JANUARY 2013

ISTANBUL KULTUR UNIVERSITY
INSTITUTE OF SCIENCES

SENTIMENT ANALYSIS FOR TURKISH LANGUAGE

MS Thesis by

Hakan ÇELİK

0909102001

Date of submission : 15 January 2013
Date of defence examination : 21 January 2013

Supervisor and Chairperson : Prof. Dr. Coşkun BAYRAK

University	**:**	**Istanbul Kültür University**
Institute	**:**	**Institute of Science**
Department	**:**	**Computer Engineering**
Programme	**:**	**Computer Engineering**
Supervisor	**:**	**Prof. Dr. Coşkun Bayrak**
Degree Awarded and Date	**:**	**MS – January 2013**

ABSTRACT

SENTIMENT ANALYSIS FOR TURKISH LANGUAGE

Hakan Çelik

Sentiment analysis a.k.a. opinion mining is an application of natural language processing to extract subjective information from given texts. Consumer comments, evaluation of films, stock exchange predictions and political researches are some examples of subjective declarations that sentiment analysis involves.

Subjective data have hugely increased parallel to web 2.0 and social media growth in recent years. People generate petabytes of data every day. Processing of this amount of raw data became more and more important for both companies and individuals.

Many sentiment analysis studies conducted in the past are focused on English Language only. Therefore, for Turkish Language, sentiment analysis is a blue ocean yet. There is a single study focused on only one input domain and tested with only one machine learning algorithm.

Within the focus of this thesis new datasets were constructed in various domains, a Turkish specific input preparation algorithm was introduced, and different machine learning algorithms were applied over prepared data. ~85% accuracy was achieved with Naïve Bayes Classifier.

Keywords: Sentiment Analysis, Machine Learning, Natural Language Processing, Opinion Mining, Data Mining, Turkish, Text polarity, Naïve Bayes, Support Vector Machines, Logistic Regression, Decision Tree

Üniversite	**:**	**İstanbul Kültür Üniversitesi**
Enstitü	**:**	**Fen Bilimleri Enstitüsü**
Dal	**:**	**Bilgisayar Mühendisliği**
Program	**:**	**Bilgisayar Mühendisliği**
Tez Danışmanı	**:**	**Prof. Dr. Coşkun Bayrak**
Tez Türü ve Tarihi	**:**	**Yüksek Lisans – Ocak 2013**

KISA ÖZET

TÜRK DİLİ İÇİN DUYGU ANALİZİ

Hakan Çelik

Duygu yada düşünce analizi metinlerden kişisel değerlendirmelerin çıkarımını sağlayan doğal dil işleme branşıdır. Duygu analizinin ilgi alanına giren değerlendirmelere örnek olarak tüketici yorumları, film değerlendirmeleri, borsa tahminleri ve siyasi araştırmalar gibi geniş ölçekte örnekler verilebilir.

Son yıllarda web 2.0 ve sosyal medyanın artış hızına bağlı olarak öznel değerlendirmeler müthiş derecede arttı. Her gün insanlar petabaytlarca veri girişi yapıyor. Bu boyuttaki ham verinin analizi gerek bireyler gerekse de şirketler için gittikçe daha fazla önem arz etmeye başladı.

İngilizce için birçok duygu analizi çalışmaları mevcut. Ancak bu çalışma alanı Türkçe için hala bakir. Gerçekleştirilen tek çalışma tek bir alan üzerine yoğunlaşıp tek tipte veri üzerinde tek tip makina öğrenmesi algoritması kullanmış.

Bu tez kapsamında farklı alanlara ait yeni veri grupları oluşturulmuş, Türkçe'ye özel veri hazırlama algoritması tanıtılmış ve oluşturulan veri üzerinde farklı makina öğrenmesi algoritmaları uygulanmıştır. Naïve Bayes sınıflandırıcı kullanılarak başarılı sayılabilecek %85'lik doğruluk oranı yakalanmıştır.

Anahtar Kelimeler: Duygu Analizi, Makine Öğrenmesi, Doğal Dil İşleme, Fikir Madenciliği, Veri Madenciliği, Türkçe, Metin Kutbu, Naïve Bayes, Destek Vektör Makineleri, Mantıksal Regresyon, Karar Ağaçları

TABLE OF CONTENTS

ABSTRACT ii

KISA ÖZET iii

TABLE OF CONTENTS iv

ABBREVIATIONS vii

LIST OF TABLES viii

LIST OF FIGURES ix

INTRODUCTION 1

1.1 DEFINITON 2

1.2 MOTIVATION & IMPORTANCE 2

1.3 OUTLINE 3

BACKGROUND 4

2.1 LITERATURE SURVEY 5

2.2 TURKISH LANGUAGE ANALYSIS 6

2.3 BACKGROUND 7

2.3.1 Machine Learning **7**

2.3.2 Support Vector Machines **7**

2.3.3 Decision Tree **10**

2.3.4 Maximum Entropy (Logistic Regression) **10**

2.3.5 Naïve Bayes **11**

2.3.6 K-fold Cross Validation **12**

2.3.7 Natural Language Processing **12**

2.3.8 Bag of Words and n-gram Model **12**

2.3.9 Negation Handling **13**

2.3.10 Thresholding **13**

2.3.11 Spell Check and Manipulation 14

2.3.12 WEKA Machine Learning Library 14

2.3.13 Zemberek NLP Library 14

METHODOLOGY 15

3.1 DATA GATHERING 16

3.2 DATA SELECTION 16

3.3 WORD PREPARATION 17

3.4 DATA PREPARATION 20

3.5 TRAINING AND TEST 21

RESULTS 22

4.1 RESULTS 23

4.1.1 ML Algorithms Comparison 23

4.1.2 Input Type Comparison 25

4.1.3 n-gram Size Comparison 26

4.1.4 Data Manipulation Methods Comparison 27

CONCLUSION 29

5.1 CONCLUSION 30

5.2 FUTURE WORK 30

REFERENCES 32

APPENDICES 34

APPENDIX A 35

EXAMPLE DATA 35

A.1 Example Book Comments 35

A.2 Example Film Comments 35

A.3 Example Shopping Comments 35

APPENDIX B 36

TOP 100 n-grams 36

B.1 Top 100 Unigrams 36

B.2 Top 100 Bigrams 37

B3. Top 100 Trigrams 38

ABBREVIATIONS

ML : Machine Learning

NLP : Natural Language Processing

SVM : Support Vector Machines

NB : Naïve Bayes

NBM : Naïve Bayes Multinomial

LR : Logistic Regression

DT : Decision Tree

ARFF : Attribute Relation File Format

LIST OF TABLES

TABLES

Table 2.1: Turkish suffixes 6

Table 3.1: Data counts 16

Table 4.1: Comparison of ML algorithms with no threshold 24

Tables 4.2: Comparison of ML algorithms with various threshold values 24

Table 4.3: Input type comparison with various threshold values 25

Table 4.4: n-gram comparison with different threshold values. 27

Table 4.5: Input manipulation methods comparison 28

Table B.1: Top 100 Unigrams 36

Table B.2: Top 100 Bigrams 37

Table B.3: Top 100 Trigrams 38

LIST OF FIGURES

FIGURES

Figure 2.1: SVM Hyper planes for linear classification 8

Figure 2.2: Complex data 9

Figure 2.3: Mapping data to higher dimension 9

Figure 2.4: A decision tree 10

Figure 2.5: Logistic Regression 11

Figure 2.6: Naïve Bayes Classifier - independent features 11

Figure 3.1: Data preparation flow 19

Figure 4.1: Accuracy graph of comments with various threshold values 25

Figure 4.2: Accuracy with increasing n-gram size 26

1

INTRODUCTION

“Most people are other people. Their thoughts are someone else's opinions, their lives a mimicry, their passions a quotation. ”

- Oscar Wilde

1.1 DEFINITON

Sentiment is the attitude toward something, i.e. opinions, feelings, emotions, subjective analysis about anything.

