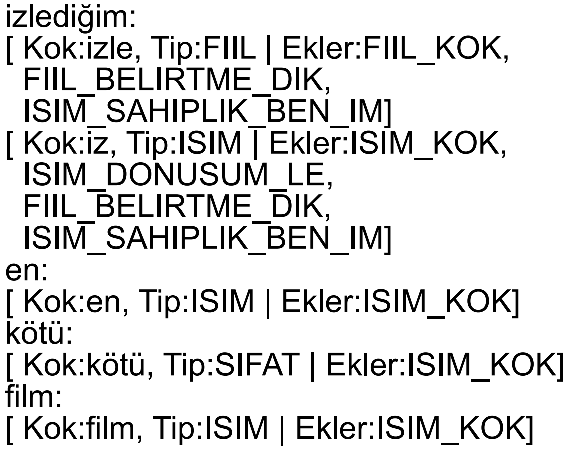
Figure 4.9: An example of words with the same root

4.2.4 Part of Speech

We also analyzed the effect of nouns, adjectives, verbs, adverbs on the sentiment of the reviews. The Zemberek library is used to get part-of-speech information of a given word. An example analysis of a sentence is given in Figure 4.10. In this experiment, words that are not part of that POS are eliminated, and the result of a certain POS is acquired in the experiments. The experiments are conducted with using only “Nouns”, “Verbs”, “Adjectives”, “Prepositions”. After analyzing the results, the experiments are continued with “Nouns+Verbs”, “Nouns+Adjectives” and “Verbs+Adjectives”, which means that words that are “Noun” or “Verbs” are used as features in “Nouns+Verbs” experiment.

4.2.5 Presence vs Frequency

One of the other methods discussed for using text based features in machine learning is the presence versus frequency question. As discussed by Pang et Al.(2002) [28], rather than giving the frequency of each feature to the SVM, giving 1 or 0 according to the presence of the feature in the input, generates better results. After seeing that the same property is acquired in our dataset, we have proceeded experiments by giving binary values for representing the presence of a feature. We have got a 70.43% accuracy on our dataset when we used the



```
izlediğim:
[ Kok:izle, Tip:FIIL | Ekler:FIIL_KOK,
  FIIL_BELIRTME_DIK,
  ISIM_SAHIPLIK_BEN_IM]
[ Kok:iz, Tip:ISIM | Ekler:ISIM_KOK,
  ISIM_DONUSUM_LE,
  FIIL_BELIRTME_DIK,
  ISIM_SAHIPLIK_BEN_IM]
en:
[ Kok:en, Tip:ISIM | Ekler:ISIM_KOK]
kötü:
[ Kok:kötü, Tip:SIFAT | Ekler:ISIM_KOK]
film:
[ Kok:film, Tip:ISIM | Ekler:ISIM_KOK]
```

Figure 4.10: Analyzing a sentence by Zemberek

frequency of words in the data, whereas the accuracy is 85.20% when we used presence of a word. We used presence of a feature throughout this thesis. We think that the performance gap is created by the SVMs ability to handle binary data.

4.2.6 Multiclass data

We also performed multiclass labeling in Turkish polarity scale data. In order to handle multiclass classifying with the use of binary classifiers, we have used “one-versus-all” and “one-versus-one” techniques. In “one-versus-all”, a classifier is trained for each label to the rest of the labels. In “one-versus-one”, a classifier is trained for each pair of labels. Selecting the correct classifier between multiple classifiers is done by a “winner-takes-all” strategy in “one-versus-all” case, where the classifier with the highest output is considered. In “one-versus-one” method, a voting mechanism is applied, where each classifier has a vote and the label that gets the maximum votes is the winner.

Another approach for the multiclass labeling is the linear regression method. In this method the classifier aims to find relationship between one or more independent variables and get the result using least square functions. The resulting function returns real values representing the labels.

CHAPTER 5

RESULTS AND DISCUSSIONS

In this chapter, the results are given and analyzed for the experiments described in Chapter 4.

In this chapter, we have used certain statistical terms to represent the results of our experiments. We have used accuracy, precision, recall and F1 values for each experiment. True Positive, True Negative, False Positive and False Negative is described in Table 5.1.

Accuracy is the degree of closeness of a measured or calculated quantity to its actual value. For our case, it's the number of examples classified correctly by our classifier, divided by the total number of examples tested, which can be calculated using Equation 5.1a.

Precision for a class is the probability that a (randomly selected) positively classified document is indeed positive. A perfect Precision score of 1.0 means that every review that was classified as positive was indeed a positive review, but says nothing about whether all positive reviews were retrieved. The precision can be calculated with Equation 5.1b.

Recall measures the proportion of actual positives which are correctly identified as such. A perfect Recall score of 1.0 means that all positive reviews were classified as positive, but says nothing about how many negative reviews were also classified as positive. The recall value can be calculated with Equation 5.1c.

The need for precision and recall can be explained with a classifier that classifies every data as negative; such a classifier will get an accuracy of 90%, precision of 0% and a recall of 0%, if the input has 1 positive and 9 negative data.

A popular measure that combines Precision and Recall is the weighted harmonic mean of precision and recall, namely the balanced F-score or the F1 measure. F1 measure can be calculated using precision and recall as shown in Equation 5.1d.

