DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

IMPROVING MACHINE LEARNING METHODS FOR SOCIAL MEDIA DATA IN TURKISH

by
Buket ERŞAHİN

January, 2021
İZMİR

IMPROVING MACHINE LEARNING METHODS FOR SOCIAL MEDIA DATA IN TURKISH

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Computer Engineering

by
Buket ERŞAHİN

January, 2021
İZMİR

Ph.D. THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **“IMPROVING MACHINE LEARNING METHODS FOR SOCIAL MEDIA DATA IN TURKISH”** completed by **BUKET ERŞAHİN** under the supervision of **ASSIST. PROF. DR. ÖZLEM AKTAŞ** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy.

Assist. Prof. Dr. Özlem AKTAŞ

Supervisor

Assoc. Prof. Dr. Deniz KILINÇ	Assoc. Prof. Dr. Zerrin IŞIK
Thesis Committee Member	Thesis Committee Member
Doç. Dr. Akın ÖZÇİFT	Prof. Dr. Ayşegül ALAYBEYOĞLU
Examining Committee Member	Examining Committee Member

Prof. Dr. Özgür ÖZÇELİK
Director
Graduate School of Natural and Applied Sciences

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my supervisor Assist. Prof. Dr. Özlem AKTAŞ and to my thesis committee member Assoc. Prof. Dr. Deniz KILINÇ and Assoc. Prof. Dr. Zerrin IŞIK for their support, guidance, supervision, and useful suggestions throughout this study. Their guidance helped me so much while writing this thesis.

Also, I would like to thank my parents for their love and support throughout my life. You are always there for me.

Then, I would like to thank my husband and love of my life, Mustafa, for his love and encouragement, without whom I would never have enjoyed so many opportunities. He has guided me and encouraged me to carry on through these years. I would also express my love to my children Mert and Burcu, who have been the light of my life who has given me the extra strength and motivation to get things done.

Finally, I am grateful to Scientific and Technological Research Council of Turkey (TÜBİTAK), who provided me full Ph.D. fellowship.

This thesis is dedicated to my grandmother and grandfather, who bring me up with endless love. I owe so much to them. May they rest in peace.

Buket ERŞAHİN

IMPROVING MACHINE LEARNING METHODS FOR SOCIAL MEDIA DATA IN TURKISH

ABSTRACT

In this thesis, we have presented a hybrid methodology, which combines the lexicon-based and machine learning (ML)-based approaches for sentiment analysis in Turkish. To use on the lexicon-based side, we have generated a sentiment dictionary by extending SentiTürkNet with a synonym dictionary, ASDICT. Besides this, we have tackled the classification problem with three supervised classifiers, Naive Bayes, Support Vector Machines, and J48, on the ML side.

Our hybrid methodology combines these two approaches by generating a new lexicon-based value according to our proposed feature generation algorithm and feeds it as one of the features to ML classifiers. We have experimented on three different datasets such as Movie, Hotel, and Twitter. Despite the linguistic challenges caused by the morphological structure of Turkish, the experimental results show that it improves the accuracy by 7% on average.

In conclusion, we have achieved these contributions in our study: It is the first hybrid approach for Turkish sentiment analysis. We have also adapted lemmatization in natural language processing for Turkish SA to preserve the positive and negative meanings of tokens. Finally, we have generated eSTN by extending STN, which is the first comprehensive polarity lexicon for Turkish.

Keywords: Sentiment analysis, opinion mining, social media, natural language processing

TÜRKÇE SOSYAL ORTAM VERİLERİ İÇİN MAKİNE ÖĞRENME YÖNTEMLERİNİN GELİŞTİRİLMESİ

ÖZ

Bu çalışmada Türkçe duygu analizi için sözlük ve makine öğrenmesi tabanlı yaklaşımları birleştiren hibrit (karma) bir yöntem geliştirilmiştir. Sözlük tabanlı kısımda kullanılmak üzere, SentiTürkNet eş anlamlılar sözlüğü olan ASDICT ile genişletilerek bir duygu analizi sözlüğü oluşturulmuştur. Bunun yanında, makine öğrenmesi tarafında Naïve Bayes, Support Vector Machines ve J48 adlı üç gözetimli öğrenme algoritması ile sınıflandırma sorunu çözülmüştür.

Hibrit yöntemimiz bu iki yaklaşımı özellik üretimi algoritmamızı kullanarak yeni bir sözlük tabanlı değer hesaplayıp ve bunu makine öğrenmesi sınıflandırıcılarına yeni bir özellik olarak ekleyerek birleştirmektedir. Film, otel ve Twitter olmak üzere üç farklı veri seti üzerinde sınamalar gerçekleştirilmiştir. Türkçe'nin morfolojik yapısından kaynaklı dilbilimsel zorluklara rağmen, deneysel sonuçlar çalışmamızın doğruluk oranını diğer çalışmalara göre ortalama %7 artırdığını göstermektedir.

Sonuç olarak, çalışmamızın katkıları şunlardır: Bu çalışma Türkçe duygu analizi için geliştirilmiş ilk hibrit yaklaşımdır. Ayrıca, pozitif ve negatif anlamı kaybetmemek için kök çözümleme algoritması iyileştirilmiştir. Son olarak, ilk kapsamlı polarite sözlüğü olan STN genişletilerek eSTN adında daha kapsamlı bir sözlük oluşturulmuştur.

Anahtar kelimeler: Duygu analizi, fikir madenciliği, sosyal medya, doğal dil işleme

CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER ONE
INTRODUCTION

1.1 General

During the last decade, the usage of social media applications has increased. With the spread of the Internet, people tend to use social media applications such as microblogging sites, social networks, and forums instead of newspapers and television. As a result of this, people have been active by sharing the information instead of being only observers. Twitter is one of the most used applications to share information especially the opinions and emotions about the news and products.

The amount of data shared by the active users on social media platforms are enormous; therefore, it is named big data. This data is also a collective intelligence created by the opinions of the users. It is not convenient to analyze and understand this big data. With the usage of social media, people's feeling on things has become available to everyone. Moreover, companies and organizations also need to be aware of their employees' and customers' feelings about their organizations. Human resources also would like to discover whether a potential employee will be loyal or leave after receiving training and benefits. Besides, the tweets about the candidates are used to predict the results of elections by the government. People read the customer reviews about the products and decide whether it is satisfiable or not for them. There is much usage of social media data like these.

Sentiment analysis (SA) is a text classification field that determines people's opinions and attitudes on different products, services, and topics. It is a discipline that started as a research topic in Natural Language Processing (NLP) in Computer Science and now transitioned to other departments like business and management schools since everyone wants to increase their profits and their customers' feelings.

1.2 Purpose

The increasing popularity of social media in recent years has led to the explosion of data on the Web. The activities of users of social networking and friendship sites (e.g., Facebook), blogging and microblogging sites (e.g., Twitter), content and media sharing sites (e.g., YouTube), and shopping sites (e.g., Amazon, AliExpress) generate huge amounts of data. As it is almost impossible to read and interpret all these data manually, SA is required to automate such an exhaustive process.

This thesis aims to develop a new hybrid SA tool combining lexicon-based and machine learning-based approaches that runs on different social media datasets for improving the results of machine learning-based sentiment analysis in Turkish.

To realize this research, we also need well-known datasets that are used in other successful studies. We have looked for these datasets and collected them to compare our experimental results with the studies using the same datasets.

Additionally, there is a need for a sentiment lexicon to realize the lexicon-based side of the hybrid algorithm. We also aim to expand the STN in order to create a new comprehensive sentiment lexicon. For this reason, we make use of ASDICT, which is a synonyms dictionary for Turkish.

In short, this thesis proposes a hybrid sentiment analysis framework to improve the results of ML-based sentiment analysis by supporting a new lexicon and specialized lexicon-based approach.

1.3 Novel Contributions of this Thesis

Our approach combines ML-based methods with lexicon-based methods as a hybrid approach and improves the results of SA. As far as we know, no previous research has investigated a hybrid approach in Turkish. In this study, we present a hybrid method for Turkish SA that is tested using three different datasets of Movie, Hotel, and Twitter. The main contributions of this study are as follows:

- To the best of our knowledge, it is the first study proposing and testing a hybrid SA method in Turkish.

- The first comprehensive Turkish SA dictionary, SentiTurkNet (STN) (Dehkharghani, Saygin, Yanikoglu, & Oflazer, 2015) is expanded using the Automated Synonym Dictionary (ASDICT) (Aktaş, Birant, Aksu, & Çebi, 2013).

- Lemmatization in natural language processing (NLP) is adapted for Turkish SA to preserve the positive and negative meaning of tokens.

As a result, in this thesis, (i) a new hybrid approach for sentiment analysis was proposed, (ii) a new tool for SA in Turkish is developed, (iii) a new polarity dictionary for SA, was introduced.

1.4 Organization of the Thesis

This thesis includes seven chapters, and the remaining of this thesis is organized as follows.

In Chapter 2, general information about related works, literature review, and field research about sentiment analysis are given.

In Chapter 3, background information about lexicon-based, machine learning-based, and hybrid sentiment analysis approaches are represented.

In Chapter 4, the lexicon expansion process, the dictionaries SentiTurkNet and ASDICT are explained in detail.

In Chapter 5, the new hybrid sentiment analysis approach is explained in its five main steps.

In Chapter 6, experiments were performed for the proposed hybrid sentiment analysis approach with well-known and widely used social media datasets.

Finally, in Chapter 7, the concluding remarks and future works are presented.

CHAPTER TWO
RELATED WORK

Sentiment analysis is the study of deciding opinions, emotions, and attitudes of people for an individual, events, or products. It is essential both for industry and academia. Customers' opinions are valuable for producers and consumers. Producers need it to improve their service or products. Customers benefit from it while they are deciding on an entity.

In this chapter, research projects, literature, and field reviews are explained, and research results are discussed. There is a lot of research in this field, however not many in hybrid Turkish SA in Turkish. According to the approach, the related work is categorized into four classes: ML-based, lexicon-based, hybrid approaches, and sentiment analysis in Turkish.

2.1 ML-based Approaches

In ML-based approaches, supervised techniques are mostly used. In the studies of Zhang et al. (2014), Chinese mobile reviews were used as a dataset. Statistical data analysis was applied only for mobile user reviews. This work showed that the mobile reviews have 17 Chinese characters on average, which are shorter than other short texts such as microblogs with 45 words on average. Labeling is done using iTunes scores. A series of experiments have been conducted to discover more appropriate methods for short texts. One of these experiments is polarity classification algorithms comparison. As a result of it, Naive Bayes (NB) and support vector machine (SVM) algorithms are selected, and the results show that NB is better than SVM. Another experiment is about the comparison of text feature representations. They applied the N-gram model and compared the results for N from 1 to 4. As a result of it, after the Chinese word segmentation, N-Gram is applied as N=2 and obtained the best result. The last experiment is on the influence of a different number of words in comments to find the difference between short and long texts. If the number of words in reviews is more than 150, feature extraction is necessary and improves the sentiment classification accuracy. As a summary, they found that the reviews have four properties: Short average length, large span of length, power-law distribution, and significant difference in polarity. So that, they have discovered that phone reviews are different from PC reviews.

Vinodhini & Chandrasekaran (2013) examined the effect of principal component analysis (PCA), which is used for feature reduction to improve the performance of learning algorithms. They experimented with PCA on SVM and NB algorithms. The high dimensionality of the features in the long texts raises problems in applying learning algorithms for text SA. Feature reduction aims to remove some irrelevant features. So that, they tend to improve the accuracy and decrease the running time of learning algorithms. The experiments were done on product reviews. The results were improved using PCA for feature reduction in both algorithms.

Pang, Lee & Vaithyanathan (2002) studied the effectiveness of applying ML algorithms for SA. He compared NB, SVM, and maximum entropy on a movie review dataset and showed that binary representation is better than frequency representation. He also found that NB has the worst and SVM has the best performance, but the difference is not large.

The Opinion Corpus for Arabic (OCA) was proposed by Rushdi Saleh et al. (2011). The corpus consists of 250 positive and 250 negative movie reviews collected from a variety of web pages. Various experiments were conducted on this corpus with NB and SVM to determine the polarity of reviews. They observed that the best result using SVM over the OCA improved on the best result obtained with the Pang corpus, using trigrams to generate the word vectors.

Govindarajan (2013) proposed a hybrid method coupling NB and a genetic algorithm (GA), and experimented on movie reviews. The results showed that GA has better performance than NB Besides, a comparison between the individual classifier and hybrid classifier shows that hybrid classifier has better performance than the other.

Duwairi (2015) tested Arabic tweets including dialectical words, with NB and SVM. Two versions of the dataset were studied; one was Tweets with dialectical words, and second was with dialectical words as translated. The accuracy of the dataset with translated dialectical words was 3% better.

2.2 Lexicon-based Approaches

Lexicon-based approaches need a lexicon, which is generated either from an existing dictionary or extracted opinion words from a corpus. According to the polarity values in the lexicon, the general sentiment of the document is predicted.

Baloglu and Aktas (2010) introduced an opinion-mining application, which creates movie scores from blog pages. They got the sentiment scores from SentiWordNet (SWN) (Esuli A, Sebastiani F., 2006) and declared that they produced accurate results close to IMDB results.

The document-based Sentiment Orientation System (Sharma, Nigam & Jain, 2014) uses WordNet (Fellbaum, 1998) to identify synonyms and antonyms, so it gives the summary of the total number of positive and negative documents. Negation is also handled in the system. That work classified the document as positive if the number of positive words is greater; otherwise, the polarity is negative. If the number of positive and negative words is equal, it is classified as neutral. They experimented on movie reviews and obtained an accuracy of 63%.