Sentiment analysis a.k.a. opinion mining is an application of natural language processing to extract subjective information from given texts. General attitude for some topic or overall polarity of a text are some applications of sentiment analysis.

As stated in [1], "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".

Sentiment analysis as the name refers, focuses on subjective declarations. Consumer comments, evaluation of films, stock exchange predictions and even political researches are some examples for these kinds of subjective declarations.

1.2 MOTIVATION & IMPORTANCE

Other people's opinions and feedbacks are always important for us. People generally learn best through experience. This is either personal experience or someone else's.

Sentiment analysis became more and more important in parallel to web 2.0 and social media growth. Twitter messages and trends, personal blogging, product commenting are very common today. Everyday petabytes of data are created and released to internet. For instance, people generate over 400 million tweets daily on twitter as of October 2012. This amount of data can't be read, understood or analyzed manually. Automatic analysis is needed in order to maximize benefit from data.

A recent study [2] shows that "58% of Americans have researched a product or service online before buying". Another perspective revealed in [3] indicated that Twitter messages predict

the stock market with 87.6% accuracy. Additional research report [4] indicates “66% of social media users have used internet to post their thoughts about civic and political issues”. According to a recent study [5] there is a strong correlation between the amount of attention for a forthcoming movie and its ranking in the future. Nonetheless, these recent studies and reports show that sentiment analysis is very important for information accuracy and sentimentally analyzed data become reliable information source.

There are many studies about sentiment analysis for English Language. However, there is only one [6] pioneering study focusing on movie reviews accomplished for Turkish Language. However, the drawback/weakness of this work comes from the single domain usage and based on support vector machines as machine learning methodology. Therefore, this thesis aims to apply proven methods of sentiment analysis used in English language to Turkish language in multiple domains and assess how the accuracy can be maximized in comparing some machine learning methods with some natural language methods.

1.3 OUTLINE

Chapter 2 contains literature survey, a brief history of sentiment analysis, background information of sentiment analysis, machine learning, natural language processing and Turkish language structure analysis. In Chapter 3, methodology details like data gathering, preparation and selection, training and test are described. Chapter 4 gives the results and discussions. Summary and conclusions are described in Chapter 5.

2

BACKGROUND

“Too often we enjoy the comfort of opinion without the discomfort of thought.”
― John F. Kennedy

2.1 LITERATURE SURVEY

Sentiment analysis studies goes quite back in time. An earlier work done in [7] tried to model human understanding of natural language using politics and political events back in 1979. Even multiple computer programs were written for modeling this study.

Many other studies [8][9] focused on manual construction of word lexicons by tagging words. These early studies were tried to limit lexicon sizes because large lexicons were consuming too much computer resources which are scant in those days.

Sentiment analysis thrived in 2000s with the World Wide Web. Research areas and opportunities were seen with internet spreading globally and hundreds of papers and researches were performed. Web 2.0 permitted users to participate the internet world and that provided datasets for machine learning algorithms. Also with Web 2.0, commercial and intelligent application areas of sentiment analysis were realized.

One of the milestone studies, "Thumbs up? Sentiment Classification using Machine Learning Techniques" by Pang et al. (2002) conducted supervised learning algorithms over IMDB (Internet Movie Database) film review data. They tried different machine learning algorithms (Naïve Bayes, Support Vector Machines, Maximum Entropy) over Bag of Words natural language processing approach. They had achieved 82.9% accuracy with use of support vector machines.

Same year Turney et al. tried a different unsupervised learning approach for word sentiment analysis [11], in which they sent queries to a search engine and using point-wise mutual information to analyze the results. They achieved 80% accuracy with their 3596 item test set.

Pang et al. (2005) tried to extend class numbers. They had achieved 70% accuracy on 3-class classification and 50% accuracy on 4-class classification.

Some researchers such as [13], [14] and [15] tried to accomplish cross-lingual sentiment classification i.e. translate and use English corpus for sentiment analysis in Romanian and Chinese. However, they generally scored below English language in their tests because of the language gap.

There is a word-level sentiment analysis study in Turkish [26] which used WordNets to construct a word relatedness graph for a multilingual approach for word polarity detection.

The only text based study [6] about Turkish Language has focused on some supervised and unsupervised approaches and uses film review data. This pioneer study has achieved 85% accuracy on single domain.

2.2 TURKISH LANGUAGE ANALYSIS

Turkish is an agglutinative and extensively inflected language. Many suffixes can be added to words and a word can form a sentence, as shown in Table 2.1.

Table 2.1: Turkish suffixes

Turkish	English
Evimdeyim.	I am at my house.

Despite the basic word order of Turkish is subject–object–verb, Turkish sentences can be formed with irregular order.

2.3 BACKGROUND

2.3.1 Machine Learning

Simply, machine learning (ML) is constructing systems that can learn. It is a subfield of artificial intelligence that aims to develop programs which can teach themselves to expand and improve when exposed to new data.

More formally, as indicated in [16] “a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E”.

There are basically 2 main types of learning, supervised and unsupervised. Data labels are known during supervised learning and ML algorithm uses these labels. However, unsupervised learning methods try to classify or cluster data without guidance of labeled training set.

Within the scope of this thesis some supervised ML algorithms are considered and tested.

2.3.2 Support Vector Machines

A Support Vector Machine (SVM) constructs an *N*-dimensional hyper plane that separates the data into two categories to classify data. This hyper plane is chosen as the distance between nearest points on each side is maximized, as shown in Figure 2.1.