Table 5.1: Precision, Recall and F1

	Positive Example	Negative Example
Classified as Positive	True Positive	False Positive
Classified as Negative	False Negative	True Negative

$$\text{Accuracy} = \frac{TruePositive + TrueNegative}{TruePositive + FalsePositive + FalseNegative + TrueNegative} \tag{5.1a}$$

$$\text{Precision} = \frac{TruePositive}{TruePositive + FalsePositive} \tag{5.1b}$$

$$\text{Recall} = \frac{TruePositive}{|PositiveExample|} \tag{5.1c}$$

$$F = 2 \cdot (\text{precision} \cdot \text{recall})/(\text{precision} + \text{recall}) \tag{5.1d}$$

5.1 Equivalence in English Data

Pang et Al.(2004) [26] stated that the performance, F1-measure, with the polarity 2.0 dataset is around 85%. Our initial tests produced an accuracy of 84.5% on the same dataset without the use of negation handling methods. It shows that we have achieved a baseline equivalence with the previous study in English.

5.2 Threshold Experiments

We have analyzed the effect of different thresholds on the number of features and performance in Figures 5.1 & 5.2 respectively. Figure 5.1 shows that not applying a threshold, generated so many unnecessary features, while generating better results. This can be explained with words (features) that occur only two or three times overall the dataset, and this can create special cases for the SVM to exploit. We have selected a threshold of 20, which creates a feature number of 5568 and a F1-measure of 85.2% for our data. This threshold setting will be used throughout the experiments in this research.

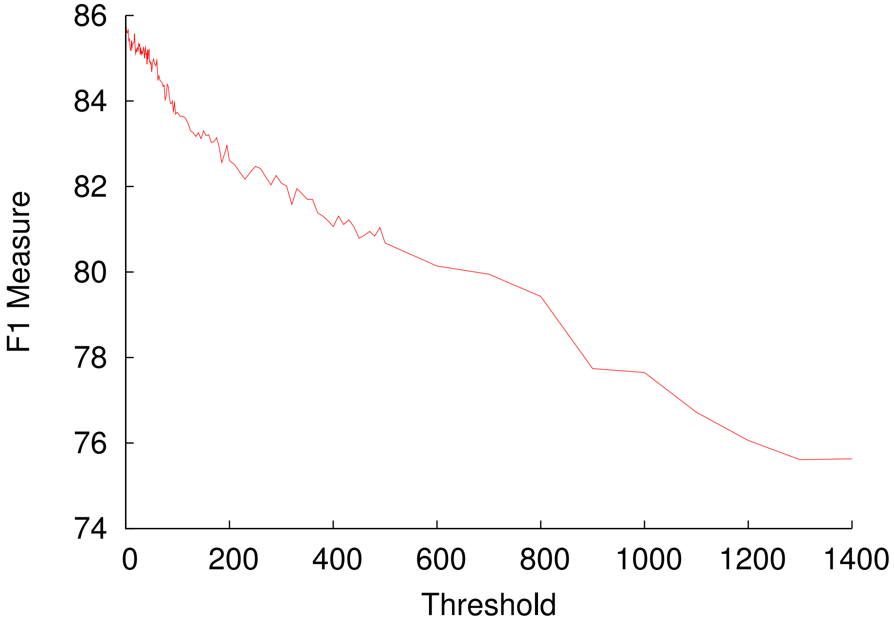
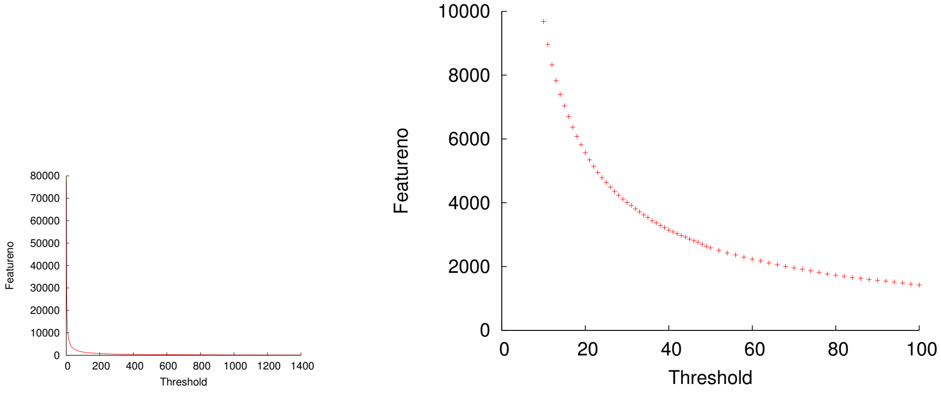


Figure 5.1: Threshold vs Performance



(a) Threshold 0-1400

(b) Threshold 0-100

Figure 5.2: Threshold vs Feature Number

5.3 Baseline

If the words that can not be recognized by the NLP libraries are not eliminated, the methods which do not use NLP tools, gain an unfair advantage over the other methods. Because the methods that needs NLP tools to function has no ability to work on those features. In order to see the effect of the methods more clearly, we eliminated words unrecognizable by our tools. We applied standard spellchecking methods that are available by the Zemberek library, in order to decrease the amount of information lost.

We applied spellchecking and elimination methods, even if we did not use any function the NLP tools offered in an experiment, for fair baseline for comparison. We ran the experiments with threshold set to 20, using only unigrams, standart spellchecking and elimination, with no NLP features to set as a baseline for other experiments on the same dataset.

The results before the elimination process and the resulting baseline with fixed threshold is shown in Table 5.2. As we can see, the performance is not directly proportional to feature number, but to the unique information each feature represents.