2.3 Hybrid Approaches

Hybrid approaches use lexicon-based and ML-based approaches in combination. The language processing operations are done before the learning of ML algorithms.

Appel et al. (2016) have shown that a hybrid method using NLP techniques, semantic rules, and fuzzy sets performed well on movie reviews and improved the results of NB and Maximum Entropy. Their hybrid approach achieved an accuracy of 76%. Another benefit of their proposed system is identifying different strengths in the polarity degree of the input sentences regarding a specific base-case. By utilizing fuzzy sets, they determine that a given sentence has a stronger or weaker intensity in terms of polarity than another one in the dataset.

Ohana & Tierney (2009) calculated the sentiment direction using SWN and then applied the SVM classifier. They presented the results of applying the SWN lexical resource to the SA of film reviews. Their approach involves positive and negative term scores to determine sentiment, and they presented an improvement by building a dataset of relevant features using SWN as a source. Then they applied ML algorithms. The results indicated that SWN can be used as an important resource for SA. They obtained the best accuracy of 69.35%, with SWN scores used as features.

2.4 Approaches on Sentiment Analysis in Turkish

Most of the research in the SA field focuses on English. There are a few works on SA on Turkish.

Akgul et al. (2016) compared the results of lexicon-based and character-based n-gram models. The dataset consists of the tweets gathered with a keyword and labeled as positive, negative, and neutral, suitable for both lexicon and n-gram models. They preprocessed their Twitter dataset and ran n-grams. As a result, the lexicon method obtained an accuracy of 70% and the n-gram model 69%, respectively.

Turkmenoglu & Tantug (2014) made a comparison of lexicon-based and ML-based approaches. After some preprocessing on Twitter and movie datasets, they obtained an accuracy of 75.2% on Twitter and 79% on the movie dataset by the lexicon-based approach. On the other hand, the best accuracy results of the ML-based approach were 85% for the Twitter dataset by SVM and 89.5% with SVM and NB on the movie dataset.

Oğul & Ercan (2016) performed their experiments on hotel reviews with NB, SVM, and random forest (RF) algorithms. According to the results of the experiments, using the document term matrix as input gives better results than the TFIDF matrix. They also observed that best results are obtained with RF classifier with the Area Under Curve (AUC) metric 89% on both positive and negative comments.

Kaynar et al. (2016) conducted their experiments on a Twitter dataset with NB, center-based classifier, multi-layer perceptron (MLP), and SVM. According to the results, the best performance was achieved with MLP and SVM with accuracy values of 86% and 81% on the movie review dataset, respectively. Also, it is seen that neural networks and SVM outperforms with both training and test sets.

Yildirim et al. (2017) experimented on Tweets in the telecommunication area. NLP was used for normalization, stemming, and negation handling. Ternary classification was achieved with an accuracy of 79% using SVM. It is the first paper in the literature that investigates and reports the impact of the natural language preprocessing layers on the SA of Turkish social media texts.

Çoban, Özyer & Özyer (2015) employed a Twitter dataset with some different classification algorithms: SVM, NB, multinomial naive Bayes (MNB), and kNN. The features represented by vector space are extracted from two different models: Bag of Words and N-Gram. The results showed that the best accuracy was achieved with MNB at 66.08%.

Vural et al. (2014) done their studies on Turkish movie reviews using unsupervised learning techniques. They used SentiStrength (Thelwall, Buckley & Paltoglou, 2012) to classify the texts by translating them. The results showed that the accuracy was 76% for binary classification. Although their framework is unsupervised, they obtained good accuracy approaching the performance of supervised polarity classification techniques.

Kaya et al. (2012) observed the sentiment analysis of Turkish political news. They used four different classifiers: NB, maximum entropy (ME), SVM, and character-based n-gram models. Their experimental results showed that ME with the n-gram language model was more effective than SVM and NB. An accuracy of 76% was achieved in binary classification of political news.

Boynukalın (2012) studied emotion analysis on Turkish texts by using an ML-based approach. It is the first study on emotion analysis of Turkish texts. Another important contribution is the generation of a new data set in Turkish for emotion analysis. It is generated by combining two types of sources. Several classification algorithms are applied to the dataset, and results are compared. Four types of emotions, joy, sadness, fear, and anger were examined on her own dataset, and an accuracy of 78% was achieved.

Eroğul (2012) developed a system for SA of movie reviews in Turkish. Their approach combined supervised learning and lexicon-based approaches as hybrid approach used a recently constructed Turkish polarity lexicon called SentiTurkNet. They investigated the contribution of different feature sets, as well as the effect of lexicon size on the overall classification performance. They also investigated the impact of part-of-speech (POS) tags, word unigrams and bigrams, and negation handling. NLP processing was done with Zemberek, obtaining an accuracy of 85% on the binary classification of Turkish movie reviews.

Dehkharghani et al. (2017) proposed and evaluated a SA system for Turkish. Their system used STN and NLP techniques such as dependency parsing. They also covered different levels of granularities with some linguistic issues such as conjunction and intensification. Their system was evaluated on Turkish movie reviews, and the obtained accuracies ranged from 60% to 79% in ternary and binary classification. In the aspect level classification, they compared the proposed method with a baseline, which considers only the neighbor words around an aspect in a window with a size of five words as two words from each side. They also experimented with their approach with the restaurant dataset, a benchmark in SemEval-2016, including 1233 sentences, and each sentence includes at least one aspect word. In the sentence and document level classifications, polar words are the most effective features, but they are ineffective in ternary classification. Generally, the system is more successful in classifying positive sentences and documents than negative or neutral ones because positive opinions are written more clearly than negative ones. We have used the STN lexicon and enhanced it to enlarge its scope.

CHAPTER THREE
BACKGROUND INFORMATION

In this chapter, we are going to explain and discuss background information about SA basics and details of lexicon-based, ML-based, and hybrid approaches used for SA.

3.1 Sentiment Analysis Basics

SA is studied at four granularity levels: document level, sentence level, word level, and aspect level. Depending on how detailed one would like to get into an issue one would have to decide the level of analysis. The levels of SA are shown in Figure 3.1.

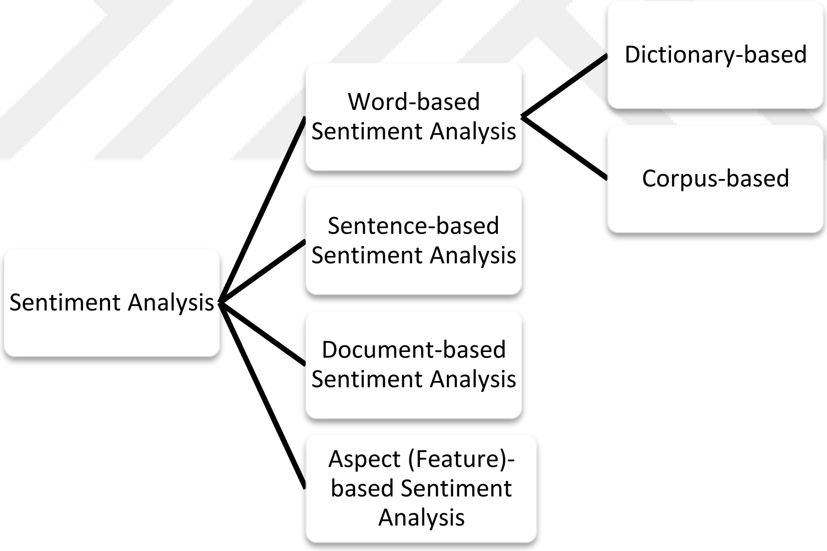

Sentiment Analysis
Word-based Sentiment Analysis
Sentence-based Sentiment Analysis
Document-based Sentiment Analysis
Aspect (Feature)-based Sentiment Analysis
Dictionary-based
Corpus-based

Figure 3.1 Types of SA

At the document level, the full text is considered an atomic unit and is assigned to a positive, negative, or neutral class. Some assumptions should be made before applying this level. For example, all objects in the document and the opinion holder of each sentence in the document should be same. There will be a problem if there are more than one opinion and several opinion holders in the document. We have developed our study in this level because almost every document in the dataset has only one opinion holder and object, according to our inspection.

At the sentence level, a sentence is identified as objective or subjective (holding an opinion). If it is subjective, it is assigned to a class; otherwise, it is ignored. The main challenge in this level is to differentiate the sentiment of some short sentences. The sentence-level and document-level SA approaches cannot discover more than one sentiment in a sentence or document.

At the word level, mostly the adjectives in the text are focused. Nevertheless, not only adjectives have an opinion, nouns, adverbs, and verbs may also carry opinions. Word level SA is either corpus-based or dictionary-based. We have also utilized this level in our study as we evaluate the words' sentiments to decide the sentiment of the document. To refer, we have built a sentiment dictionary since there is not any comprehensive sentiment dictionary in Turkish. The details of how we create it is explained in Section 4.

Aspect (feature) level SA can discover different sentiments in a text with their related targeted terms. It can identify opinion tuples, which consist of a target term, target attribute (feature), and target sentiment (Boudad, Faizi, Rachid & Chiheb, 2018). A good example of this level is a customer who has both positive and negative reviews about a product. As expected, the customer may like some features and dislike other features of a product. The other levels cannot handle such a review text, and it is preferred in this condition.

Considering the literature, methods used for SA are divided into three categories as ML-based approaches, lexicon-based approaches, and hybrid approaches as depicted in Figure 3.2 (Maynard & Funk, 2012). In ML-based approaches, some famous ML algorithms are applied to predict the sentiment. ML-based text classification approach can be divided into two categories such as supervised and unsupervised learning. Supervised methods need a training dataset which should be labeled. The unsupervised methods are used when there is no labeled dataset.

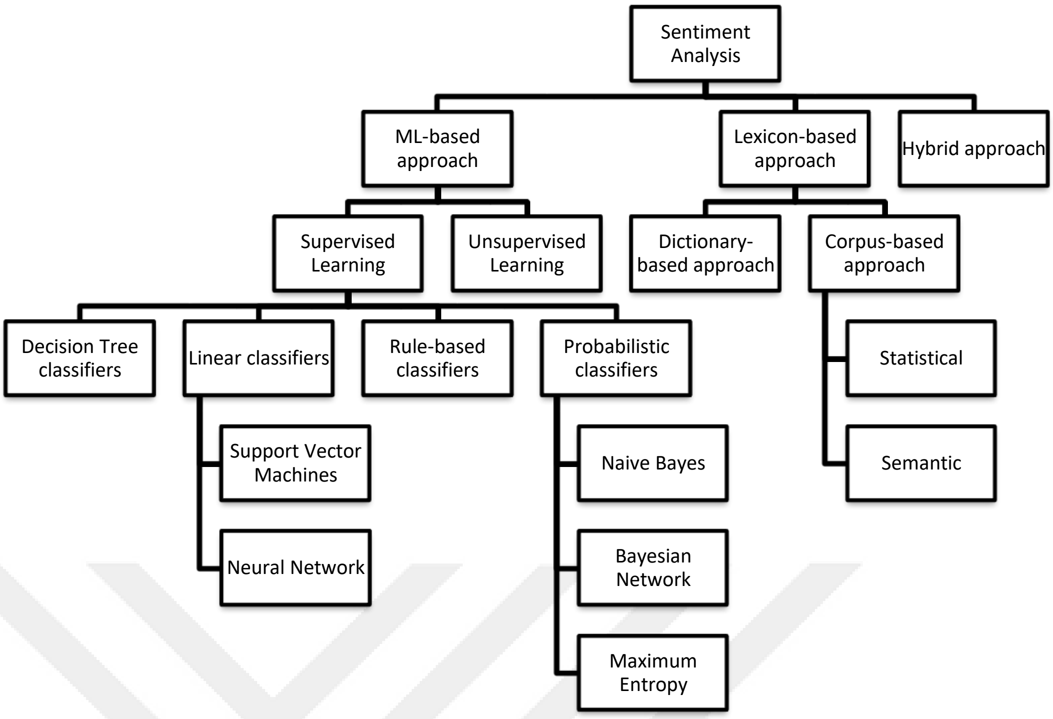


Figure 3.2 Sentiment Classification Techniques

On the other hand, the lexicon-based SA approach relies on sentiment lexicons to analyze the text. As for the hybrid methods, combination of the ML-based and lexicon-based approaches are used to improve the accuracy above both the single approaches. The details of each approach are given below.

3.2 ML-based Approaches

ML-based approach applies ML algorithms and uses linguistic features. It is divided into two categories: supervised and unsupervised learning. Supervised learning methods make use of labeled training datasets. Unsupervised learning methods are used when there is not any labeled dataset available. SA is a text classification problem which uses linguistic and syntactic features for ML algorithms.

3.2.1 Supervised Learning

In the following subsections, we explain some classification algorithms which are generally used in this field.

3.2.1.1 Probabilistic Classifiers

A probabilistic classifier does not only output the most likely class of the observation. It can predict a probability distribution over a set of classes. They are useful when combining classifiers into ensembles or on their own. We introduce three of them.

3.2.1.1.1 Naïve Bayes Classifier (NB). NB is the most used classifier. It is simple and computes the probability of a class based on the distribution of the words in the document. It ignores the position of the terms as in the Bag of Words feature extraction. NB is Bayes theorem (Equation 3.1) to predict the probability.

$$P(label|features) = \frac{P(label) * P(features|label)}{P(features)} \tag{3.1}$$

P(label) is the likelihood that a random feature set the label. P(features | label) is the probability that a given feature set is classified as a label. P(features) is the probability of a given feature set is occurred.