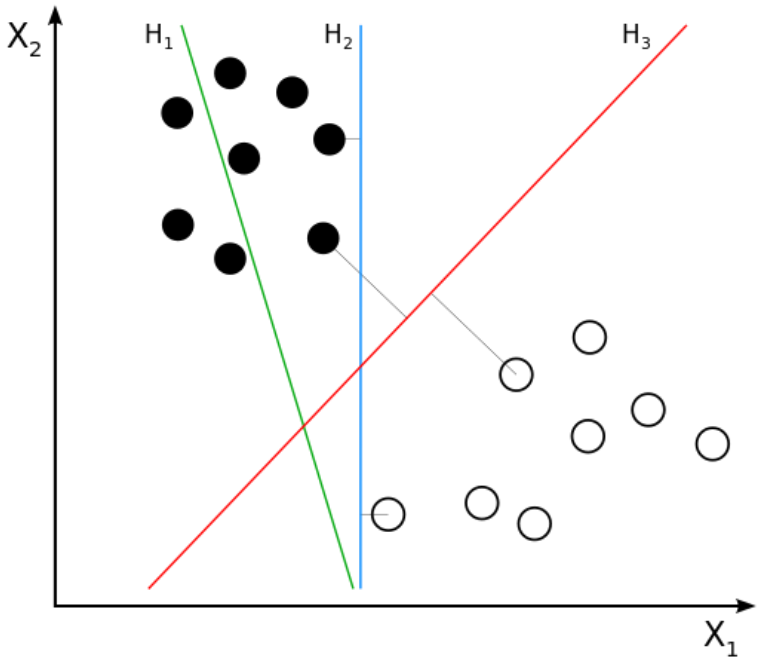

Figure 2.1: SVM Hyper planes for linear classification [23]

Generally complex data, like in Figure 2.2., can't be separated linear. In these cases kernel functions are introduced to map the data into a different (higher) space as shown in Figure 2.3.

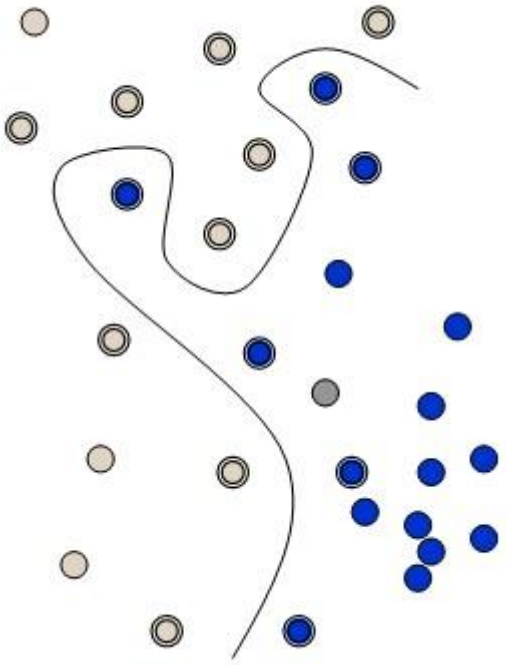

Figure 2.2: Complex data [23]

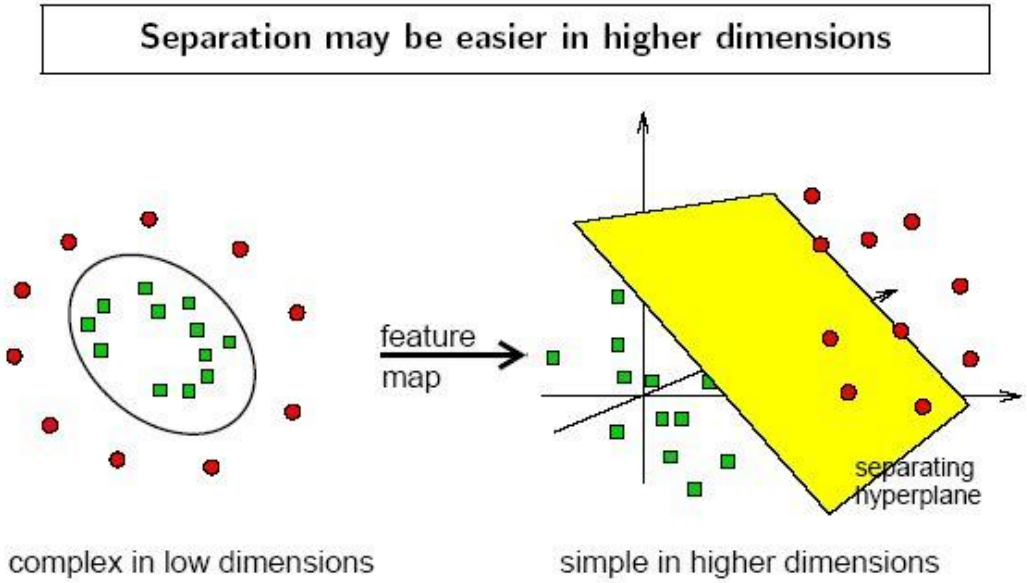

Figure 2.3: Mapping data to higher dimension [17]

2.3.3 Decision Tree

Decision trees are structures which maps observations about input to conclusions about target. In these tree structures, class labels are represented as leaves and compositions of features are represented as branches. An example is shown in Figure 2.4.

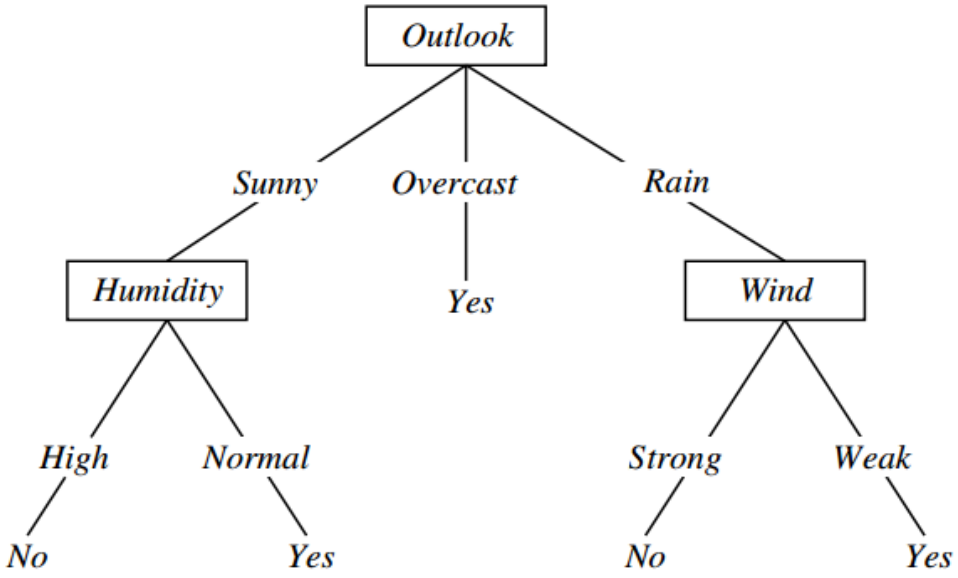

Figure 2.4: A decision tree

2.3.4 Maximum Entropy (Logistic Regression)

Logistic regression is a ML method to analyze the relation between dependent variable and predictor variables. As Figure 2.5 shows, an S-type separator is drawn between class labels and cuts the sides equally.

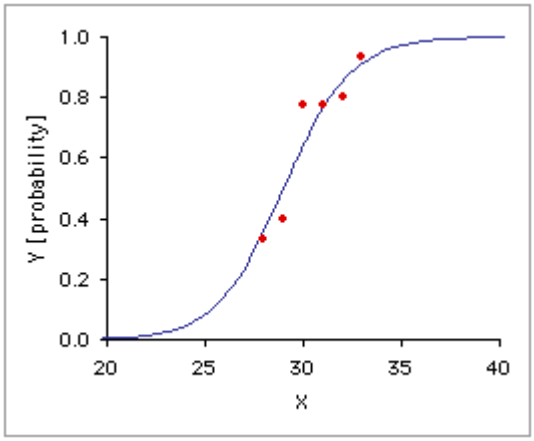

Figure 2.5: Logistic Regression [25]

2.3.5 Naïve Bayes

A Naïve Bayes classifier assumes that each feature of a class is independent of any other feature, i.e. all features independently affect the probability of the result. Due to this simplifying assumption, Naïve Bayes classifier particularly is suitable when the dimension of the input is high. A sample application practice with independent features is shown in Figure 2.6.