Table 5.2: General test results for acquiring baseline

Spellcheck	#Distinct Words	#Features	Accuracy	Precision	Recall	F1
None	131857	6605	85.11	85.69	84.30	84.99
Standard	77598	5568	85.20	86.02	84.06	85.03
Extensive	89166	6269	85.77	86.18	85.20	85.69

5.4 Using Roots of Words

One of the hypothesis we have tested is that the main opinion of a review can be expressed solely with the use of the roots of the words. To test this hypothesis, we have stemmed each word and analyzed the results. Then we have experimented with the negation affix and try to increase performance by negating the word roots in the sentence.

The results are given in Table 5.3. Looking at the results, we can see a drop in the F1-measure when we applied the negation spanning in the sentence level (from 84.99% to 83.92%). The negation spanning in word level is not tested when we do not use stemming because the negation suffix is not deleted in the process.

Eliminating all the suffixes, gives an F1-measure of 83.99%, which shows that most of the information for sentiment analysis can be captured within the roots of each word in a sentence. We did not expect such a high score, since many of the studies in Turkish parsing [6] stated that using only stems should create an information loss, which should degrade the performance. When we used the negation suffixes, there is a performance gain of 0.25% (from 83.99% to 84.24%) which can not be interpreted as a significant improvement.

Table 5.3: Results using roots of words

Stemming	Polarity	Accuracy	Precision	Recall	F1
None	None	85.11	85.69	84.30	84.99
None	Sentence	83.91	83.87	83.98	83.92
Roots	None	84.12	84.65	83.35	83.99
Roots	Only Word	84.34	84.82	83.67	84.24
Roots	Sentence	83.42	83.71	82.99	83.34

The results in Table 5.4, shows the effect of stemming more clearly. Using the roots of the words has decreased the feature set by half, while giving a similar performance. This shows us that the significant information is kept in the root, and stemming is a valid feature refining method for Turkish.

Table 5.4: Detailed feature information using roots of words

Stemming	Polarity	#Distinct Words	#Features
None	None	131857	6605
None	Sentence	93474	5983
Roots	None	8485	2698
Roots	Only Word	9320	2876
Roots	Sentence	14014	3713

5.5 Part-of-Speech

We want to see the effect of each part-of-speech on the polarity of the review, so we conducted experiments, where we tested the model with including only the certain part-of-speech tags as features . The results are shown in Table 5.5. From the results, we can see that "Noun" is the most significant part-of-speech for sentiment analysis in Turkish.

One interesting result is using only the "Verbs", gives a better F1-measure than using only the "Adjectives" (70.37% to 67.30%). Whereas when coupled with "Nouns", "Adjectives" gives a better performance than "Verbs" (83.34% to 80.54%). This result confirms the intuition that the adjective and noun couples has more information for the sentiment analysis.

We can also see that the prepositions do not have any positive effect on the performance, which can be expected.

Table 5.5: Results using part-of-speech data

Part of Speech	Accuracy	Precision	Recall	F1
Verb	73.33	79.14	63.35	70.37
Noun	77.44	79.31	74.25	76.70
Adjective	72.17	81.57	57.29	67.30
Preposition	50.84	52.89	15.34	23.78
Verb + Adjective	81.11	85.09	75.43	79.97
Verb + Noun	81.24	83.67	77.65	80.54
Noun + Adjective	83.55	84.41	82.31	83.34

Table 5.6 is given to show the distribution of parts-of-speech in our dataset.

5.6 N-gram tests

The use of N-grams had a positive effect in English, and the results of our experiments are shown in Table 5.7. We can see that using bigrams coupled with unigrams give a significant improvement over using only unigrams, whereas adding trigrams and 4-grams do not add a statistically significant improvement. When we look at Table 5.8, we can see that adding trigrams increase the feature set without adding much to performance while greatly increasing

Table 5.6: Detailed feature information using part-of-speech data

Part of Speech	#Distinct Words	#Features
Verb	38260	5156
Noun	36297	6886
Adjective	2455	494
Preposition	292	20
Verb + Adjective	40449	5641
Verb + Noun	74291	12033
Noun + Adjective	38486	7371

the complexity.

Table 5.7: Results in ngram

n-gram	Accuracy	Precision	Recall	F1
unigram	85.20	86.02	84.06	85.03
bigram	80.57	83.96	75.57	79.54
trigram	68.28	62.82	89.59	73.85
4-gram	54.71	94.83	09.95	18.02
unigram+bigram	86.20	86.90	85.25	86.07
unigram+bigram+trigram	86.15	86.85	85.20	86.02
uni+bi+tri+4-gram	86.24	86.71	85.61	86.16

5.7 Extracted Features

The SVM creates a model file, where the effect of each feature can be calculated. We have combined that information with the feature list and came up with the effect of each feature (word or word phrase) on the classification. The most powerful 20 word phrases for English and Turkish is given in the tables 5.10 and 5.9 respectively, and the most powerful 40 word phrases for English and Turkish is given in the tables A.2 and A.1 in the Appendix. In the tables, the left column has the word phrase with the most negative effect, the second column has the magnitude of the related word phrases, the fourth and the third columns have the positive word phrases respectively. This data is automatically generated from the training

Table 5.8: Detailed feature information in N-gram tests

n-gram	#Distinct Words	#Features
unigram	77598	5568
bigram	507865	4137
trigram	709510	985
4-gram	667210	161
unigram+bigram	585462	9704
unigram+bigram+trigram	1294973	10688
uni+bi+tri+4-gram	1962183	10848

data without any supervision.