3.2.1.1.2 Bayesian Network (BN). BN model is a directed acyclic graph whose nodes represent random variables, and the edges represent the conditional dependencies. There is a complete joint probability distribution over all variables in the model. The computation complexity of BN very expensive, so it is not widely used.

3.2.1.1.3 Maximum Entropy (ME). The Maxent Classifier is a conditional exponential classifier, which converts labeled feature sets to vectors through encoding. This vector is then used to calculate weights for each feature. Combining these weights determines the most likely label for a feature set. It is parameterized by a set of X{weights}, which combines the joint features generated from a feature set by an X{encoding}. The encoding maps each C {(feature set, label)} pair to a vector.

3.2.1.2 *Linear Classifiers*

Given $\vec{X}$ is the normalized document word frequency and $\vec{A}$ is a vector of linear coefficients with the same dimensionality with the feature space, and b is the output of the linear predictor. The calculation of the predictor is the output of the linear classifier. The predictor is calculated according to Equation 3.2, and it is a hyperplane between classifiers.

$$p = \vec{A} \cdot \vec{X} + b \tag{3.2}$$

There are many linear classifiers. We have explained here Support Vector Machines and Neural Network.

3.2.1.2.1 Support Vector Machines (SVM). SVM algorithms aim to find a hyperplane in an N-dimensional space. There are many hyperplanes to separate classes. The objective is to find the plane with a maximum margin, which is the distance between data points of classes. SVM model determines the linear separators in the search space, which can best separate the classes. Text data is suitable for SVM because it is sparse.

3.2.1.2.2 Neural Network. NNs are multi-layer networks of neurons. $\vec{X}$i is the word frequencies in the ith document and $\vec{A}$ is a set of weights related to each neuron. The linear function of NN is pi given in Equation 3.3.

$$p_i = \vec{A} \cdot \vec{X_i} \tag{3.3}$$

NNs imitate the function of the human brain. The output of one layer is the input to the following layer. It can adapt to changing input and is used in various applications such as finance and marketing, especially for fraud detection and risk evaluation.

3.2.1.3 Decision Tree Classifiers

Decision tree classifier is one of the modeling approaches for predicting in the field of statistics. It creates classification or regression models in the tree structure. It breaks data into smaller parts so that the tree is created. It does not require domain knowledge, and it is popular. While deciding the nodes, the condition is the presence or absence of the words. The decomposition is done recursively until the leaf nodes contain minimum numbers of records used for classification.

3.2.1.4 Rule-based Classifiers

In rule-based classifiers, the knowledge is obtained in the form of rules from the model. It is suitable for the data containing both numerical and qualitative attributes. It makes use of IF-THEN rules for classification. The left side of the rule represents a condition on the feature set, while the right side is the class label. The conditions are generally on the presence of terms because rules on term absence are not meaningful for sparse data.

3.3 Unsupervised Learning

Text classification is the process of assigning classes to a text according to its content. Labeled training documents are required for supervised learning. Sometimes it is difficult to find such labeled documents, and it is impossible to apply supervised learning techniques for classifying them. In this way, it is easy to find unlabeled documents. Therefore, the unsupervised learning methods overcome these difficulties. Clustering and association are two types of unsupervised learning.

3.4 Lexicon-based Approaches

Opinion words are utilized to find the sentiment of a text. They are usually adjectives in the sentences. Positive opinion words express desirable and negative opinion words express undesirable conditions. There are some approaches to compile or collect the opinion word list. Managing them manually is very time-consuming. Therefore, it is usually used together with two other automated systems to avoid the mistakes resulting from automated methods. These two automated approaches are explained in the following subsections.

3.4.1 Dictionary-based Approach

In the dictionary-based approach, the process starts with selecting some opinion words, which are collected manually. This set is then expanded by searching and including the synonyms or antonyms of the selected words in different popular corpora or thesaurus such as WordNet or SentiTürkNet. New words are appended in the word list and this iterative process is repeated until any new words are discovered. After the process is ended it is possible to check the consistency and errors manually. The success of the approach is dependent on the scope of the dictionaries.

3.4.2 Corpus-based Approach

The corpus-based approach enables us to create a context-specific lexicon of opinion words. There are some methods for creating corpus for SA. These are label propagation, domain adaptation, pointwise mutual information, matrix factorization, polar phrase extraction, social media hashtags and emoticons, and conjunction rules on adjectives. It has the advantage of creating a domain or context-specific lexicon. Also, it can capture informal terms and slang words. However, it has some disadvantages. It is not efficient for formal texts and computation intensive.

3.5 Hybrid Approaches

The hybrid approach is a combination of the ML-based approach and lexicon-based approach (Medhat, Hassan & Korashy, 2014). A common strategy used to study SA is to apply either ML-based or lexicon-based approaches. On the other hand, some studies try to apply both, but not together in a hybrid approach, and compare the results of them. There are also studies which use hybrid methods combining lexicon-based and ML-based approaches. Our framework is classified in this category.

CHAPTER FOUR
LEXICON EXPANSION

Polarity lexicons are used to estimate the sentiment polarity of a review based on the polarities of the words constituting it. Some of the studies use dictionaries, and some of them use corpus. Corpus is domain-specific, but dictionaries are general.

General-purpose lexicons such as SentiWordNet are domain-independent and have shortcomings that they cannot handle different aspects and cannot differentiate domains and nations, but they are fast and scalable. For example, the word “big” is positive for hotel rooms’ size, but it is negative if used for battery size in the camera.

It is difficult to keep these lexicons up-to-date manually. Some automatic techniques are required for it. Another approach for building polarity lexicons for languages other than English is translating it from English. Another approach is starting with seed words and expanding them with their synonyms. We have used this approach in our study and used SentiTürkNet as seed words. Then, we have used ASDICT synonyms dictionary and expand our polarity lexicon. We give the details of these dictionaries in the following subsections.

4.1 SentiTürkNet

STN proposed a semi-automatic approach for assigning the polarity values to the Turkish WordNet synsets, which has only the polarity classes. Their method uses the information from Turkish WordNet and the polarity strength from SentiWordNet. They applied this method for Turkish, but it can be used for any language. Then, they controlled the assigned polarity results using three different methods. They show that their results are better than only translating the words directly from SentiWordNet. It is the first sentiment polarity for Turkish evaluating its accuracy.

4.2 ASDICT

Automated Synonym Dictionary Generation Tool for Turkish (ASDICT) is a synonym dictionary gathered by applying it on the data of Contemporary Turkish Dictionary published by Turkish Linguistic Association (TDK: Türk Dil Kurumu).

The synonym dictionary generation process was accomplished in four steps. As a result of these steps, the definite synonyms were classified as Definite Synonym (Dn) and inserted into the Synonym List (SLi). Some of the words were not classified as Dn. They were classified as Ambiguity and stored in a file called Ambiguity File (AF). Then, they were controlled by supervised methods to build a more reliable synonym database. The synonym database for Contemporary Turkish Dictionary, called Definite Synonyms Database (DSDB), was constructed by applying ASDICT.

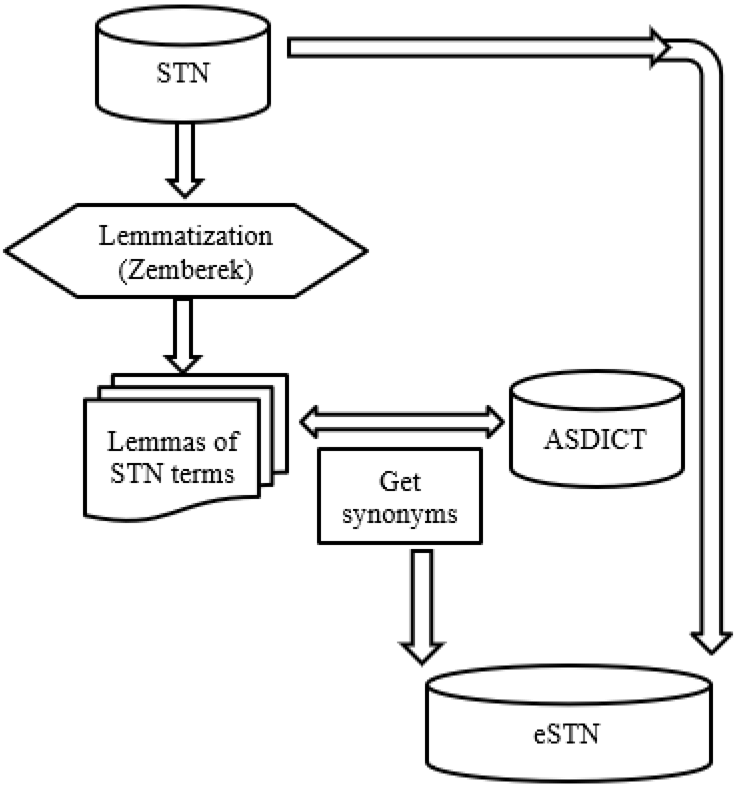


Figure 4.1 Expansion methodology

4.3 Expansion Methodology

We have used SentiTürkNet and ASDICT as resources to create our new lexicon. The steps of expansion methodology are given in Figure 4.1. We have started with the terms in STN and selected these words as seed words. Then, we have lemmatized them in Zemberek to reduce the diversity of terms because of Turkish's agglutinative feature and increase the matching ratio of the words. After lemmatization, we have searched the lemmatized words of STN in ASDICT and appended the synonyms of the words with the same polarity value. This process aims to create eSTN. A simple scenario of lexicon expansion is also given in Figure 4.2. In this example, a seed word is selected as "acemi". Then, the positive and negative polarity values are taken from STN. Next, synonyms of "acemi" are searched in ASDICT, and "amatör" is found which does not exist in STN. Therefore, this term is added in eSTN with the same polarity values of "acemi". If it were a word with suffixes such as "acemilik" then it would be lemmatized with Zemberek to be able to match it with synonyms.

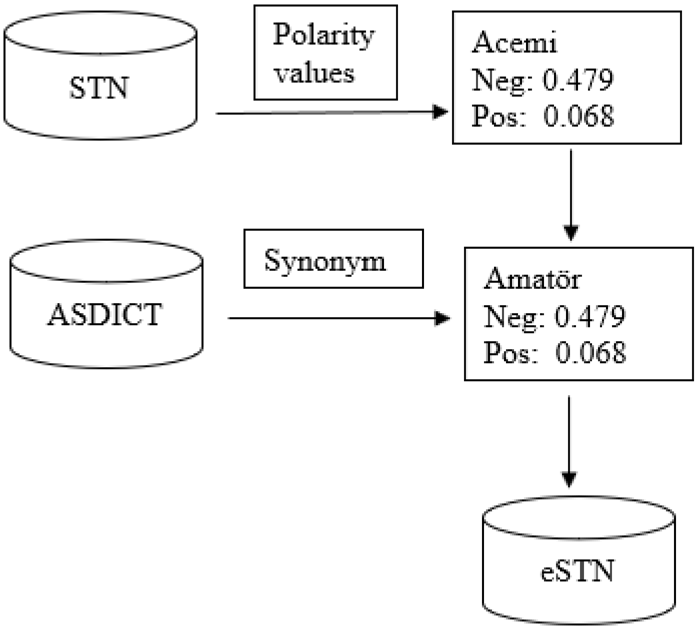


Figure 4.2 Lexicon expansion scenario

We have also defined some rules to make the matching possible. For example, all verbs are translated in their infinitive form. Inflection suffixes in verbs are so many that they produce many new features. To translate them into their infinitive form, first they are lemmatized by Zemberek. Then -mek or -mak is appended at the end of them so that they are standardized into a simpler form.

CHAPTER FIVE
A NEW HYBRID SENTIMENT ANALYSIS TOOL

In this chapter, our hybrid approach is explained in detail. In Section 5.1, dataset collection, in Section 5.2 preprocessing and used tools, in Section 5.3, lexicon expansion and its statistical results, in Section 5.4, our feature selection with lemmatization and lexicon usage, in Section 5.5, our feature generation algorithm, and in Section 5.6, selected ML algorithms and user interface of our tool is presented.