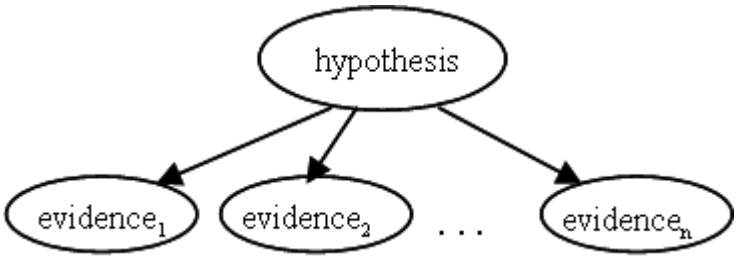

Figure 2.6: Naïve Bayes Classifier - independent features

2.3.6 K-fold Cross Validation

Cross validation is a technique for predictive model performance estimation. Data set is partitioned into subsets randomly for analysis within this validation technique.

In k-fold cross validation, data set is randomly divided into k subsets. Each subset is considered as test data while remaining are used as training data. This process is repeated k times as all parts of data are used once for validation. Finally, the validation results of all runs are averaged for a single and final estimation.

2.3.7 Natural Language Processing

Natural language processing (NLP) is an area of artificial intelligence which aims to enable computers to understand human natural language. In order to achieve that, generally ML algorithms are used with NLP methods. NLP is used widely for tasks like question answering, machine translation, optical character recognition, speech recognition, text to speech, sentiment analysis, etc.

2.3.8 Bag of Words and n-gram Model

Bag of words is a simplifying natural language processing model that a text is represented as an unordered collection of words. This model ignores word order, grammar and punctuations. For example:

Input - 1: Bob is a hardworking student.

Input - 2: Sally is a lazy student.

Bag of words - 1: "hardworking", "a", "Bob", "student", "is"

Bag of words - 2: "lazy", "Sally", "student", "a", "is"

Dictionary: {"Bob":1, "is":2, "a":3, "hardworking":4, "student":5, "lazy":6}

An n-gram is continuous n items from a given text. n-gram of size 1 is referred to as a "unigram"; size 2 is a "bigram"; size 3 is a "trigram".

n-grams are generally used with bag of words approach. Firstly, n-grams are generated from given text, then this n-grams are considered elements in the bag.

For example:

Input: "Bob is a hardworking student"

Unigrams: {"Bob", "is", "a", "hardworking", "student"}

Bigrams: {"Bob is", "is a", "a hardworking", "hardworking student"}

Trigrams: {"Bob is a", "is a hardworking", "a hardworking student"}

2.3.9 Negation Handling

Bag of words approach ignores word order and can lose negation information. For English, "not" is the negation indicator. Previous studies [10][24] show that handling negation by adding NOT keyword next to all words afterwards improves accuracy slightly.

In Turkish language, most of the negations are handled by suffixes or prefixes. However, similar to “not” in English, "değil" is used as the negation indicator in Turkish. Therefore, negation handling will be conducted accordingly within the scope of this thesis.

2.3.10 Thresholding

Thresholding is the feature ranking of the bag of words approach. n-grams are thresholded with respect to their frequencies. Although a slight increase in accuracy is achieved, due to the exclusion of uncommon n-grams within training and decision process, a huge performance gain is obtained.

2.3.11 Spell Check and Manipulation

Data are being checked for spelling errors or typos before constructing n-grams. Wrong written words can be corrected or disregarded completely. This approach also increases accuracy slightly.

2.3.12 WEKA Machine Learning Library

Weka [18] is a collection of machine learning algorithms for data mining tasks which has a well-documented Java API widely used in academic studies.

2.3.13 Zemberek NLP Library

Zemberek [19] is a general purpose Natural Language Processing library for Turkish. This Java library is used for spell checking and corrections in this thesis.

3

METHODOLOGY

“Sentiment without action is the ruin of the soul.”

— Edward Abbey

3.1 DATA GATHERING

Unfortunately, there was no Turkish sentiment training or test data available. Therefore, http crawlers were written for three sites [20][21][22] and book, film, and shopping comments were gathered along with their rating points. Namely, over 20K book comments, 13K film comments, and 51K shopping comments were collected. For more information on sample data see Appendix A.

3.2 DATA SELECTION

Many of the comments are positive oriented. For instance negative comments are 4-5% of all data according to domain. In order to prevent training data to be positively emphasized, only equal number of positive and negative comments is taken into consideration. On the other hand, all comments of three types were merged for cross-domain check and it was considered as 4th data type with the name of "ALL".

The data set sizes are shown in Table 3.1.

Table 3.1: Data counts

	Raw Data	**After Equalization of Polarity**
Book	20623	1548
Film	13156	2248
Shopping	51879	5256
ALL	85658	9624

3.3 WORD PREPARATION

There are many typos, shortening, etc. in comments. Best grammatical comments are book site's as expected. Film comments take the second place, right before the general shopping comments. Input manipulation is needed in order to achieve higher success ratio. In order to do that Zemberek library was used for spell checking and character correction. Zemberek does simple character correction with changing ASCII counterparts of Turkish specific characters and tries to find word in dictionary. For example:

Input: *Bu okudugum en harika kitap.*

Output: *Bu okuduğum en harika kitap.*

However, more complex typos and some proper nouns fail in Zemberek spell check. These failed n-grams were disregarded and excluded. For example:

Input: *İşte Marian Keyes en sevdiğim yazadır.*

Output: *İşte en sevdiğim*.

Input was converted to lowercase in order to avoid some matching and evaluation errors. For Example:

Input: *Tamam hepsi birbirinden güzel ama GURBETİ BEN YAŞADIM bambaşka...*

Output: *tamam hepsi birbirinden güzel ama gurbeti ben yaşadım bambaşka...*

Punctuation is not important for bag of words approach. However, many people do not care correct spacing between lines. So, a process of removing all punctuation from input may

mislead because of possibility of merging two separate words. Therefore punctuation characters should be replaced with empty space characters except apostrophe. For example:

Input:

Kitap yalın anlatımlı ve sürükleyici bir günde bitebilir.Biraz Suriye'deki yaşam tarzından biraz Eset ailesinden bahsediyor.

Output: (full stops are replaced with empty spaces but apostrophe is not replaced, just removed)

Kitap yalın anlatımlı ve sürükleyici bir günde bitebilir Biraz Suriyedeki yaşam tarzından biraz Eset ailesinden bahsediyor

On the other hand, some very short words, probably typos, should be disregarded as well. So, one or two character words were excluded. For Example:

Input: *Güzel bi anlatım geniş hikaye.*

Output: *Güzel anlatım geniş hikaye.*

For Turkish Language, the negation keyword is "*değil*". In our approach, if the word appears before the negation keyword or one of its variations (i.e., *değildi*) is an adjective, it is merged with negation keyword. For example:

Input: *Yeni anı oluşturmak konusunda nedense o kadar da başarılı değil.*

Output: *Yeni anı oluşturmak konusunda nedense o kadar da başarılıDeğil.*

The data preparation process flow used so far is given in Figure 3.1.