It should be noted that for Turkish, 15 of 40 most powerful word phrases are “Adjectives”, and 11 of 40 word phrases are “Verbs”. The “Nouns” made the 14 of 40 most powerful word phrases. It can be said that the “Adjectives” are the most effective part-of-speech in Turkish, while their frequency in the test data is limited.

It is worth mentioning that while 28 of 20 of the most effective word phrases are not “Nouns”, the high performance of “Nouns” can be explained with the number of features available for the SVM, which can be seen in Table 5.6.

5.8 Turkish Polarity Scale Data

We also performed multiclass labeling techniques on Turkish Polarity Scale data. The results of “one-vs-all” training is given in Table 5.11. We can see that most of the classifiers were unable to extract a significant feature for an input, and 9 out of 11 classifiers return negative (meaning that the input is not in this class) for every input. This behavior can be explained with the percentage of true and false inputs in the training process of each classifier. Since we have around 400 reviews for rating 0, and around 42000 reviews for the rest of the ratings, the classifier chooses to label everything false generating an accuracy of 99%.

When we applied linear regression methods in our data, we get a Mean Square Error of 8.09. MSE measures the average of the square of the “error”, error is the amount by which the

Table 5.9: Most powerful 20 word phrases for Turkish

Negative Feature	Effect	Effect	Positive Feature
berbat	-1.35	1.03	beğendim
vasat	-1.21	1.02	mükemmel
beğenmedim	-1.18	0.94	harika
kötüydü	-1.16	0.93	süper
olmamış	-1.15	0.86	muhteşem
sıkıcı	-1.07	0.81	müthiş
değmez	-1.06	0.79	bekliyorum
kötü	-1.04	0.75	mükemmeldi
başarısız	-1.04	0.74	numara
rezalet	-1.03	0.74	eğlenceli
kaybı	-1.02	0.72	başyapıt
sıkıldım	-0.99	0.69	muhteşemdi
yazık	-0.98	0.67	değer
gereksiz	-0.95	0.66	sıkılmadan
sıkıcıydı	-0.94	0.64	mutlaka
berbattı	-0.91	0.64	harikaydı
etmiyorum	-0.88	0.64	kaçırmayın
kötüsü	-0.86	0.64	bayıldım
fiyasko	-0.85	0.63	izlenmeli
saçma	-0.83	0.62	güzeldi

estimator differs from the quantity to be estimated, which means that the difference between the classifier's rating for a given review, and the actual rating of the review is 2.84 on average.

The rating distribution of Turkish Polarity Scale data is given in Figure 5.3. A strange behavior that can be seen from the figure, is that the ratings have concentrated on a general 5 ratings, namely 1,4,6,8,10. This can be explained as, most of the people gives ratings on a 5-star scale when given a 10-star choice.

The distribution of the labels given by the linear regression method is given in Figure 5.4. The pattern can be interpreted as a Gaussian curve, and it is clear that such a pattern will not generate a good classification on this data. We have tested a classifier that gives a rating of 7 regardless of input, which generated a mean square error of 10.85.