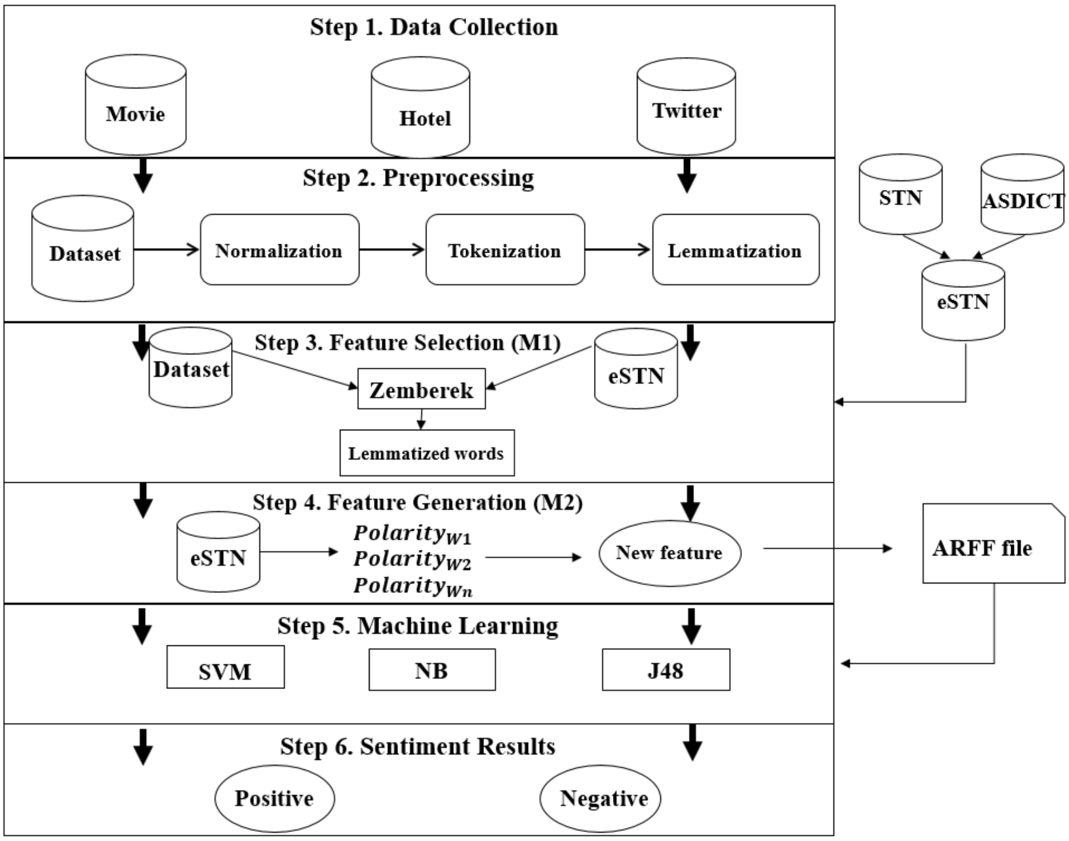


Figure 5.1 Our proposed framework

Our approach has two main steps: In the first stage, a lexicon score is calculated according to the polarities of the words which compose the document. In the second stage, this polarity is added as a new feature according to the lexicon score and the ML algorithms are learnt. Figure 5.1 shows the proposed hybrid method that aims to improve the accuracy of ML algorithms for SA by feeding them with a new lexicon-based feature. We apply five main steps, which consist of data collection, preprocessing and lexicon expansion, feature extraction with lemmatization (M1), polarity-based feature generation (M2), and ML. All steps are presented in the following subsections.

5.1 Data Collection

As for the dataset, we have chosen three different datasets. Two of them are customer review, and one is Twitter dataset. Customer review datasets are about hotel and movie reviews. In the beginning, we have started the research with customer review datasets. Then, to analyze our method's effect on short and misspelled words, we included Twitter dataset.

Table 5.1 Sample texts from datasets

Datasets	**Text**
Hotel	pislikten kalamadik tatilbudur com mikaturdan da kisilik gecelik rezervasyonumuz sonucunda kerasus hotele gittik katta bir oda verildi tambir pislik yuvasi halindeki oda nem kokuyor havlular sapsari tuvalet berbat yatak ve koltuk berbatti kalmadan dk icinden otelden ayrildik resepsiyon muduresi adi altinda bir bayan gecelik ucreti kredi kartimdan kestirmis yetkili birinle gorusmeye calistim fakat muhatap bulamadim kesinlikle kerasus otele ve tatilbudur com mikatur a inanmayin paranizla rezil oluyorsunuz bogle tatil olmaz olsun
Movie	Rahmetli Kemal Sunal demek buradan almış senaryoyu. İzlerken ne büyük keyif aldığımı tarif edemem. Ne kadar muhteşem bir film ya. O dönemler de böyle filmler yapılıyormuş işte. Seni asla unutmayacağız Chaplin...
Twitter	#vodafone icin Buyuk eksiklik. Apple urunlerinde kullanilmak uzere #turkcell online islem gibi bir app yapmadilar, yapamadilar.

In this study, experiments are done by using three different datasets to evaluate the results of the methodology on different types of data. Movie review and hotel review datasets are downloaded from the Hacettepe University Multimedia Information Retrieval Group's website. Movie reviews on this website are collected from beyazperde.com and hotel reviews are collected from otelpuan.com. All extracted movie reviews are rated by their authors according to stars. One or two stars is classified as negative, while 4 or 5 stars is classified as positive. Similarly, hotel reviews are rated between 0 and 100 instead of stars. The negative reviews are selected from 0 to 40-point reviews and the positive from 80 to 100-point reviews. A completely different dataset consisting of Tweets is also used in the experiments to control the accuracy of the proposed methodology. This dataset is taken from the website of the Kemik NLP group of Yıldız Technical University. It consists of 3000 Turkish tweets having three classes for SA as positive, negative, and neutral. We have chosen these datasets to compare their experimental results with ours and due to their balanced nature. Some text examples from the datasets are given in Table 5.1. As it is seen, hotel reviews are long and have more negative words. Movie reviews are shorter than hotel reviews. Twitter texts are about telecommunication. They are short texts and have many misspelled words, mentions, retweets, emoticons, and abbreviations.

5.2 Preprocessing

For a given dataset, the first step of SA is preprocessing, which involves a series of methods to improve the following phases. First, we normalize the input document utilizing the ITU NLP tool (Eryiğit & Torunoğlu-Selamet, 2017) and break it into tokens by using Zemberek. Then we lower the tokens to prevent mismatches because of case sensitivity. We also remove the tokens shorter than two characters to reduce the stop words.

5.2.1 ITU NLP Tool

ITU NLP tool is an NLP platform developed by the NLP group of Istanbul Technical University. It operates as Software as a Service, and it enables the researchers to do many NLP operations such as preprocessing, syntax, and entity recognition. The users can use this platform by file uploads, web interface, and Web APIs. The tool provides these components: Tokenizer, Deasciifier, Vowelizer, Spelling Corrector, Normalizer, isTurkish, Morphological Analyzer, Morphological

Disambiguator, Named Entity Recognizer, and Dependency Parser. We have used the Normalizer component in our method.

5.2.2 Zemberek

Zemberek is an NLP library for Turkish written with Java. It has many functions such as morphological analysis, disambiguation, word generation, tokenization, sentence boundary detection, spell checker, normalization, named entity recognition, text classification, and language identification. It is open source and can be enhanced. In our study, we have made use of it for tokenization as it is free.

5.3 Lexicon Expansion

STN, the lexicon used in our study, is the first comprehensive polarity lexicon for Turkish, and it is constructed using a semi-automatic approach. It is based on Turkish WordNet (Ehsani, Solak & Yildiz, 2018) and is mapped to both SentiWordNet and WordNet. It contains polarity values for all 15,000 synsets of Turkish WordNet, but the coverage size is small. To improve the performance of our matching process for lexicon-based feature generation, ASDICT is explored and utilized. The basic data source used in ASDICT is the Contemporary Turkish Dictionary (CTD), which includes more than 70,000 words and was published by the Turkish Linguistic Association (Turkish abbreviation: TDK). Supervised methods are used to generate a reliable synonym dictionary and handle the ambiguities arising from the different meanings of words. For the synonym dictionary, all ambiguities are examined and finalized by the experts of the TDK and the College of Social Sciences and Literature of Dokuz Eylül University (DEU). In our lexicon expansion step, all words in ASDICT are searched in STN. If there is a match, the synonyms are added to STN with the polarity values that are already in STN. The new lexicon is called extended STN (eSTN). Objective terms are excluded from eSTN because binary classification is the goal. Multiword terms are also removed since our features are words as unigrams. The coverage rates of STN and eSTN are compared on all datasets after applying the lemmatization to evaluate the effectiveness of the expansion process.

Table 5.2 Coverage rates of lexicons

Coverage rates	STN	eSTN	Increase
Movie	449	685	53%
Hotel	415	780	88%
Tweet	145	284	96%

According to the results given in Table 5.2, the performance of eSTN varies depending on the type and size of the dataset. The average increase in the coverage rate is approximately 78%.

5.4 Feature Selection with Lemmatization

After preprocessing, the datasets and eSTN are lemmatized using Zemberek. The aim of the lemmatization is to convert the word into a standard format by removing sentimentally insignificant suffixes. In this way, the number of tokens is reduced. Lemmatization is done by preserving negations in the word. For this, Turkish suffixes such as -me/-ma and -sız/-siz are conserved. The verbs are also translated into infinitive form, as seen in Table 5.3.

Table 5.3 Term lemmatization example

Term before lemmatization	Term after lemmatization
Akılsız	akılsız
anlaşmazlık	anlaşmamak
beğenilmeyen	beğenmemek
dumanlı	dumanlı
gürültülü	gürültülü

The main challenge of text classification is dealing with a massive number of tokens. They prolong the learning time and affect the ML algorithms' performance negatively. Feature extraction with our lemmatization approach is proposed to overcome this problem. It is implemented by lemmatizing tokens of texts and eSTN terms, and it also reduces the dimensionality, as seen in Table 5.4.

Table 5.4 Feature extraction with lemmatization

Before lemmatization	Kesinlikle izlenip desteklenmesi gereken müthiş bir film konu olarak orjinal bir film olduğunu da söylemeliyim (16 tokens)
After lemmatization	Kesin izlemek desteklemek gerek müthiş film konu olmak orjinal film olmak söylemek (12 tokens)
After feature selection	Kesin desteklemek gerek müthiş orjinal (5 tokens)

5.5 Polarity-based Feature Generation

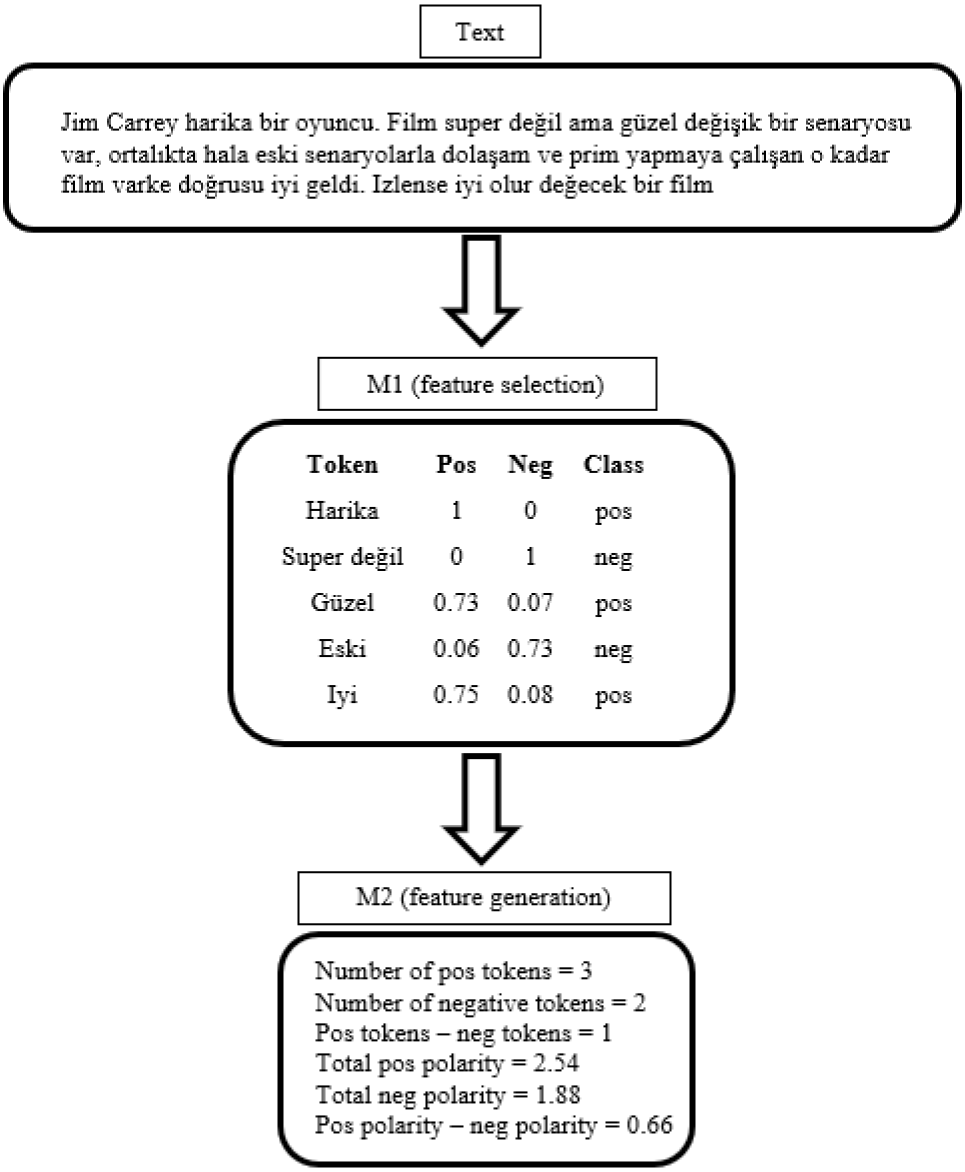
One of the contributions of this thesis is the generation of a new polarity-based feature, which improves the results significantly. In the feature extraction step, the tokens are lemmatized. In this step, the lemmatized tokens of a document are searched in eSTN and matching tokens are used to create the polarity-based feature. The number of positive tokens and the number of negative tokens is calculated using eSTN, and the value of the new feature is calculated considering the algorithm in Figure 5.2.

```
Feature Generation Algorithm
Input : S1 - Document as String
Output: polarity_prediction - predicted sentiment class

procedure GENERATE_FEATURE ( S1 )
BEGIN
  $polarity\_score_{S1} \leftarrow 0$ //initialize polarity score
  for $i \leftarrow 0$, numberOfTokens do
  if ($S_1[i]$ is positive) then  // result of STN matching
     $pos_{S1} \leftarrow pos_{S1} + 1$     // number of positive tokens
     $polarity\_score_{S1} \leftarrow polarity\_score_{S1} + polarity_{S1[i]}$
  else if ($S_1[i]$ is negative) then
     $neg_{S1} \leftarrow neg_{S1} + 1$     // number of negative tokens
     $polarity\_score_{S1} \leftarrow polarity\_score_{S1} - polarity_{S1[i]}$
  if ( $pos_{S1} - neg_{S1} >= 2$ ) then
     $polarity\_prediction_{S1} \leftarrow$ pos
  else if ( $neg_{S1} - pos_{S1} >= 2$ ) then
     $polarity\_prediction_{S1} \leftarrow$ neg
  else
     if ($polarity\_prediction_{S1} < 0$) then
        $polarity\_prediction_{S1} \leftarrow$ neg
     else
        $polarity\_prediction_{S1} \leftarrow$ pos
  return $polarity\_prediction_{S1}$
END
```