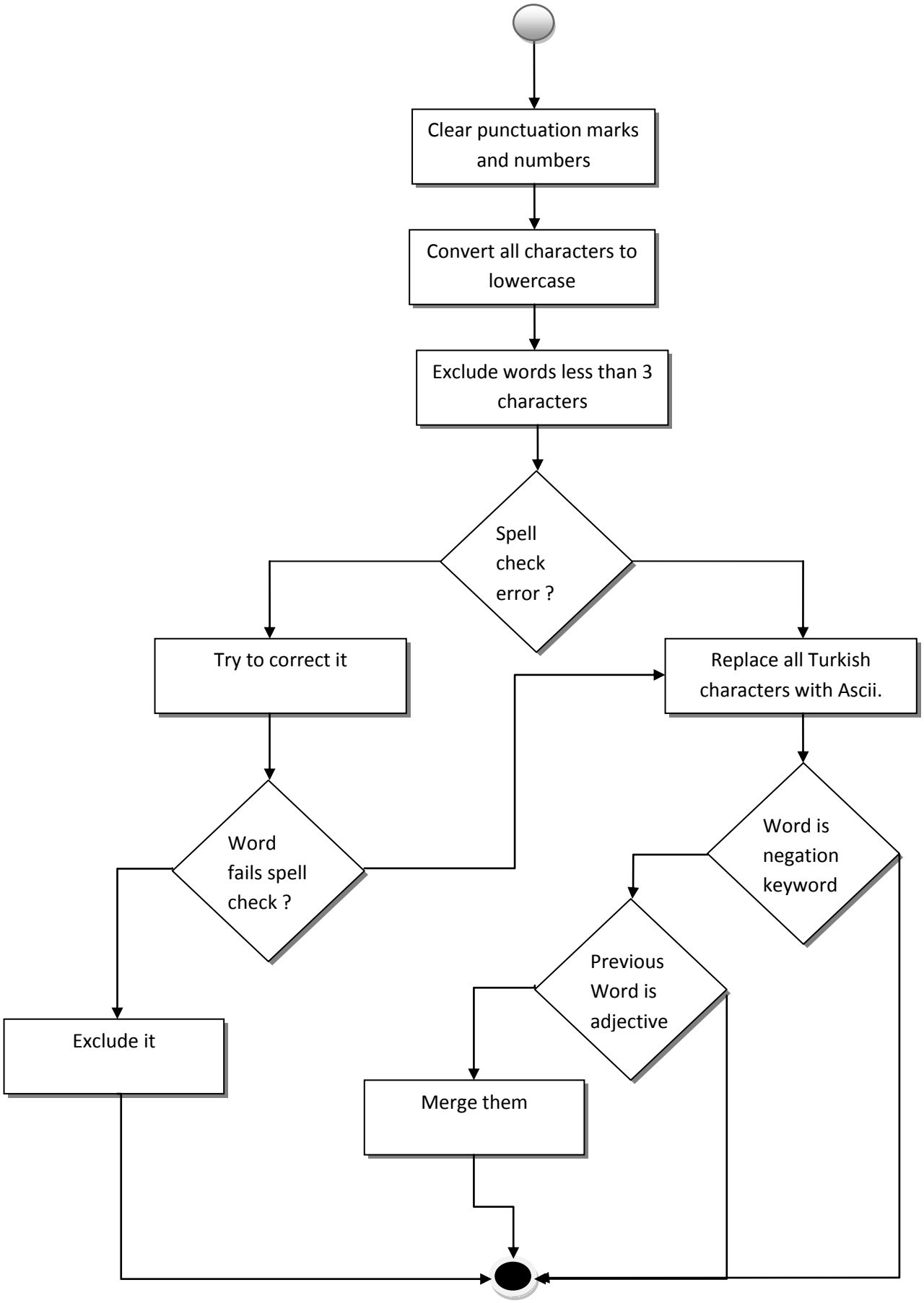

Figure 3.1: Data preparation flow

3.4 DATA PREPARATION

n-grams are prepared after manipulation of words. n-gram size is configurable and should be given as a positive integer number. Thresholding is applied on n-grams if configured so. Finally attribute relation file format (arff) file is prepared in order to be used in Weka library.

Input file starts with relation name. After that all n-grams are listed as attributes (features). Attribute type is integer because value of an attribute is the frequency in the data instance. Dataset instances exist after attribute definitions. Each dataset instance starts with CLASS_LABEL attribute (index 0) with value 0 (negative) or 1 (positive) and goes on as mappings of attribute ids with their frequencies.

A sample arff file is as below:

@RELATION ALL_1GRAM

@attribute CLASS_LABEL {0,1}

@attribute "muhtesem" integer

@attribute "bir" integer

@attribute "kitap" integer

@attribute "insan" integer

…

@attribute "kaydirmazlik" integer

@attribute "tutabildigini" integer

@attribute "tozlanarak" integer

@attribute "kullanilacagindan" integer

@data

{0 1,1 1,2 1,3 2,4 1,5 1,6 1,7 1,8 1,9 1,10 1,11 1,12 1,13 1,14 1,15 1,16 1,17 1,18 1,19 1,20 1,21 1}

{0 0,15 1,101 1,102 1,103 1,104 1,105 1,106 1,107 1,108 1,109 1,110 1,111 1,112 1,113 1,114 1,115 1}

{0 1,2 1,3 1,26 1,33 1,116 1,117 1,118 1,119 1,120 1,121 1}

{0 0,201 1,202 1,203 1,204 1,205 1,206 1,207 1,208 1,209 1,210 1}

{0 1,26 1,81 1,139 1,211 1,212 1,213 1,214 1,215 1,216 1,217 1,218 1,219 1,220 1}

For example, a dataset instance like {0 1,1 1,2 1,3 2,4} means a positive sentence consists of n-grams with ids 1, 2, 3 and 4. These n-grams have frequency of 1except 4^{th}.

3.5 TRAINING AND TEST

Logistic Regression, Naïve Bayes (Multinomial), Support Vector Machines, and Decision Tree ML algorithms were used for classification. Furthermore 10-fold cross validation technique was applied on training and test process.

4

RESULTS

"Let us beware of common folk, common sense, sentiment, inspiration, and the obvious."

— Charles Baudelaire

4.1 RESULTS

The results obtained are given and analyzed in this chapter. The fundamental aim of results is mainly to focus on increasing the accuracy. There are analysis criteria, such as precision, recall, and F1, used in previous studies but they are generally for handling cases of heterogeneous distribution of polarity input data. However, in this study the positive and negative input counts are equalized and these indicators are not considered.

However, the execution time of training and test process was especially important for the study. Very large datasets were generated in data preparation phase, which causes processing phase length to be unreasonable. Some ML algorithms respond much faster than the others.