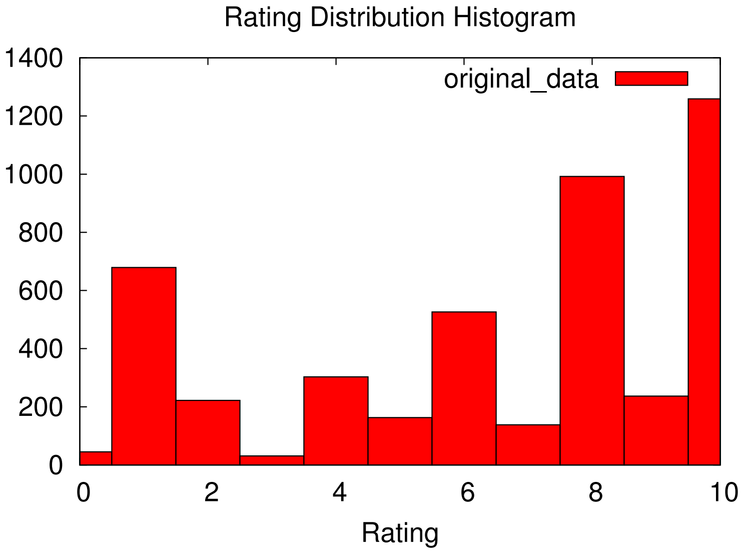


Figure 5.3: Turkish Polarity Scale Data Histogram

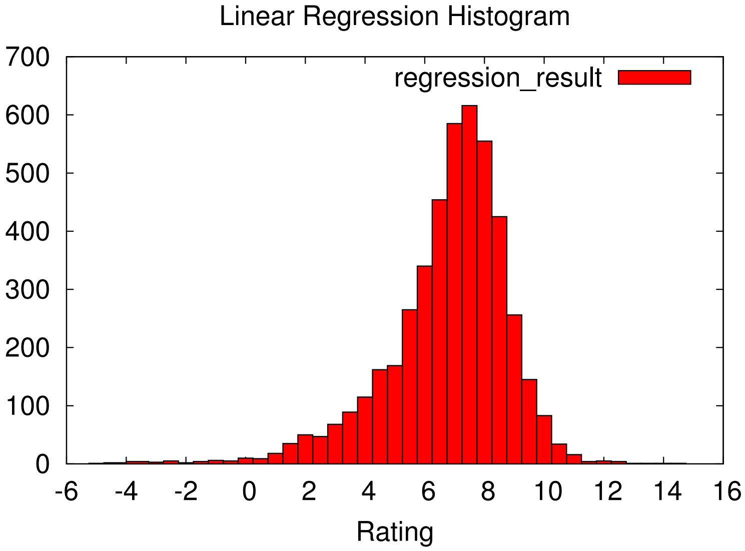


Figure 5.4: Linear Regression Histogram

Table 5.10: Most powerful 20 word phrases for English

Negative Feature	Effect	Effect	Positive Feature
dull	-1.32	1.05	unique
boring	-1.3	1.04	entertain
fails	-1.2	1.01	masterpiece
neither	-1.18	1	diverting
supposed	-1.09	1	provides
unless	-1.09	0.99	skin
too	-1.05	0.98	cinema
disappointment	-1.04	0.97	ages
tedious	-1.02	0.96	thanks
badly	-1.02	0.95	refreshing
plodding	-1	0.93	remarkable
pie	-1	0.91	wonderful
ill	-0.99	0.91	treat
flat	-0.99	0.9	gem
worst	-0.98	0.89	works
unfunny	-0.97	0.89	always
schneider	-0.96	0.89	entertaining
lack	-0.94	0.88	colorful
inept	-0.93	0.86	breathtaking
routine	-0.9	0.85	engrossing

Table 5.11: Results in one-vs-all training

One-vs-all	Accuracy	Precision	Recall	F1
0	99.00	0.00	0.00	0.00
1	85.80	64.77	8.38	14.84
2	95.16	0.00	0.00	0.00
3	99.31	0.00	0.00	0.00
4	93.40	0.00	0.00	0.00
5	96.46	0.00	0.00	0.00
6	88.56	0.00	0.00	0.00
7	96.98	0.00	0.00	0.00
8	78.39	0.00	0.00	0.00
9	94.83	0.00	0.00	0.00
10	76.35	64.77	29.76	40.78

Table 5.12: Results with linear regression

Ratings	0	1	2	3	4	5	6	7	8	9	10
True Distribution	45	679	222	31	303	163	526	138	992	237	1259
True Positives	1	42	35	8	74	37	220	86	628	99	296
Accuracy %	2	6	15	25	24	22	41	62	63	41	23

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Conclusion

This thesis aimed to initiate sentiment analysis study in Turkish. We have introduced two datasets (one for binary classification, one for multiclass classification) tagged with polarity information, and performed previously tested methods on the data. The initial results has shown that, the proven methods are applicable to Turkish.

When we applied bag-of-words method to our dataset, we have acquired a F1-measure of 85% with standard spellchecking methods.

Using only the roots of words as features for machine learning, we have achieved a F1-measure of 84%. Considering previous studies in Turkish parsing, which suggests major performance drops when using roots of words, a 1% drop in performance was better than we have expected. Using the roots of the words have reduced the feature set by 4000 features, from 6600 to 2600, while generating a F1-measure of 84%.

Creating a special case for the negation suffix in the stemming process, has increased the F1-measure by 0.3%. We were expecting a more significant increase in this method, since it should decrease the information loss in the stemming process. Our second approach for the negation issue, the polarity spanning, had a negative effect on the performance (-0.7%) in Turkish. This result is also interesting since it is shown to be effective in English.

We have seen that the N-gram methods works for Turkish. Using both unigrams and bigrams, have generated a F1-measure of 86%, which is a 1% increase in F1-measure. The use of bigrams in English did not generate a positive effect on the performance, in the movie review domain.

From the Part-of-Speech experiments we have seen that “Nouns” are the most effective (76%) Part-of-Speech for sentiment analysis. We have seen that “Nouns” are the most effective because of its high recall score. When we analyze the POS in couples, the best F1-measure is acquired with “Nouns + Adjectives”(83%).

The performance boost in “Nouns + Adjectives” can be described with “Adjectives” having high precision, and a low recall score, which means that adjectives can classify reviews better than the other POS labels, but there are lots of reviews which can not be classified due to the unavailability of adjectives in the review. When coupled with “Nouns”; this creates a good combination, because “Nouns” have the highest recall score among the Part-of-Speech.

The multiclass experiments showed that, classifying between 10 classes is a hard task, and this task cannot be handled with a single classifier which uses regression methods. The one-vs-all method created only two classifier instances that have learned anything other than a tautology. These experiments can be simplified by merging classes, which should decrease the complexity of the problem.

This thesis proposed that, the use of linguistic features should increase the performance in Turkish sentiment analysis. However this was not the case, none of the new methods proposed, have increased the overall performance. We think that this is caused by the informal use of language on the web site that we collected the data from.

The simple use of language had decreased the importance of NLP tools. In Turkish polarity dataset, there are 2.3 suffixes per word on the average, and 105 different suffix types are used. While the different suffix types are indeed a measure of the complexity of the language used, 61 suffix types are used only in 1% of the words. We think that the use of NLP features would increase the performance on a more formal writing, such as newspaper columns.

6.2 Future Work

This work can be used in an automated market research application, which should crawl certain blogs, forums and reviews collecting text that have a certain keyword. After the data gathering, this study can give a polarity score to the reviews, to be further analyzed by domain experts.

This work can be improved further by using a simple grammar for parsing the sentences. A grammar would help us identify which words form groups in a sentence, letting us further enhance the effect of the sentiment.

The most powerful word phrases list, is also a good resource to be analyzed. It can be used to find critical features when analyzing a new domain.

The creation of new tagged datasets should flourish this field, considering the need for high quality data in machine learning techniques. Our current datasets are in the movie review domain, and current methods may learn domain specific features. The creation of new datasets in different domains would enable the domain transfer studies in Turkish, generating more generic features for sentiment analysis.