Figure 5.2 Feature Generation Algorithm

As seen in Figure 5.3, the proposed feature generation algorithm takes the text as input and creates the lexicon-based new feature as output. After preprocessing and feature extraction, “harika”, “süper değil”, “güzel”, “eski”, and “iyi” are the selected features for the given example text. As it was mentioned, when “değil” is encountered, the polarity value of the token just before it is negated. This means the values of negative and polarity scores are interchanged, as in Table 5.5. Then the polarity values and class labels of the tokens are taken from eSTN and processed according to our proposed algorithm. Based on the results of the algorithm, the



Jim Carrey harika bir oyuncu. Film super değil ama güzel değişik bir senaryosu var, ortalıkta hala eski senaryolarla dolaşam ve prim yapmaya çalışan o kadar film varke doğrusu iyi geldi. Izlense iyi olur değecek bir film

Token	Pos	Neg	Class
Harika	1	0	pos
Super değil	0	1	neg
Güzel	0.73	0.07	pos
Eski	0.06	0.73	neg
Iyi	0.75	0.08	pos

Number of pos tokens = 3
Number of negative tokens = 2
Pos tokens – neg tokens = 1
Total pos polarity = 2.54
Total neg polarity = 1.88
Pos polarity – neg polarity = 0.66

Figure 5.3 Feature Generation Scenario

number of positive tokens, the number of negative tokens, the difference between them, total positive polarity, total negative polarity, and the difference between them are calculated. Since the difference between positive tokens and negative tokens is not greater than or equal to 2 in this example text, the difference between positive polarity values and negative polarity values is calculated. It is found as 0.66, and since it is a positive value, a new feature is generated as positive.

Table 5.5 Handling negations

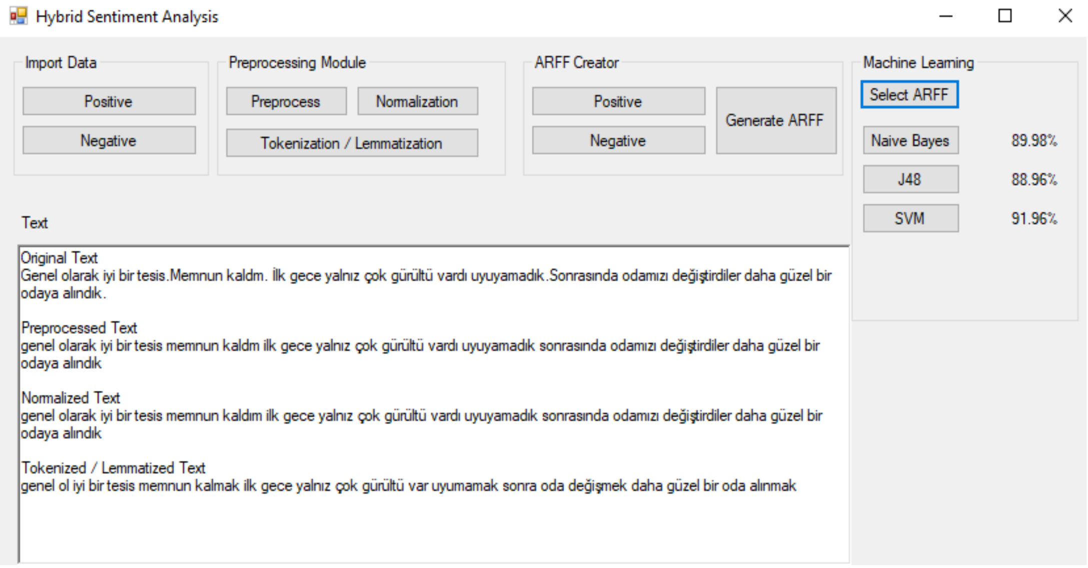
Term	Positive polarity	Neutral polarity	Negative polarity
güzel (beautiful)	1	0	0
güzel değil (not beautiful)	0	0	1
fena (bad)	0.035	0.02	0.945
fena değil (not bad)	0.945	0.02	0.035

The threshold value in this algorithm is selected with Grid search (Thisted RA., 1988). It is a technique that scans the data to configure the optimal parameters for a given model and works in an iterative way. In our model, we experiment with parameters 1 to 3. The grid search iterates through each of them and compares the result for each value. To evaluate the results, NB is selected as ML algorithm, and all configurations are run on all datasets. 5-fold cross-validation is selected because it is computationally intensive. The results are evaluated for accuracy. The average accuracy values for all datasets are 87.25% for parameter=1, 87.43% for parameter=2, and 87.36% for parameter=3, respectively. It finds the best parameter as 2 for our model.

5.6 Machine Learning

As the last step of our proposed approach, we have run NB, J48, and SVM algorithms with 10-fold cross-validation. We have used WEKA for the execution of the algorithms. We have implemented a desktop application on the Visual Studio .Net framework to apply processes of the proposed approach. The user interface of our tool can be seen in Figure 5.4.

In our tool, there is an import function to include the learning data as negative and positive texts. Another function is for preprocessing of the dataset. We remove stop words and transform all words into lower case. For normalization, we use ITU NLP tool. To tokenize and lemmatize the normalized text, we use Zemberek . Then, we create ARFF file from the resulting texts. Finally, using this file on WEKA, the results of algorithms are obtained. In the text box, the results of each step of the proposed approach are listed.



Original Text
Genel olarak iyi bir tesis.Memnun kaldm. İlk gece yalnız çok gürültü vardı uyuyamadık.Sonrasında odamızı değiştirdiler daha güzel bir odaya alındık.

Preprocessed Text
genel olarak iyi bir tesis memnun kaldm ilk gece yalnız çok gürültü vardı uyuyamadık sonrasında odamızı değiştirdiler daha güzel bir odaya alındik

Normalized Text
genel olarak iyi bir tesis memnun kaldım ilk gece yalnız çok gürültü vardı uyuyamadık sonrasında odamızı değiştirdiler daha güzel bir odaya alındik

Tokenized / Lemmatized Text
genel ol iyi bir tesis memnun kalmak ilk gece yalnız çok gürültü var uyumamak sonra oda değişmek daha güzel bir oda alınmak

Figure 5.4 User interface of the proposed tool

CHAPTER SIX
EXPERIMENTAL STUDY

6.1 Dataset Statistics

In this study, experiments are done by using three different datasets to evaluate the results of the methodology on different types of data. Movie review and hotel review datasets are downloaded from the Hacettepe University Multimedia Information Retrieval Group's website. Movie reviews on this website are collected from beyazperde.com, and hotel reviews are collected from otelpuan.com. All extracted movie reviews are rated by their own authors according to stars. One or two stars is classified as negative, while 4 or 5 stars is classified as positive. In a similar way, hotel reviews are rated between 0 and 100 instead of stars. The negative reviews are selected from 0 to 40-point reviews and the positive from 80 to 100-point reviews (Oğul & Ercan, 2016). A completely different dataset consisting of Tweets is also used in the experiments to control the accuracy of the proposed methodology. This dataset is taken from the website of the Kemik NLP group of Yıldız Technical University. It consists of 3000 Turkish tweets having three classes for SA.

Table 6.1 Statistics of the datasets

Datasets	# of instances	# of sentences	# of tokens
Movie	49,476	106,813	1,345,726
Hotel	11,164	17,874	738,216
Tweets	1,756	2,535	19,056

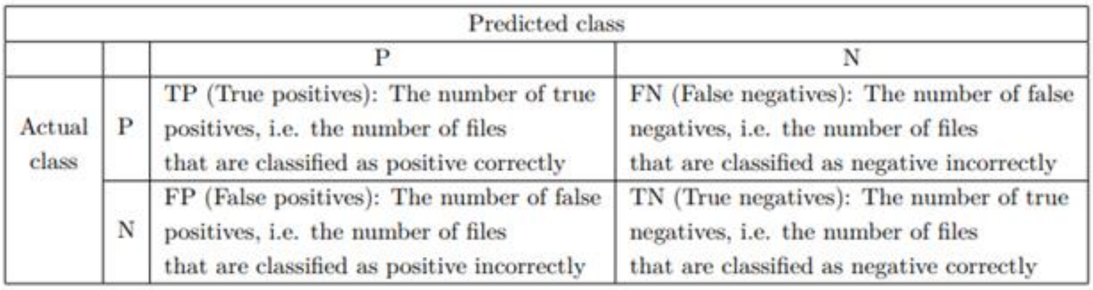
The statistics of the datasets, including the number of instances, sentences, and tokens are represented in Table 6.1. The dataset having the most instances is Movie dataset, and the less is Twitter. Twitter dataset has 3000 tweets, but we have removed neutral instances from it, so it has 1756 instances. Hotel dataset's reviews are longer than the others.

The texts in tweets are informal and have many misspelled words, but hotel and movie reviews are usually well-written and have sentimental words. Movie dataset has the largest number of sentences. Although Hotel dataset is fewer sentences, the length of its sentences is too longer than other datasets. According to the number of tokens, the words in Movie dataset are shorter than others.

6.2 Evaluation Metrics

The algorithms used in the study are NB, SVM, and J48. NB is selected as a probabilistic classifier, SVM is selected as a linear classifier, and J48 is selected as a decision tree classifier. NB is one of the simplest and most used machine learning algorithms used for text classification and based on the statistical Bayes theorem and conditional probability. The NB classifier presumes that the impact of a feature’s value on a given class is independent of the values of other attributes. SVMs are based on the structural risk minimization principle (Vapnik, 1995), which is the idea of finding a hypothesis (h) with the lowest error (Joachims, 1998). The error is the probability that h will have when it encounters new or randomly selected data. They can learn the dimensionality of features independently and therefore work well for text categorization. J48 is a C4.5 decision tree algorithm for classification based on binary trees. The main idea is to divide the data into ranges based on the attribute values in the training set (Goyal & Mehta, 2012). The evaluation metrics used are accuracy, precision, recall, and f-measure, which are defined using the terms in Table 6.2.

Table 6.2 Definition of confusion matrix



		Predicted class	
		P	N
Actual class	P	TP (True positives): The number of true positives, i.e. the number of files that are classified as positive correctly	FN (False negatives): The number of false negatives, i.e. the number of files that are classified as negative incorrectly
	N	FP (False positives): The number of false positives, i.e. the number of files that are classified as positive incorrectly	TN (True negatives): The number of true negatives, i.e. the number of files that are classified as negative correctly

Accuracy (Acc) is the ratio of the number of documents that are correctly classified to the total number of documents. The calculation of accuracy is given in Equation 6.1.

$$Acc = (TP + TN)/(TP + TN + FP + FN) \tag{6.1}$$

Precision (Pr) is the probability that a randomly selected document is retrieved as relevant. It is calculated as the ratio of the total number of positive files that are correctly classified to the total number of positive classified files, as in Equation 6.2

$$Pr = TP/(TP + FP) \tag{6.2}$$

Recall (Re) is the probability that a randomly selected relevant document is retrieved in a search. It is calculated as the ratio of the total number of positive files that are correctly classified to the number of positive files that are in the dataset, as in Equation 6.3.

$$Re = TP/(TP + FN) \tag{6.3}$$

The F-measure (Fm) is the harmonic mean of precision and recall, and it is calculated as in Equation 6.4.

$$Fm = 2 * Pr * Re/(Pr + Re) \tag{6.4}$$

6.3 Experimental Results

All datasets used in the experiments are balanced and have separate training and test sets, except the Twitter dataset, and all experiments run with 10-fold cross-validation. We have applied train/test ratio as 80/20. According to the experimental results, there are improvements in all three datasets.

We have experimented on three datasets with three methods. These methods are only ML, only lexicon, and hybrid. The results show that our hybrid approach outperforms both the lexicon-based and ML-based results in all datasets, as seen in Table 6.3.

To check the effectiveness of the new feature, the attributes are ranked using a filter-based attribute selection method, with information gain (IG) as an attribute evaluator and ranker as a search method, then sorted according to IG score. The experimental results are shown in Table 6.4. It is clearly seen that our new attribute named “type” is the first ranked attribute, having by far the best IG ranking score in all three datasets. The scores are 0.17388 in Movie, 0.32817 in Hotel, and 0.04737 in the Twitter dataset, respectively. The score in the Twitter dataset is less than the others because the Tweets in the dataset are very short and there are some abbreviations and jargon, which makes finding strong sentiment words harder. Despite this, our new feature is still in the first rank. Although the second and third-ranked features are the most used and powerful sentiment words in the language, the new feature has more impact in terms of sentiment.