There are more than 60 generated data files, which has 0, 10, 50, and 100 threshold values; 1-grams, 2-grams and 3-grams; spell-check enabled/disabled; negation handling enabled/disabled etc. The largest data file in size is 58 MB. The execution time of this data file is defined in terms of *days* for ML algorithms except Naïve Bayes.

As the results in table 4.1 and 4.2 show, Naïve Bayes algorithm is the top performer in terms of accuracy. Therefore, many input combinations were tested only with Naïve Bayes to differentiate configurable properties like n-gram size or best threshold value.

4.1.1 ML Algorithms Comparison

For bigger sized inputs, some machine learning algorithms have used longer execution time (days) to complete the process. The following description is a sorting of tested algorithms according to processing time (faster to slower)

Naïve Bayes < *Support Vector Machines* < *Logistic Regression* < *Decision Tree*

Table 4.1 shows a comparison of ML algorithms in book comments.

Table 4.1: Comparison of ML algorithms with no threshold

Book Comments ML Algorithm Comparison	Correctly Classified	Incorrectly Classified	Accuracy
Naïve Bayes	1322	227	85.34 %
Support Vector Machines	1291	258	83.34 %
Logistic Regression	1181	368	76.24 %
Decision Tree	1156	393	74.62 %

Total Comments: 1549

When feature thresholding was introduced for the aforementioned four ML algorithms, no great variation is encountered as shown in Table 4.2. Naïve Bayes is the winner again.

Table 4.2: Comparison of ML algorithms with various threshold values

	Accuracy			
Book Comments ML Algorithm Comparison w/ thresholding	**Threshold 0**	**Threshold: 10**	**Threshold: 50**	**Threshold: 100**
Naïve Bayes	85.34 %	85.80 %	83.93 %	80.35 %
Support Vector Machines	83.34 %	81.52 %	80.40 %	78.11 %
Logistic Regression	76.24 %	73.88 %	67.40 %	72.37 %
Decision Tree	74.62 %	74.33 %	73.08 %	73.76 %

Additional experiments were conducted on Naïve Bayes algorithm to test the possible degree of fluctuation. This algorithm is both fastest and the most accurate between four of them.

4.1.2 Input Type Comparison

Four input types were tested, namely book, film, shopping, and all comments. Results of all kinds of input types with variable threshold, unigram structure, and classified in Naïve Bayes are given in Table 4.3.

Table 4.3: Input type comparison with various threshold values

	Accuracy			
Commenty Type Comparison w/ thresholding	**Threshold 0**	**Threshold: 10**	**Threshold: 50**	**Threshold: 100**
Book Comments	85.34 %	85.80 %	83.93 %	80.35 %
Film Comments	83.99 %	82.65 %	80.21 %	79.07 %
Shopping Comments	79.62 %	79.74 %	79.78 %	77.87 %
All Comments	80.86 %	80.71 %	79.73 %	79.18 %

The results with increasing threshold are shown in Figure 4.1.

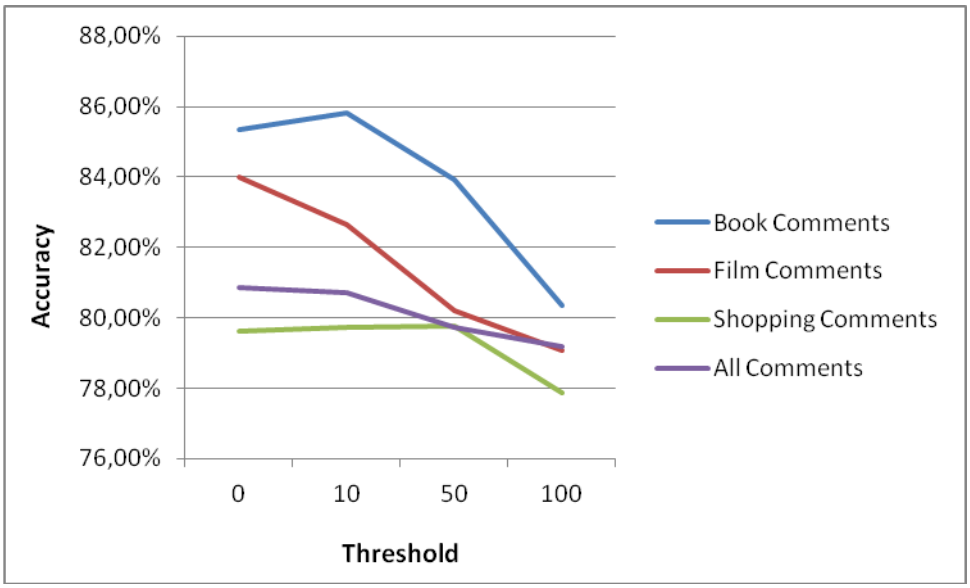

Figure 4.1: Accuracy graph of comments with various threshold values

Best results are book site's and worst results belong to shopping site comments. When these comments are reviewed, it is obvious that use-of- language, socio-educational, and age level provide different outcomes. Book readers use language much more effectively, care about grammar and make less typo. However, shopping site contains all kinds of comments from all sections of the society, age groups, and educational levels.

Additional experiments were done with only two extreme examples: book and shopping comments, which gave best and worst accuracy rates.

4.1.3 n-gram Size Comparison

Generally accuracy decreases with increase in n-gram size as shown in Figure 4.2. Actually expected results were opposite, bigrams was thought to be more successful than unigrams because they seem to be more meaningful.

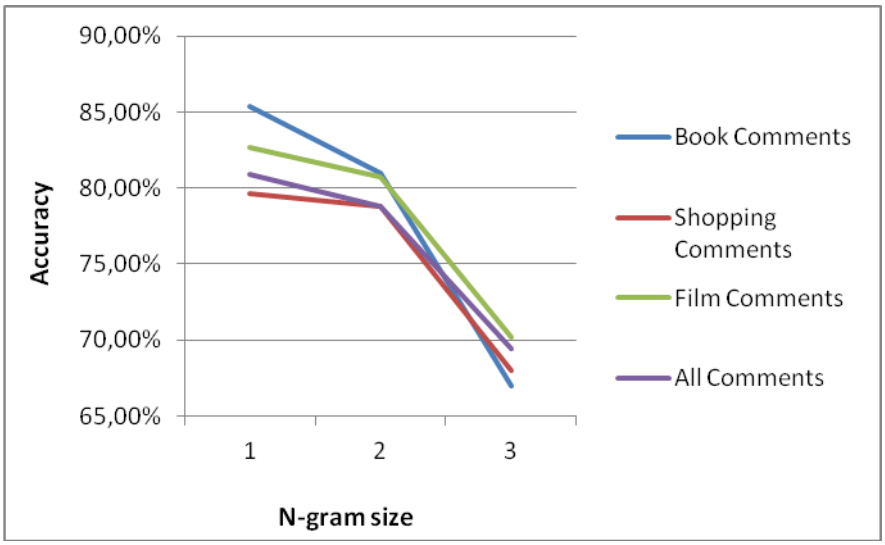

Figure 4.2: Accuracy with increasing n-gram size

Increasing threshold value does not change results. As Table 4.4 shows, accuracy decreases with n-gram size increasing.