Also creating some relation between a feature and its negated form may improve the performance.

Since this work is so dependent on the ability of current NLP tools, every update to the NLP tools could create new features to test on this dataset.

REFERENCES

[1] Edoardo M. Airoldi, Xue Bai, and Rema Padman. Markov blankets and meta-heuristic search: Sentiment extraction from unstructured text. *Lecture Notes in Computer Science (Advances in Web Mining and Web Usage Analysis)*, 3932:167–187, 2006.

[2] Anthony Aue and Michael Gamon. Customizing sentiment classifiers to new domains: a case study. In *Submitted to RANLP-05, the International Conference on Recent Advances in Natural Language Processing*, Borovets BG, 2005.

[3] Anthony Aue and Michael Gamon. Customizing sentiment classifiers to new domains: A case study. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, 2005.

[4] John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Association for Computational Linguistics*, Prague, Czech Republic.

[5] Lillian Lee. Movie Review Data Bo Pang. http://www.cs.cornell.edu/people/pabo/movie-review-data/. Last visited on: May 2009.

[6] Ruket Cakici. Wide-coverage parsing for turkish. In *PhD Thesis, University of Edinburgh*, 2008.

[7] Gobinda G. Chowdhury. Natural language processing. *Annual Review of Information Science and Technology*, 37:51–89, 2003.

[8] comScore/the Kelsey group. Online consumer-generated reviews have significant impact on offline purchase behavior. Press Release, November 2007.

[9] Sanjiv Das and Mike Chen. Yahoo! for Amazon: Extracting market sentiment from stock message boards. In *Proceedings of the Asia Pacific Finance Association Annual Conference (APFA)*, 2001.

[10] Kushal Dave, Steve Lawrence, and David M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of WWW*, pages 519–528, 2003.

[11] Ann Devitt and Khurshid Ahmad. Sentiment polarity identification in financial news: A cohesion-based approach. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 984–991, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[12] Boiy Erik, Hens Pieter, Deschacht Koen, and Moens Marie-Francine. Automatic sentiment analysis of on-line text. In *11th International Conference on Electronic Publishing*, pages 349–360, 2007.

[13] Andrea Esuli and Fabrizio Sebastiani. PageRanking WordNet synsets: An application to opinion mining. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, pages 424–431, Prague, CZ, 2007.

[14] Anindya Ghose, Panagiotis Ipeirotis, and Arun Sundararajan. Opinion mining using econometrics: A case study on reputation systems. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 416–423, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[15] Marti Hearst. Direction-based text interpretation as an information access refinement. In Paul Jacobs, editor, *Text-Based Intelligent Systems*, pages 257–274. Lawrence Erlbaum Associates, 1992.

[16] John A. Horrigan. Online shopping. Pew Internet & American Life Project Report, 2008.

[17] Zemberek 2 is an open source NLP library for Turkic languages. http://code.google.com/p/zemberek/. Last visited on: May 2009.

[18] Thorsten Joachims. A support vector method for multivariate performance measures. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pages 377–384, New York, NY, USA, 2005. ACM.

[19] Koppel, Moshe, Schler, and Jonathan. The importance of neutral examples for learning sentiment. *Computational Intelligence*, 22(2):100–109, May 2006.

[20] Edward Loper and Steven Bird. Nltk: The natural language toolkit. In *Proceedings of the ACL demonstration session*, pages 214–217, Barcelona, July 2004.

[21] Shotaro Matsumoto, Hiroya Takamura, and Manabu Okumura. Sentiment classification using word sub-sequences and dependency sub-trees.

[22] Ryan McDonald, Kerry Hannan, Tyler Neylon, Mike Wells, and Jeff Reynar. Structured models for fine-to-coarse sentiment analysis. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 432–439, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[23] Rada Mihalcea, Carmen Banea, and Janyce Wiebe. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 976–983, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[24] Tony Mullen and Nigel Collier. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 412–418, July 2004. Poster paper.

[25] Beyazperde Mynet. http://beyazperde.mynet.com/. Last visited on: May 2009.

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

[27] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*, pages 115–124, 2005.

[28] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86, 2002.

[29] Ellen Riloff, Siddharth Patwardhan, and Janyce Wiebe. Feature subsumption for opinion analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 440–448. Association for Computational Linguistics, 2006.

[30] Warren Sack. On the computation of point of view. In *Proceedings of AAAI*, page 1488, 1994. Student abstract.

[31] Hui Yang, Luo Si, and Jamie Callan. Knowledge transfer and opinion detection in the TREC2006 blog track. In *Proceedings of TREC*, 2006.

[32] Hong Yu and Vasileios Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *EMNLP2003*, pages 129–136, 2003.

APPENDIX A

EXAMPLE DATA

A.1 Polarity 2.0 Data Examples

One positive and one negative movie review examples from the polarity 2.0 dataset [28] are given in the following sections.