Table 6.3 Summary of experimental results

Dataset	Classifier	Method	Average			Accuracy
			Pr	Re	Fm	
Movie	NB	ML	0.83	0.804	0.8	80.35%
		Hybrid	0.891	0.889	0.889	**88.93%**
	SVM	ML	0.799	0.799	0.798	79.85%
		Hybrid	0.863	0.863	0.863	**86.31%**
	J48	ML	0.689	0.674	0.667	67.35%
		Hybrid	0.781	0.779	0.779	**77.92%**
	Lexicon		0.67	0.79	0.725	70.93%
Hotel	NB	ML	0.875	0.838	0.834	83.80%
		Hybrid	0.909	0.9	0.899	**89.98%**
	SVM	ML	0.912	0.911	0.911	91.14%
		Hybrid	0.92	0.92	0.92	**91.96%**
	J48	ML	0.869	0.861	0.86	86.10%
		Hybrid	0.892	0.89	0.889	**88.96%**
	Lexicon		0.73	0.91	0.81	78.88%
Twitter	NB	ML	0.7	0.702	0.701	70.21%
		Hybrid	0.834	0.834	0.834	**83.37%**
	SVM	ML	0.716	0.708	0.71	70.84%
		Hybrid	0.822	0.818	0.819	**81.83%**
	J48	ML	0.672	0.667	0.647	66.69%
		Hybrid	0.729	0.727	0.728	**72.72%**
	Lexicon		0.53	0.81	0.64	62.81%

To improve the results' generalizability, they are tested using three different algorithms, i.e., NB as a probabilistic classifier, SVM as a linear classifier, and J48 as a decision tree classifier. The results of lexicon-based experiments are also included to compare. As a result of nine runs with three algorithms, the minimum difference between baseline and our approach's accuracy was 1.12% in the Hotel dataset with SVM. On the other hand, the maximum difference was 13.33% in the Twitter dataset with NB, as seen in Figure 6.1. The average improvement in all datasets with all algorithms was 7%.

Table 6.4 IG score of new generated feature

Datasets	**Id**	**Name**	**Score**
Movie dataset	4200	**type**	0.174
	103	kötü (bad)	0.036
	26	harika (wonderful)	0.032
Hotel dataset	3053	**type**	0.328
	1251	berbat (terrible)	0.151
	19	güzel (beautiful)	0.124
Twitter dataset	2468	**type**	0.047
	68	güzel (beautiful)	0.038
	66	hayat (life)	0.038

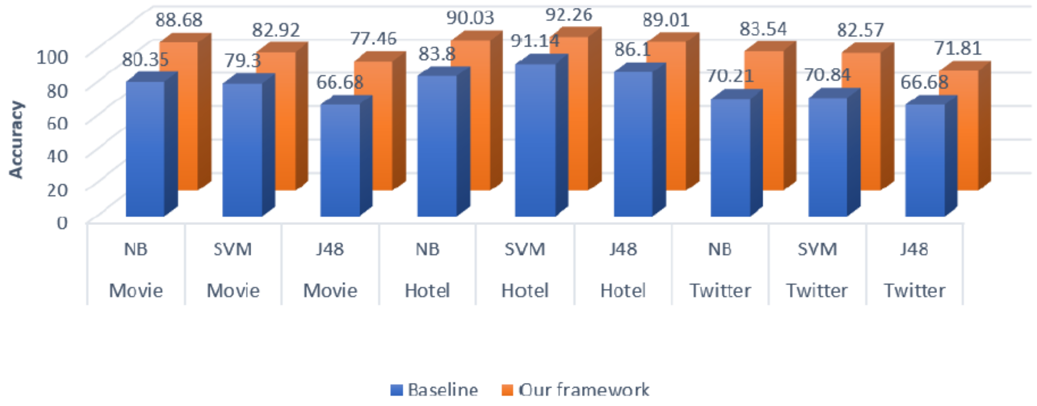
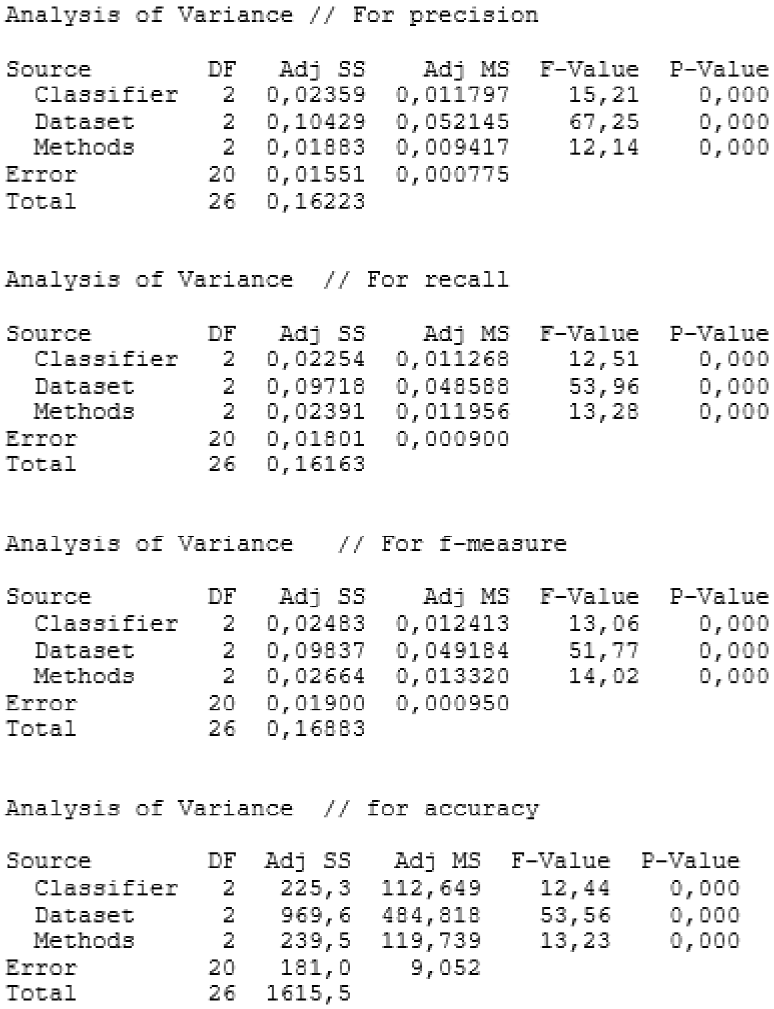


Figure 6.1 The experimental results of different ML algorithms

To evaluate the statistical significance of the results, we have performed a two-way ANOVA test. The statistical test results can be examined in Figure 6.2. In this figure, DF, SS, MS, and F denote degrees of freedom, the adjusted sum of squares, mean squares, F-statistics, and probability value, respectively. As it can be observed from the results, there is statistically significant difference (P < 0.001) for the means of the compared classifiers, datasets, and methods. Also, the 95% confidence interval for the compared algorithms based on the pooled standard deviation is presented in Figure 6.3 through Figure 6.6, which supports the results shown in Figure 6.2. Based on the statistical significances between the empirical results on three datasets, Figure 6.3, 6.4, 6.5, and 6.6 are divided into two regions denoted by red dashed lines for precision, recall, f-measure, and accuracy values. An interval plot shows a 95% confidence interval for the mean of each group. It is revealed that precision, recall, f-measure, and accuracy are all above the red line for this confidence interval.



```
Analysis of Variance // For precision

Source        DF   Adj SS    Adj MS  F-Value  P-Value
  Classifier   2  0,02359  0,011797    15,21    0,000
  Dataset      2  0,10429  0,052145    67,25    0,000
  Methods      2  0,01883  0,009417    12,14    0,000
Error         20  0,01551  0,000775
Total         26  0,16223

Analysis of Variance  // For recall

Source        DF   Adj SS    Adj MS  F-Value  P-Value
  Classifier   2  0,02254  0,011268    12,51    0,000
  Dataset      2  0,09718  0,048588    53,96    0,000
  Methods      2  0,02391  0,011956    13,28    0,000
Error         20  0,01801  0,000900
Total         26  0,16163

Analysis of Variance   // For f-measure

Source        DF   Adj SS    Adj MS  F-Value  P-Value
  Classifier   2  0,02483  0,012413    13,06    0,000
  Dataset      2  0,09837  0,049184    51,77    0,000
  Methods      2  0,02664  0,013320    14,02    0,000
Error         20  0,01900  0,000950
Total         26  0,16883

Analysis of Variance  // for accuracy

Source        DF  Adj SS   Adj MS  F-Value  P-Value
  Classifier   2   225,3  112,649    12,44    0,000
  Dataset      2   969,6  484,818    53,56    0,000
  Methods      2   239,5  119,739    13,23    0,000
Error         20   181,0    9,052
Total         26  1615,5
```

Figure 6.2. ANOVA results

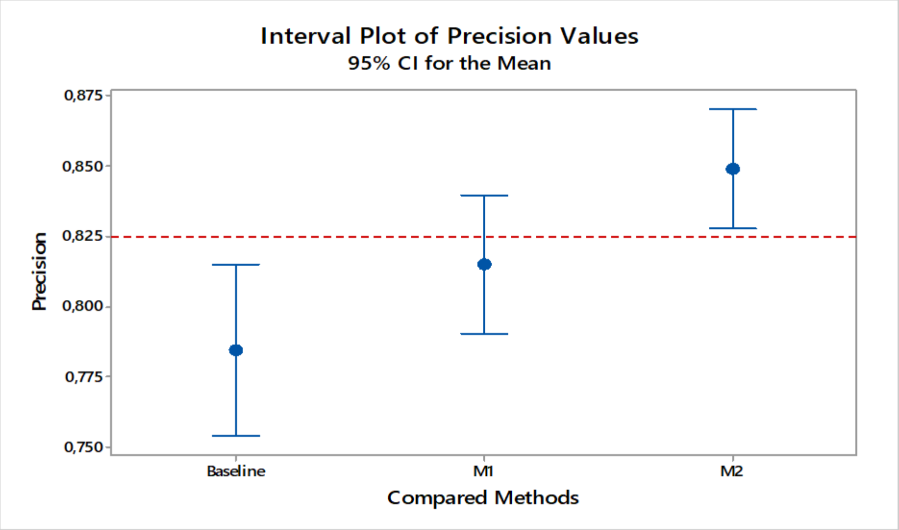


Figure 6.3 Interval plot of precision values

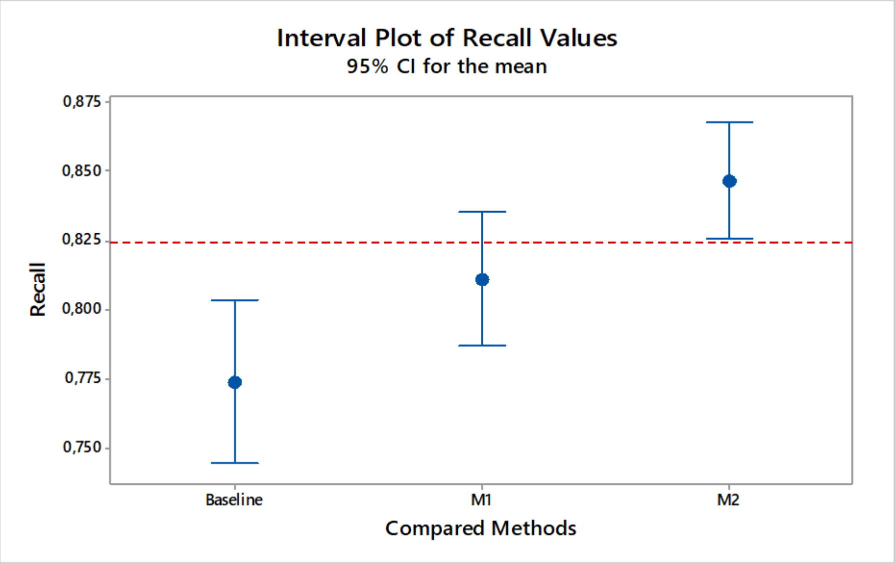


Figure 6.4 Interval plot of recall values

Hence, it is indicated that the differences between the results obtained by the proposed scheme (M2) are statistically significant compared to the results obtained by the baseline methods. There are significant improvements achieved with our hybrid SA framework in Turkish in all runs. SVM usually has the highest accuracy of all classification algorithms due to its robust nature, but it requires an extensive training set and a very long training time. The NB method is improved with our approach and surpassed the SVM and J48 in all cases except the Hotel dataset.