Table 4.4: n-gram comparison with different threshold values.

	Accuracy			
n-gram Comparison w/ thresholding	**Threshold 0**	**Threshold: 10**	**Threshold: 50**	**Threshold: 100**
Book Comments - 1gram	85.34 %	85.80 %	83.93 %	80.35 %
Book Comments - 2gram	81.02 %	81.54 %	76.93 %	79.28 %
Book Comments - 3gram	66.95 %	73.56 %	75.34 %	27.27 %
Shopping Comments - 1gram	79.62 %	79.74 %	79.78 %	79.07 %
Shopping Comments - 2gram	78.81 %	77.56 %	75.46 %	72.59 %
Shopping Comments - 3gram	68.01 %	67.05 %	66.89 %	69.23 %

It should be noted that the accuracy achieved in trigram book comments with the threshold value of 100, is extremely low (27.27%) due to the decreasing level of test instances (only 11 test instances left after thresholding trigrams with 100).

4.1.4 Data Manipulation Methods Comparison

Spell check, negation handling and ASCII conversion increases the accuracy by 1% all together which is far below the expectations. Especially, negation handling seems to be ineffective in this approach. Therefore, further investigation is necessary. Table 4.5 shows the effects of data manipulation methods.

Table 4.5: Input manipulation methods comparison

Data Manipulation Methods Comparison	Correctly Classified	Incorrectly Classified	Accuracy
No input manipulation	1314	245	84.28 %
Only conversion to English characters disabled	1320	229	85.21 %
Only spell check disabled	1319	240	84.60
Only negation handling disabled	1322	227	85.34
Full input manipulation	1322	227	85.34

Total Comments: 1549

5

CONCLUSION

“People with opinions just go around bothering one another.”

-Gautama Buddha

5.1 CONCLUSION

This study aimed to apply proven methods of sentiment analysis used in English language to Turkish language; expand previous studies by applying different machine learning algorithms over different domains of data; and maximize accuracy.

First of all, thousands of labeled data were gathered from various domains, book, film, and shopping comments. Four new datasets were generated from the labeled data. Largest dataset contains nearly 5000 positive and 5000 negative comments.

A custom data preparation algorithm was developed for Turkish language in order to maximize the accuracy. Naïve Bayes, Support Vector Machines, Logistic Regression, and Decision Tree machine learning algorithms were applied on the generated data sets. Results show that Naïve Bayes is more successful than others with maximum 85.80% accuracy. Besides the accuracy, the execution time of Naïve Bayes algorithm is much faster than the others. More specifically, Naïve Bayes is 20 or 30 times faster than its nearest competitor, Support Vector Machines algorithm.

Some surprising results were also encountered. Positive comments about everything are 20-25 times more than negative comments. Accuracy is decreasing while n-gram size increases. Negation handling seems to have no effect. Thresholding has generally little positive effect on accuracy but improves execution time gradually on algorithms other than Naïve Bayes.

5.2 FUTURE WORK

Naïve Bayes algorithm with bag of words approach can be regarded as successful with 85% accuracy. Any achievement is not expected by tweaking machine learning algorithm. However, a hybrid methodology of supervised learning and linguistic analysis can achieve higher results. Sentence can be processed linguistically before bagging all words. Some words like adjectives or verbs can be more valuable for sentiment analysis. For instance their frequencies can be multiplied.

On the other hand negation handling methodology should be improved or changed because it had no effect over accuracy.

REFERENCES

[1] 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.

[2] Jim Jansen, Online Product Research, Pew Research Center's Internet & American Life Project, http://www.pewinternet.org/Reports/2010/Online-Product-Research.aspx

[3] Johan Bollen, Huina Mao, Xiao-Jun Zeng, Twitter mood predicts the stock market, 2010.

[4] Lee Rainie, Aaron Smith, Kay Lehman Schlozman, Henry Brady, Sidney Verba, Social Media and Political Engagement, http://pewinternet.org/Reports/2012/Political-engagement.aspx

[5] Sitaram Asur, Bernardo A. Huberman, Predicting the Future With Social Media, 2010.

[6] Umut Eroğul, Sentiment Analysis in Turkish, METU Master's Thesis, 2009.

[7] Jaime Carbonell. Subjective Understanding: Computer Models of Belief Systems. PhD thesis, Yale, 1979.

[8] 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.

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

[10] 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.

[11] Turney, Peter; Littman, M.L, Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus, NRC/ERB-1094. May 15, 2002, 9 pages.

[12] 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.

[13] Mihalcea, C. Banea and J. Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In Proceedings of ACL-2007.

[14] Banea, R. Mihalcea, J. Wiebe and S. Hassan. 2008. Multilingual subjectivity analysis using machine translation. In Proceedings of EMNLP-2008.

[15] Wan, X. 2009. Co-training for cross-lingual sentiment classification. In Proceedings of the ACL,

pages 235–243.

[16] Mitchell, T. (1997). Machine Learning, McGraw Hill. ISBN 0-07-042807-7, p.2.

[17] *Software For Predictive Modeling and Forecasting* , <http://www.dtreg.com>

[18] *Data Mining Software in Java*, <http://www.cs.waikato.ac.nz/ml/weka/>

[19] *Natural Language Processing library for Turkish*, <http://code.google.com/p/zemberek/>

[20] *Culture and art portal*, <http:// kitap.antoloji.com>

[21] *Cinema website*, <http://www.beyazperde.com>

[22] *Leader online shopping site*, <http://www.hepsiburada.com>

[23] *Wikipedia, the free encyclopedia*, <http://en.wikipedia.org>

[24] 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.

[25] *Website for statistical computation*, <http://vassarstats.net/>

[26] Cüneyd Murad Özsert, Word Polarity Detection Using A Multilingual Approach, BOUN Master's Thesis, 2012.

APPENDICES

“All opinions are not equal. Some are a very great deal more robust, sophisticated and well supported in logic and argument than others.”
― Douglas Adams

APPENDIX A

EXAMPLE DATA

A.1 Example Book Comments

- Çok ilginç konuların bir arada toplandığı güzel bir kitap olmuş.

- Hiç beğenmediğim kitplardan biriydi zaten doğru düzgün okuyamadım kötü içeriklerinden ötürü adeta midemi bulandırdı

A.2 Example Film Comments

- Şüphesiz serinin en güzeli ... İzlemeyen varmı hala merak ediyorum .. hele o eşsiz müziğiyle finali yokmu tekrar tekrar izleyesi geliyor insanın.

- Zaman kaybı arkadaşlar .. kesinlikle tavsiye etmiyorum..

A.3 Example Shopping Comments

- Ürünün minik adaptörü ile pil sorunu yaşamadan çalışması en sevdiğim tarafı oldu.

- Ürünün elime ulaşma hızı ile kalitesi aynı orantıda değil maalesef.