A.1.1 Positive Movie Review from Polarity 2.0

films adapted from comic books have had plenty of success , whether they're about superheroes (batman , superman , spawn) , or geared toward kids (casper) or the arthouse crowd (ghost world) , but there's never really been a comic book like from hell before .

for starters , it was created by alan moore (and eddie campbell) , who brought the medium to a whole new level in the mid '80s with a 12-part series called the watchmen . to say moore and campbell thoroughly researched the subject of jack the ripper would be like saying michael jackson is starting to look a little odd . the book (or " graphic novel , " if you will) is over 500 pages long and includes nearly 30 more that consist of nothing but footnotes . in other words , don't dismiss this film because of its source . if you can get past the whole comic book thing , you might find another stumbling block in from hell's directors , albert and allen hughes .

getting the hughes brothers to direct this seems almost as ludicrous as casting carrot top in , well , anything , but riddle me this : who better to direct a film that's set in the ghetto and features really violent street crime than the mad geniuses behind menace ii society ? the ghetto in question is , of course , whitechapel in 1888 london's east end . it's a filthy , sooty place where the whores (called " unfortunates ") are starting to get a little nervous about this mysterious psychopath who has been carving through their profession with surgical precision .

when the first stiff turns up , copper peter godley (robbie coltrane , the world is not enough) calls in inspector frederick abberline (johnny depp , blow) to crack the case . abberline , a widower , has prophetic dreams he unsuccessfully tries to quell with copious amounts of absinthe and opium . upon arriving in whitechapel , he befriends an unfortunate named mary kelly (heather graham , say it isn't so) and proceeds to investigate the horribly gruesome crimes that even the police surgeon can't stomach .

i don't think anyone needs to be briefed on jack the ripper , so i won't go into the particulars here , other than to say moore and campbell have a unique and interesting theory about both the identity of the killer and the reasons he chooses to slay . in the comic , they don't bother cloaking the identity of the ripper , but screenwriters terry hayes (vertical limit) and rafael yglesias (les mis ? rables) do a good job of keeping him hidden from viewers until the very end .

it's funny to watch the locals blindly point the finger of blame at jews and indians because , after all , an englishman could never be capable of committing such ghastly acts . and from hell's ending had me whistling the stonecutters song from the simpsons for days (" who holds back the electric car/who made steve guttenberg a star ? ") . don't worry - it'll all make sense when you see it .

now onto from hell's appearance : it's certainly dark and bleak enough , and it's surprising to see how much more it looks like a tim burton film than planet of the apes did (at times , it seems like sleepy hollow 2) . the print i saw wasn't completely 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 . ians holm

(joe gould's secret) and richardson (102 dalmatians) log in great supporting roles , but the big surprise here is graham . i cringed the first time she opened her mouth , imagining her attempt at an irish accent , but it actually wasn't half bad . the film , however , is all good . 2 : 00 - r for strong violence/gore , sexuality , language and drug content

A.1.2 Negative Movie Review from Polarity 2.0

plot : two teen couples go to a church party , drink and then drive . they get into an accident . one of the guys dies , but his girlfriend continues to see him in her life , and has nightmares . what's the deal ? watch the movie and " sorta " find out . . .

critique : a mind-fuck movie for the teen generation that touches on a very cool idea , but presents it in a very bad package . which is what makes this review an even harder one to write , since i generally applaud films which attempt to break the mold , mess with your head and such (lost highway & memento) , but there are good and bad ways of making all types of films , and these folks just didn't snag this one correctly . they seem to have taken this pretty neat concept , but executed it terribly . so what are the problems with the movie ? well , its main problem is that it's simply too jumbled . it starts off " normal " but then downshifts into this " fantasy " world in which you , as an audience member , have no idea what's going on .

there are dreams , there are characters coming back from the dead , there are others who look like the dead , there are strange apparitions , there are disappearances , there are a looooot of chase scenes , there are tons of weird things that happen , and most of it is simply not explained . now i personally don't mind trying to unravel a film every now and then , but when all it does is give me the same clue over and over again , i get kind of fed up after a while , which is this film's biggest problem . it's obviously got this big secret to hide , but it seems to want to hide it completely until its final five minutes . and do they make things entertaining , thrilling or even engaging , in the meantime ? not really .

the sad part is that the arrow and i both dig on flicks like this , so we actually figured most of it out by the half-way point , so all of the strangeness after that did start to make a little bit of sense , but it still didn't the make the film all that more entertaining . i guess the bottom line with movies like this is that you should always make sure that the audience is " into it " even before they are given the secret password to enter your world of understanding . i mean

, showing melissa sagemiller running away from visions for about 20 minutes throughout the movie is just plain lazy ! ! okay , we get it . . . there are people chasing her and we don't know who they are . do we really need to see it over and over again ?

how about giving us different scenes offering further insight into all of the strangeness going down in the movie ? apparently , the studio took this film away from its director and chopped it up themselves , and it shows . there might've been a pretty decent teen mind-fuck movie in here somewhere , but i guess " the suits " decided that turning it into a music video with little edge , would make more sense . 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 . oh , and by the way , this is not a horror or teen slasher flick . . . it's just packaged to look that way because someone is apparently assuming that the genre is still hot with the kids . it also wrapped production two years ago and has been sitting on the shelves ever since . whatever . . . skip it ! where's joblo coming from ? a nightmare of elm street 3 (7/10) - blair witch 2 (7/10) - the crow (9/10) - the crow : salvation (4/10) - lost highway (10/10) - memento (10/10) - the others (9/10) - stir of echoes (8/10)

A.2 Sentence Polarity Data Examples

Two positive and two negative movie review examples from the Sentence Polarity Dataset v1.0 [26] are given in the following sections.

A.2.1 Positive Movie Review from Sentence Polarity Dataset v1.0

the rock is destined to be the 21st century's new " conan " and that he's going to make a splash even greater than arnold schwarzenegger , jean-claud van damme or steven segal .

the gorgeously elaborate continuation of " the lord of the rings " trilogy is so huge that a column of words cannot adequately describe co-writer/director peter jackson's expanded vision of j . r . r . tolkien's middle-earth .