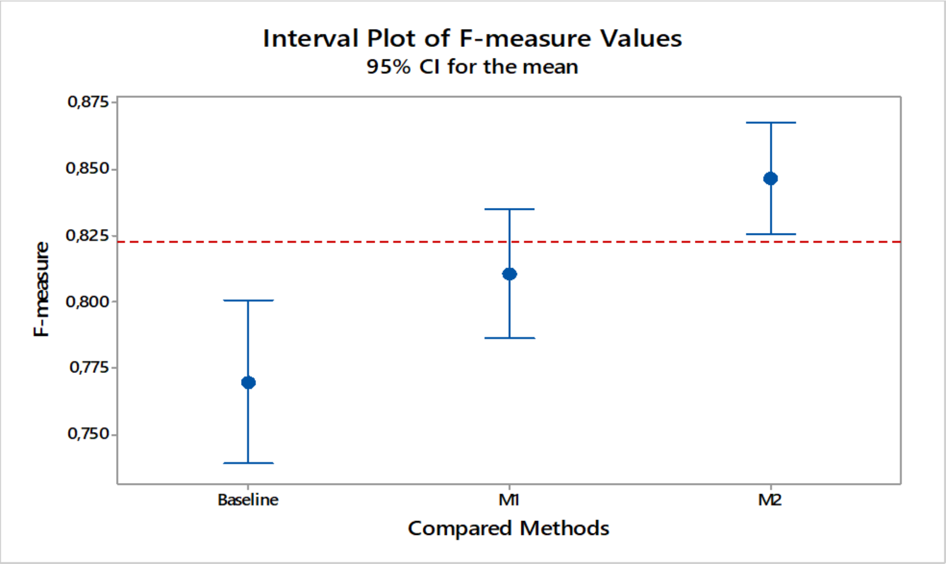


Figure 6.5 Interval plot of f-measure values

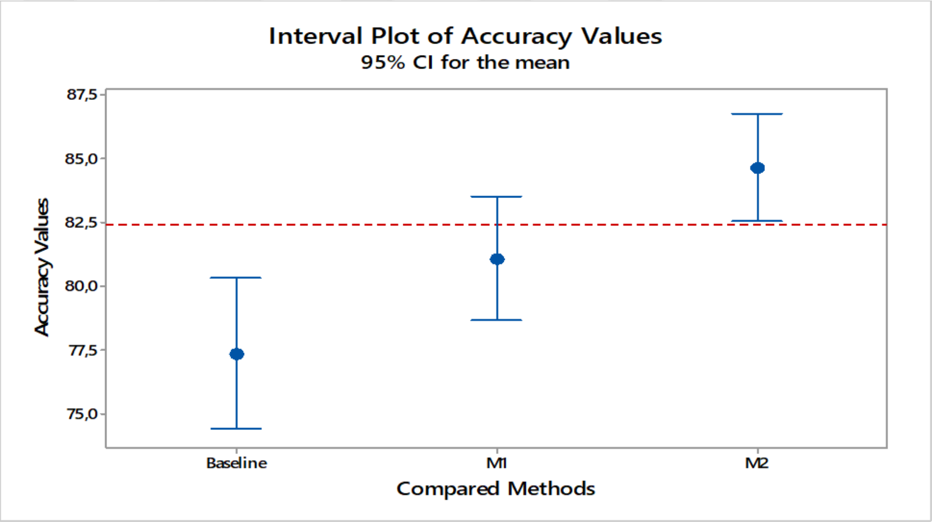


Figure 6.6 Interval plot of accuracy values

Finally, we compare our approach to previous SA studies using the same datasets. These studies, their techniques, and accuracy values are given in Table 6.5. First, in (Cetin & Fatih, 2013), the authors investigated the feasibility of active learning for Turkish SA. The aim of active learning is to get the same or better results with smaller amounts of training data. They experimented with the Twitter dataset that we used and the NB method. The results of the system with active learning were better than only NB with accuracy values 64% and 62.6%, respectively. Another study (Parlar, Sarac & Ozel, 2017) using the same Twitter dataset compared the performance of four feature selection methods using logistic regression. They showed that query expansion ranking (QER) and ant colony optimization (ACO) methods outperformed other traditional feature selection methods for SA. They evaluated their results with Fm using 5-fold CV and got the best results with QER. Movie and hotel datasets were prepared and used in (Ucan, Naderalvojoud, Sezer & Sever, 2016). They proposed an automatic translation approach to creating a lexicon for a new language. They used English resources mapping automatically to Turkish and constructed three different lexicons using different methods. Finally, they experimented with their lexicons and got the best accuracy value of 70.35% for Movie and 80.68% for Hotel utilizing TSDp, which is a lexicon prepared by parallel-based translation approach. Their ML-based results with SVM were 84.6% and 79.7% in the Movie and Hotel datasets, respectively. By all accounts, our hybrid method performs better on all the same datasets.

Table 6.5 Comparison of studies

Method	**Dataset**	**Technique**	**Results**
Cetin M., Fatih A.M.	Twitter	NB	Acc: 62.6%
Cetin M., Fatih A.M.	Twitter	NB + active learning	Acc: 64%
Parlar T., Sarac E., Ozel S.A.	Movie	Logistic regression + QER	Fm: 0.779
Ucan A. et al.	Movie	Lexicon	Acc: 70.35%
Ucan A. et al.	Movie	SVM	Acc: 84.6%
Ucan A. et al.	Hotel	Lexicon	Acc: 80.68%
Ucan A. et al.	Hotel	SVM	Acc: 79.7%
Our method	**Twitter**	**Hybrid (NB + eSTN)**	**Acc: 83.37%**
Our method	**Hotel**	**Hybrid (SVM + eSTN)**	**Acc: 91.96%**
Our method	**Movie**	**Hybrid (SVM + eSTN)**	**Acc: 86.31%**
Our method	Movie	Lexicon	Acc: 70.93%
Our method	Hotel	Lexicon	Acc: 78.88%

6.4 Threats to Validity

This subsection considers threats to validity. The types of them are threats to construct validity, threats to internal validity, and threats to external validity. Threats to construct validity is about the qualification of the evaluation metrics. In this study, precision, recall, F-measure, and accuracy is used like most of the past studies, therefore threats to the construct validity is minimized.

Threats to internal validity are biases that may be done by experimenters. For instance, when using supervised learning techniques, the dataset must be labeled. The labeling process may be subjective, and therefore it is better to involve some people looking as an outsider to double-check the labels. In our study, there are known datasets that are created considering this internal validity. The Movie and Hotel dataset is taken from HUMIR. These datasets are selected from two popular websites. The movie reviews are collected from “beyazperde.com” and hotel reviews from “otelpuan.com”. The Movie reviews are investigated, and they were already rated by own authors between 1 and 5 stars. The negative reviews are created from 1 and 2 stars. The positive reviews are created from 4 and 5 stars. The Hotel reviews are investigated, and they were already rated by their authors between 0 and 100 points instead of stars. The negative reviews are created from 0 to 40-point reviews. The positive reviews are created from 80 to 100-point reviews. Twitter dataset is taken from Yıldız Teknik University Kemik NLP Group. Another threat to internal validity is the selection of attributes used for classification. In this case, True Positive Rates (TPR) and False Positive Rates (FPR) can be too low or high. It is minimized by using 10-fold cross validation.

Threats to external validity is about the generalizability of the results. Our framework is tested on 3 datasets with different size and from different domains to guarantee that our results will apply to all type of datasets. It is also tested with 3 different ML algorithm each of them from different type of supervised techniques. The use of a single machine learning algorithm can be a threat to the external validity of this study. Therefore, NB is selected as probabilistic classifier, SVM is selected as linear classifier, and J48 is selected as decision tree classifier. It is believed that threats to external validity are minimized, but in the future new datasets and new algorithms will be tested additionally.

CHAPTER SEVEN
CONCLUSION AND FUTURE WORKS

7.1 Conclusion

Sentiment analysis is the study of understanding people's opinions and attitudes for an entity, people, or service. In the last decades, with the widespread usage of microblogging sites, forums, social media platforms, and e-commerce sites, people widely share their opinions on the Internet. The amount of data transmitted by the users on social media platforms is enormous; therefore, it is named as big data. It is not practical to analyze and understand this big data manually. It is better to computerize this process by using SA techniques.

There are some areas where the SA is useful. For instance, companies and organizations need to be aware of their employees' and customers' feelings about their organizations. Human resources also would like to discover whether a potential employee will be loyal or leave after receiving training and benefits. Besides, the tweets about the candidates are used to predict the results of elections by the government. People read the customer reviews about the products and decide whether it is satisfiable or not for them. There is much usage of social media data like these. In the context of analyzing big data for its sentiment, a question arises, whether it is possible to improve the existing SA results using a new hybrid approach. We have researched for it and obtained promising results.

In this thesis, we aimed to answer this question by performing experiments with our hybrid approach for SA in Turkish on three different datasets (Movie, Hotel, and Twitter) by three different ML algorithms of NB, SVM, and J48. As part of this thesis, we have developed a framework to conduct data collection, preprocessing, ARFF creation, and hybrid SA steps.

Through this research three main contributions were made: 1) to the best of our knowledge, it is the first study proposing and testing a hybrid SA method in Turkish; 2) the first comprehensive Turkish SA dictionary, STN is expanded using the

Automated Synonym Dictionary; 3) lemmatization in NLP is adapted for Turkish SA to preserve the positive and negative meaning of tokens.

We showed that the accuracy of the SA for all datasets can be improved by combining the powerful aspects of ML-based and lexicon-based approaches in our hybrid approach. To improve the experimental results, on the lexicon-based side, STN is expanded with ASDICT, and a lexicon score is calculated based on the polarity of the words in eSTN. It is performed by finding all the synonyms of terms in STN in ASDICT and including them with the same polarity scores in the eSTN. Experiments showed that by using eSTN, the matching terms increased by 53% in Movie dataset, 88% in Hotel dataset, and 96% in Twitter dataset.

As for the feature selection by lemmatization, which is one of our study's contributions, we have utilized Zemberek by customizing it with some rules. For instance, we have not stemmed all suffixes. We have preserved the meaningful suffixes such as -siz, -sız, -li, -lı. Also, we have transformed all words which have verb stem into the infinitive form. Through this method, the number of features is reduced significantly. It is a natural feature selection approach.

The other contribution in our study is new feature generation algorithm. It is generated utilizing eSTN and included in the ARFF file as a new feature. Then, we have evaluated the effect of it in all datasets. According to experimental results, the ranking of all features based on the IG scores show that the lexicon-based new feature is at the top of the list, confirming its relevance.

Another point to emphasize is the negation handling issue. We have preserved the suffixes containing positive or negative meaning to conserve the sentiment. Additionally, we have handled the negation resulting from the word "*değil*". The words' polarity values preceding this word are negated as swapping the scores of positive polarity and negative polarity.

We evaluated our method with 3 different algorithms on 3 different datasets and using 10-fold cross-validation. Experimental results show that the hybrid method achieves a minimum 77.92% accuracy with j48 and a maximum of 88.93% with NB. It is better than ML results from 7% to 10%. In Hotel reviews, it achieves a minimum 88.96% accuracy with j48 and a maximum of 91.96% with SVM. It is better than ML results up to 6%. The increase is not as much as in movie reviews because the reviews are long and well-written. Therefore, even ML techniques are successful on their own. We are glad that there is still an improvement. In Twitter reviews, it achieves a minimum of 72.72% accuracy with j48 and a maximum of 83.37% with NB. It is better than ML results, up from 6% to 13%. Tweets are informal texts and have abbreviations, hashtags, and misspellings. For this reason, ML algorithms scored only 70% accuracy at most. The hybrid approach improved it reasonably.

To evaluate the statistical significance of the results, we have performed a two-way ANOVA test. According to the results of ANOVA, there is a statistically significant difference ($P < 0.001$) for the means of the compared classifiers, datasets, and methods. In addition, the 95% confidence interval plots for the compared algorithms based on the pooled standard deviation is calculated. The interval plots show a 95% confidence interval for the mean of each group. It is showed that precision, recall, f-measure, and accuracy are all above the red line for this confidence interval.

To conclude, we have compared our experimental results with the studies using the same dataset as the benchmark. The findings of this thesis demonstrated that our hybrid approach outperforms both ML-based and lexicon-based approaches. These results have serious implications for both industry and academia.

7.2 Future Works

This study has some limitations which should be addressed for future work. One of the future directions for the proposed approach consists of research on aspect-based SA and its subtasks to improve the system’s overall performance. Datasets we have used generally have one sentiment for all documents. However, our approach may not be sufficient for long documents that have different sentiments for several entities. It is named multi-polarity. Aspect-based SA is fine-grained, so it is more appropriate for such texts.

As another future work, we would like to evaluate the proposed method on some English datasets to check its effectiveness in multilingual environments. NLP is particular and dependent on a language specifically. We have also developed our negation handling according to Turkish linguistic features, but our feature generation algorithm is a generic approach. It is applicable to all languages, and with some adaptations for NLP, we believe that this approach can also obtain successful results.

Besides, word vectors such as Word2Vec may be used to improve the quality of the feature selection process. BOW model we have used cannot capture the meaning between words. It captures the words as features only. Word embeddings such as Word2Vec use a model to map a word into vectors so that similar words will be closer to each other. This model takes the surroundings of a word according to a window size to maintain the semantical information of words. In this way, we can also apply deep learning techniques and compare the results with ours.

Furthermore, the lexicon may be improved with other methods and expanded to increase the scope of it. Our hybrid approach utilizes eSTN, which is more comprehensive than STN. The quality of the lexicon is vital for SA because the polarities are obtained from there. The better the dictionary, the more words it is caught. Therefore, sentiments are evaluated more precisely.

Finally, we may focus on the classification of negative, sarcasm, or irony containing statements. Sarcasm is hard to detect because the real sentiment is the opposite of the word’s meaning due to the irony. Sarcasm is different from negation because it contains intensified positive words to express a negative opinion. Especially, there are sarcastic sentences in tweets, which make it hard to train models for ML algorithms. If we handle this problem, the evaluation results will be better. Another issue is about determining the range of the negation. We can take negation by reversing the words’ polarity, but it is difficult to decide how many words should be affected by negation. Ambiguity is also a problem when it is impossible to decide the sentiment in advance without knowing the context because some words are dependent on the context. We are also planning to tackle such linguistic issues as future work.

REFERENCES

Akgül, E.S., Ertano, C., & Diri, B. (2016). Sentiment analysis with Twitter. *Pamukkale University Journal of Engineering Sciences*, *22*, (2), 106-110.

Aktaş, Ö., Birant, Ç., Aksu, B., & Çebi, Y. (2013). Automated synonym dictionary generation tool for Turkish (ASDICT). *BILIG - Turk Dunyasi Sosyal Bilimler Dergisi*, *65,* (9), 47-68.