APPENDIX B

TOP 100 n-grams

B.1 Top 100 Unigrams

Table B.1: Top 100 Unigrams

unigram	frequency	unigram	frequency	unigram	frequency
bir	68461	uygun	4861	fakat	2992
cok	50542	harika	4764	zaman	2934
urun	18158	mukemmel	4759	ancak	2869
guzel	17844	hem	4606	son	2859
ama	17577	herkese	4584	filmi	2797
tavsiye	17338	fiyata	4480	oldugunu	2738
icin	17158	gun	4289	geldi	2730
iyi	15968	tesekkurler	4268	hepsi	2730
daha	15254	oldu	4246	ayni	2713
ederim	13387	degil	4187	ise	2698
kitap	12734	tek	4119	ozellikle	2692
aldim	9788	kaliteli	3930	fiyat	2638
kadar	9466	olmasi	3870	arkadaslar	2630
urunu	9286	bile	3840	oldugu	2628
ben	8683	ayrica	3810	kisa	2620
gibi	8585	kullanisli	3708	benim	2615
gercekten	8313	ediyorum	3613	iki	2599
bence	6959	fazla	3607	okudum	2538
olarak	6830	kullaniyorum	3601	sadece	2537
var	6686	siparis	3548	kolay	2504
sonra	6239	olan	3460	gerek	2500
yok	6075	diger	3284	ulasti	2494
film	6072	buyuk	3281	tam	2455
kitabi	5887	super	3251	verdim	2441
gayet	5773	telefon	3106	sekilde	2436
ile	5771	hemen	3087	yeni	2412
hic	5770	oldukca	3067	baska	2391
her	5684	zaten	3047	bana	2337
kesinlikle	5569	fiyati	3036	basarili	2248
elime	5268	uzun	3035	ses	2199
ilk	5239	diye	3008	kalitesi	2192
once	5085	icinde	3005	derim	2182
biraz	5084	memnun	2998		
gore	4996	hizli	2995		

B.2 Top 100 Bigrams

Table B.2: Top 100 Bigrams

bigram	frequency	bigram	frequency	bigram	frequency
tavsiye ederim	11877	urun cok	1107	ben cok	670
bir urun	7227	gayet iyi	1103	cok sik	660
bir kitap	6850	icin aldim	1059	fiyati cok	656
cok guzel	6173	kisa surede	1037	daha cok	641
cok iyi	5017	siddetle tavsiye	1010	urun gercekten	634
guzel bir	5000	icin cok	989	siparis ettim	634
herkese tavsiye	3311	hic bir	964	cok hizli	628
bir film	3232	aldim cok	960	cok basarili	609
tavsiye ediyorum	2479	kaliteli bir	939	surukleyici bir	600
iyi bir	2359	gore cok	926	daha fazla	585
gercekten cok	2242	cok cok	902	alin derim	574
harika bir	1835	urun elime	885	gun icinde	574
elime ulasti	1759	bir sey	866	kadar guzel	574
kesinlikle tavsiye	1615	cok kaliteli	863	urunu alali	568
elime gecti	1525	bir cok	860	gun sonra	562
cok uygun	1517	tesekkur ederim	859	benim icin	562
daha iyi	1506	icin ideal	850	daha guzel	557
mukemmel bir	1472	boyle bir	806	farkli bir	557
daha once	1449	ama cok	789	okunmasi gereken	557
hepsi burada	1391	bir roman	769	cok hos	551
cok memnun	1383	tek kelimeyle	747	muhtesem bir	545
bir sekilde	1382	baska bir	744	uzun sure	541
gereken bir	1275	buyuk bir	743	cok kisa	538
fiyatina gore	1254	son derece	730	gibi bir	533
cok begendim	1231	kullanisli bir	720	bence cok	526
memnun kaldim	1228	super bir	711	diye dusunuyorum	525
siparis verdim	1223	cok fazla	711	aydir kullaniyorum	522
bir telefon	1213	biraz daha	696	muthis bir	513
gayet guzel	1194	basarili bir	690	gercekten guzel	512
cok memnunum	1184	cok kolay	685	ayri bir	511
gerek yok	1181	olmasina ragmen	681	kacirmayin derim	500
oldugu icin	1172	bir kitapti	675	bir kac	495
cok daha	1146	bir eser	674		
cok kullanisli	1139	cok rahat	673		

B3. Top 100 Trigrams

Table B.3: Top 100 Trigrams

trigram	freq	trigram	freq	trigram	freq
herkese tavsiye ederim	2483	gunde elime ulasti	269	ideal bir urun	173
cok guzel bir	1777	cok kisa surede	268	cok ama cok	170
guzel bir kitap	1119	almaya karar verdim	267	cok basarili bir	168
kesinlikle tavsiye ederim	1026	almanizi tavsiye ederim	266	super bir urun	166
guzel bir urun	878	surukleyici bir kitap	261	aldim cok memnun	166
cok iyi bir	820	okumanizi tavsiye ederim	257	okumasi gereken bir	165
cok memnun kaldim	745	mukemmel bir kitap	256	cok tesekkur ederim	164
gercekten cok guzel	567	cok daha iyi	248	muhtesem bir kitap	164
gereken bir kitap	526	arkadaslara tavsiye ederim	245	urun herkese tavsiye	162
siddetle tavsiye ederim	521	cok hosuma gitti	241	kisa bir surede	162
herkese tavsiye ediyorum	495	isteyenlere tavsiye ederim	240	bir kitap tavsiye	161
kaliteli bir urun	487	urun gercekten cok	237	hediye olarak aldim	160
iyi bir urun	484	kullaniyorum cok memnunum	236	icinde elime ulasti	159
kullanisli bir urun	446	cok cok iyi	232	iyi tavsiye ederim	158
fiyati cok uygun	433	icin ideal bir	220	kalitesi cok iyi	157
kesinlikle tavsiye ediyorum	433	urun bugun elime	219	daha iyi bir	157
harika bir kitap	421	tavsiye ederim cok	219	bir urun herkese	155
tesekkurler hepsi burada	394	almak isteyenlere tavsiye	218	yaklasik aydir kullaniyorum	155
fiyatina gore cok	367	gayet guzel bir	217	kitabi okuduktan sonra	155
harika bir urun	363	gun sonra elime	214	bir urun cok	154
mukemmel bir urun	355	ben cok begendim	213	guzel bir telefon	154
okunmasi gereken bir	335	guzel bir kitapti	209	kullanimi cok kolay	154
urun tavsiye ederim	333	hizli bir sekilde	206	dusunenlere tavsiye ederim	153
bir urun tavsiye	333	gun icinde elime	203	tavsiye ederim tesekkurler	151
guzel bir film	328	aldim cok memnunum	201	guzel tavsiye ederim	151
gercekten cok iyi	312	cok kaliteli bir	201	hic dusunmeden alin	150
siddetle tavsiye ediyorum	299	iyi bir film	199	sonra elime ulasti	149
kisa surede elime	297	cok kullanisli bir	197	dusunmeden alin derim	148
hepsi burada com	293	bir kokusu var	191	alali hafta oldu	146
gore cok iyi	281	bir kez daha	189	bir kitap cok	145
tek kelime ile	280	tek kelimeyle harika	189	basarili bir urun	144
bugun elime gecti	274	surede elime ulasti	187	bir film olmus	144
urun cok guzel	273	tek kelimeyle mukemmel	181		
gereken bir urun	273	gercekten guzel bir	177		