A.2.2 Negative Movie Review from Sentence Polarity Dataset v1.0

it's so laddish and juvenile , only teenage boys could possibly find it funny .

exploitative and largely devoid of the depth or sophistication that would make watching such a graphic treatment of the crimes bearable .

A.3 Turkish Polarity Data Examples

One positive and one negative movie review examples from the Turkish polarity dataset are given in the following sections.

A.3.1 Turkish Positive Movie Review

Açıkçası bu kadar düşük puan almayı hakketmeyecek kadar harika bir film. özellikle cem özer kendini bir kere daha kanıtlamış. düşük puan verenlerin bir kere daha izlemeleri gerkiyor bence.

A.3.2 Turkish Negative Movie Review

Bu oyun değil film..Zaten oyunda deli gibi korktuğumuz Nemesis saçma bir şekilde ölüyor bu filmde...Hele hele STARS grubunu yoketmesi yokmu...Nemesis in zaten duygusal yaratığa çevirmişler iş bitmiş...Oldu olacak evlenme teklifide etseydi...

A.3.3 Turkish Neutral Movie Review

Nicole Kidman ne Antony Hopkins ... ikisi br arada .... Sırf bu nedenle izlenecek bir film ... Filmnin ilk yarısında konu insanı pek bağlayamıyor..ikinci yarıda ise herşey aniden netleşiyor..Sakin kafayla gidilecekbir film..

A.4 Analysis of Extracted Features

Most powerful 40 word phrases for Turkish and English is given in the tables A.1 and A.2. In the tables the left column has the features with the most negative effect, the second column has the magnitude of the related word phrases, the fourth and the third columns have the positive word phrases respectively. This data is automatically generated from the training data without any supervision.

Table A.1: Most powerful 40 word phrases for Turkish

Negative Feature	Effect	Effect	Positive Feature
berbat	-1.35	1.03	beğendim
vasat	-1.21	1.02	mükemmel
beğenmedim	-1.18	0.94	harika
kötüydü	-1.16	0.93	süper
olmamış	-1.15	0.86	muhteşem
sıkıcı	-1.07	0.81	müthiş
değmez	-1.06	0.79	bekliyorum
kötü	-1.04	0.75	mükemmeldi
başarısız	-1.04	0.74	numara
rezalet	-1.03	0.74	eğlenceli
kaybı	-1.02	0.72	başyapıt
sıkıldım	-0.99	0.69	muhteşemdi
yazık	-0.98	0.67	değer
gereksiz	-0.95	0.66	sıkılmadan
sıkıcıydı	-0.94	0.64	mutlaka
berbattı	-0.91	0.64	harikaydı
etmiyorum	-0.88	0.64	kaçırmayın
kötüsü	-0.86	0.64	bayıldım
fiyasko	-0.85	0.63	izlenmeli
saçma	-0.83	0.62	güzeldi
etmem	-0.78	0.62	başarılı
kırıklığı	-0.77	0.61	keyifle
basit	-0.76	0.59	kusursuz
saçmalık	-0.76	0.59	etkileyici
vasatın	-0.76	0.58	söze
iğrenç	-0.75	0.58	sevdim
sıradan	-0.74	0.57	süperdi
izlemeyin	-0.71	0.57	güzel
rezil	-0.7	0.56	sürükleyici
sıkıntıdan	-0.68	0.55	seviyorum
para	-0.67	0.55	eğlenmek
hüsran	-0.66	0.54	manyak
vasattı	-0.65	0.54	zevkli
abartılmış	-0.65	0.54	beğenmeyen
maalesef	-0.64	0.53	izlenesi
vasatı	-0.63	0.53	sonuçta
sevmedim	-0.63	0.53	haber
sarmadı	-0.62	0.53	izlemesi
gitmeyin	-0.62	0.52	herkese
dandık	-0.61	0.51	ilaç

Table A.2: Most powerful 40 word phrases for English

Negative Feature	Effect	Effect	Positive Feature
dull	-1.32	1.05	unique
boring	-1.3	1.04	entertain
fails	-1.2	1.01	masterpiece
neither	-1.18	1	diverting
supposed	-1.09	1	provides
unless	-1.09	0.99	skin
too	-1.05	0.98	cinema
disappointment	-1.04	0.97	ages
tedious	-1.02	0.96	thanks
badly	-1.02	0.95	refreshing
plodding	-1	0.93	remarkable
pie	-1	0.91	wonderful
ill	-0.99	0.91	treat
flat	-0.99	0.9	gem
worst	-0.98	0.89	works
unfunny	-0.97	0.89	always
schneider	-0.96	0.89	entertaining
lack	-0.94	0.88	colorful
inept	-0.93	0.86	breathtaking
routine	-0.9	0.85	engrossing
waste	-0.88	0.85	marvel
intentions	-0.88	0.84	resist
suffers	-0.88	0.84	brilliant
maudlin	-0.88	0.83	russian
artificial	-0.86	0.82	affection
superficial	-0.86	0.82	delightful
lacks	-0.85	0.81	smarter
none	-0.85	0.81	smart
unfortunately	-0.85	0.8	rare
name	-0.84	0.8	imax
bore	-0.83	0.79	polished
exhausting	-0.83	0.79	warm
hasn	-0.82	0.78	beautifully
disguise	-0.82	0.78	inventive
product	-0.82	0.78	open
oh	-0.81	0.77	intimate
stupid	-0.81	0.77	spider
devoid	-0.8	0.77	culture
stunt	-0.79	0.77	grown
onscreen	-0.79	0.76	hilarious