Appel, O., Chiclana, F., Carter, J., & Fujita, H. (2016). A hybrid approach to sentiment analysis. *In: IEEE Congress on Evolutionary Computation; Sendai, Japan*, 4950-4957.

Baloglu, A. & Aktas, M.S. (2010). An automated framework for mining reviews from blogosphere. *International Journal of Advances in Internet Technology*, *3,* (4), 234-244.

Boudad, N., Faizi, R., Rachid, O.H.T., & Chiheb, R. (2018). Sentiment analysis in Arabic: a review of the literature. *Ain Shams Engineering Journal*, *9,* (4), 2479-2490.

Boynukalın, Z. (2012). *Emotion analysis of Turkish texts by using machine learning methods*. Master Thesis, Middle East Technical University, Ankara.

Cetin, M., & Fatih, A.M. (2013). Active learning for Turkish sentiment analysis. In*: IEEE International Symposium on Innovations in Intelligent Systems and Applications*, Turkey, 1-4.

Çoban, Ö., Özyer, B., & Özyer, G.T. (2015). Sentiment analysis for Turkish Twitter feeds. In: *23nd Signal Processing and Communications Applications Conference*, Malatya, Turkey, 2388-2391.

Dehkharghani, R., Saygin, Y., Yanikoglu, B., & Oflazer K. (2015). SentiTurkNet: A Turkish polarity lexicon for sentiment analysis. *Language Resources and Evaluation*, *50*, (3), 667-685.

Dehkharghani, R., Yanikoglu, B., Saygin, Y., & Oflazer, K. (2017). Sentiment analysis in Turkish at different granularity levels. *Natural Language Engineering*, *23*, (4), 535-559.

Duwairi, R.M. (2015). Sentiment analysis for dialectical Arabic. In: *Proceedings 6th International Conference on Information and Communication Systems*, Amman, Jordan, 166-170.

Ehsani, R., Solak, E., & Yildiz, O.T. (2018). Constructing a WordNet for Turkish using manual and automatic annotation. *ACM Transactions on Asian Language Information Processing*, *17*, (3), 1-15.

Eroğul, U. (2012). *Sentiment analysis in Turkish*. Master Thesis, Middle East Technical University, Ankara.

Eryiğit, G., & Torunoğlu-Selamet, D. (2017). Social media text normalization for Turkish. *Natural Language Engineering*, *23*, (6), 1-41.

Esuli, A., & Sebastiani, F. (2006). SENTIWORDNET: A publicly available lexical resource for opinion mining. In: *Proceedings of the 5th Conference on Language Resources and Evaluation*, Genoa, Italy, 417-422.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database (Language, Speech, and Communication).* Cambridge, MA: MIT Press.

Govindarajan, M. (2013). Sentiment analysis of movie reviews using hybrid method of naive Bayes and genetic algorithm. *International Journal of Advanced Computer Research*, *3*, (4), 139-146.

Goyal, A., & Mehta, R. (2012). Performance comparison of naïve Bayes and J48 classification algorithms. *International Journal of Applied Engineering Research, 7,* (11), 1389-1393 .

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. In: *European Conference on Machine Learning*, Chemnitz, Germany, 137-142.

Kaya, M., Fidan, G., & Toroslu, I.H. (2012). Sentiment analysis of Turkish political news. In: *International Conferences on Web Intelligence and Intelligent Agent Technology*, Washington, DC, USA, 174-180.

Kaynar, O., Görmez, Y., Yildiz, M., & Albayrak, A. (2016). Sentiment analysis with machine learning techniques. In: *International Artificial Intelligence and Data Processing Symposium*, Malatya, Turkey, 80-86.

Maynard, D., & Funk, A. (2012). Automatic detection of political opinions in tweets. In: *Proceedings of the 8th International Conference on the Semantic Web*, Heraklion, Greece, 88-99.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: a survey. *Ain Shams Engineering Journal, 5,* (4), 1093-1113.

Oğul, B.B., & Ercan, G. Sentiment classification on Turkish hotel reviews. (2016) In: *24th Signal Processing and Communication Application Conference*, Zonguldak, Turkey, 497-500.

Ohana, B., & Tierney, B. (2009). Sentiment classification of reviews using SentiWordNet. In: *9th IT&T Conference*, Dublin, Ireland, 10-19.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques; In: *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA, 79-86.

Parlar, T., Sarac, E., & Ozel, S.A. (2017). Comparison of feature selection methods for sentiment analysis on Turkish Twitter data. In: *25th Signal Processing and Communications Applications Conference*, Turkey, 1-4.

Rushdi-Saleh, M., Martín-Valdivia, M.T., Ureña-López, L.A., & Perea-Ortega, J.M. (2011). OCA: Opinion corpus for Arabic. *Journal of the Association for Information Science and Technology, 62,* (10), 2045–2054.

Sharma, R., Nigam, S., & Jain, R. (2014). Opinion mining of movie reviews at document level. *International Journal on Information Theory, 3,* (3), 13-21.

Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology, 63,* (1), 163-173.

Thisted, R.A. (1988). *Elements of Statistical Computing: Numerical Computation.* New York, NY, USA: Chapman & Hall.

Turkmenoglu, C., & Tantug, A.C. (2014). Sentiment analysis in Turkish media. In: *International Conference on Machine Learning*, Beijing, China, 32-42.

Ucan, A., Naderalvojoud, B., Sezer, E.A., & Sever, H. (2016). SentiWordNet for new language: automatic translation approach. In: *12th International Conference on Signal-Image Technology & Internet-Based Systems*, Naples, Italy, 308-315.

Vapnik, V.N. (1995) *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer.

Vinodhini, G., & Chandrasekaran, R.M. (2013). Effect of feature reduction in sentiment analysis of online reviews. *International Journal of Advanced Research in Computer Engineering & Technology*, *2*, (6), 2278–1323.

Vural, A.G, Cambazoglu, B.B., Senkul, P., & Tokgoz, Z.O. (2012). A framework for sentiment analysis in Turkish: application to polarity detection of movie reviews in Turkish. In: *Computer and Information Sciences III*, London, UK, 437-445.

Yildirim, E., Çetin, F., Eryigit, G., & Temel, T. (2017). The impact of NLP on Turkish sentiment analysis. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, *7,* (1), 43-51.

Zhang, L., Hua, K., Wang, H., Qian, G., & Zhang, L. (2014). Sentiment analysis on reviews of mobile users. *Procedia Computer Science, 34,* (11), 458-465.

APPENDICES

APPENDIX 1: LIST OF ACRONYMS

Acronym	Definition
ASDICT	Automated Synonym Dictionary
AUC	Area Under Curve
BN	Bayesian Network
CTD	Contemporary Turkish Dictionary
DEU	Dokuz Eylül University
eSTN	Extended SentiTürkNet
FPR	False Positive Rates
GA	Genetic Algorithm
IG	Information Gain
ML	Machine Learning
MNB	Multinomial Naive Bayes
NB	Naïve Bayes
NLP	Natural Language Processing
PCA	Principal Component Analysis
RF	Random Forest
SA	Sentiment Analysis
STN	SentiTürkNet
SVM	Support Vector Machines
TDK	Turkish Language Society
TPR	True Positive Rates

APPENDIX 2: SAMPLES FROM DATASETS

Hotel dataset

80101;Hotel Review;asla gidilmeyecek bir otel hasta oldukotel tam anlamıyla bir fiyasko satın alırken ve web sitesinde gözünüze çarpan en büyük özellik otelin tüm alanlarının yenilenmiş olması ama bunun gerçekle alakası yok odalar en az yıllık bir otel harabeliğinde yemekler ve özellikle kahvaltı tam bir hayal kırıklığı kahvaltıdaki yiyecekler asla yenmeyecek ve yedirilmeyecek kadar kötü bir tane lekesiz temiz bir tabak bardak veya çatal kaşık görmeniz olası bile değil içecek konusunda su değil zehirli su katılmış gibi gerçeği ile alakası olmayan içecekler ıce tea yada soğuk çay cinsi birşey otelde asla yok konsepte uygun değilmiş açıklama bu soğuk çay hangi konseptin ki acaba bu otele uymuyor garsonların hepsi kendi dalında bir kabadayı restaurant müdürü denen kişi inanılmaz yeteneksiz asla yeme içme kültürü yok gerçekten birsey isteyipte almanız mucize birde asıl bir mevzu varki anlatılmaz bu otelde can güvenliğiniz yok oda anahtarı kardeşimdeydi ben resepsiyona anahtar almaya gittim sırf anahtar yapmamak için elimde anahtar yok dedi benim sorunum değil bulacaksınız ben odama gireceğim anahtar kardeşimde oda otel dışında dedim sordu oda numaramı yaptı verdi ama tuhaf olan şuki ne oda numaramdan adımı kontrol etti yada hiçbirşey sormadı bizi daha öncede görmedi ki güven esaslı verdi diyeceğim yani herkes oda anahtarını alıp herseyi yapabilir otelde şampuan yok tamam kimse kullanmıyor belki ama * lı bir otelde nasıl olmaz otelde terlik yok yani yoklar oteli ama şunu söylemem gereki ki housekeeping deki çalışanlar çok iyi hk yöneticileri asla insana değer vermeyen asık suratlı insanlar tatil dönüşü kendimi kardeşimle beraber hastanede bulduk tatil boyunca azıcıkda olsa yediğimiz herseyi çıkardık ve geldiğimizde serum alacak kadar hasta olduk biz gittiğimizde otelin sahibide oteldeydi tüm şikayetleri memnuniyetsizlikleri duyuyor ama asla umurlarında olmuyor sahili çok kötü kıyısı berrak değil iskele dökülüyor asla gidilmeyecek bir otel;**Negative;train;1**

91979;Hotel Review;"Genel olarak otel hizmet ve her bölümdeki çalışanlar iyi güleryüzlü yemekler iyi imkanları iyi herhangi bir sorun yok odaların bir kısmı biraz eski olmakla beraber genel olarak iyi. ";**Positive;train;1**

83968;Hotel Review;otel igrenç eski bir yapı ve çalısanlar yetersız havuz temizlenmiyor dogru dürüst aşçı desen herşeye burnunu sokuyor ve kadınlara askıntı oluyor bulundugum sürece tüm kadınlar bundan şıkayetçi oldu tuvaletlerı berbat tamamen para avcısı bir sahibi var demedı demeyin sonra pişman olmayın paranızla gitmeden önce iyi düşünün;**Negative;test;1**

94610;Hotel Review;Yemekleri güzel servisinden memnunuz. Hoş davranıyorlar güzel bir tatil.;**Positive;test;1**

Movie Dataset

245;Movie Review;" 10/0 alan zaman kaybından başka ele bi veri bırakmayan bir film şiddetle tavsye edilmez ";**Negative;train;1**

3132;Movie Review;" Vasat ötesi. Ata DEMİRER film çevirmeyi bırakmalı bence.Kendisi Stand-Up yapmaya devam etmeli.Film çok vasat. ";**Negative;train;2**

5763;Movie Review;" çok durağan ve konu çok basit bu kadar bekledikten sonra renonun bu filmi olmamış 10/5 ";**Negative;train;3**

8420;Movie Review;" fragmanlar kalenin dıştan görünüşünü verir...azıcık bile etkilenmedim filmden.senaryo çok basit.adam tek başına çevirmiş resmen filmi diğer oyuncuları beğenmedim.özellikle de kadını...gayet sıkıcı bir film. ";**Negative;train;4**

11345;Movie Review;" Sanki bir tarantino filmi gibiydi.Uzun ve aslında gidisatı cok etkilemeyen dialoglar.Fazla uyusturucu sohbetleri ve bunlardan ziyade taxi driver ın kötü bir gölgesi gibi.Ambulance driver adını verdigim film üstad ın en kötü filmi ";**Negative;train;5**

52383;Movie Review;" filmi bugün arkadaşlarımla izledim. çok eğlendik harikaydı. başarılı bir devam filmi niteliğindeydi. şrek 10 da çekilse kesinlikle giderim. serinin diğer filminin çekilmesi taraftarıyım. mükemmeldi. ";**Positive;train;1**

54898;Movie Review;" çok özel bir film:)) 90lı yılların herşeyini seviyrm. Roxettenin it must have been love şarkısını film bittikten sonra yüksek sesle dinledm...Titanicten sonra etkilendiğim tek &quot;aşk&quot; film oldu..ayrıca gere ve roberts çifti çok yakışmış filme...bu iki karaktere aşık oldum diyebilrim filmi izlerken...özellikle julia roberts gerçekten çok özel ve çekici bir kadın:))) ";**Positive;train;2**

57497;Movie Review;" Gerçekten yönetmen kendini çok geliştirmiş. çok iyi bir filmdi bence beni çok etkiledi. Bu tarzı sevenler mutlaka izlemeli! ";**Positive;train;3**

60226;Movie Review;" Uzun lafa gerek yok.Kesinlikle bir başyapıt ve arşivlik bir film.Mutlaka izlenmesi gereken bir film... ";**Positive;train;4**

62931;Movie Review;" harika bir film muhteşemmmm.film müziğine de bayıldım ";**Positive;train;5**

13745;Movie Review;" Ya bi film nası bu kadar güzel başlayıp bun kadar saçmalayabilir sonradan çok büyük bi hewesle başladım izlemeye ama sonu hüsran oldu.Hoş bi konu yakalamışlar ama final cidden çok kötüydü bu kadar ii oyunculara yakışmamış bi film... ";**Negative;test;1**

16290;Movie Review;" 1 yıldır bu filmi bekliyordum ve sinemaya gittim. Açıkçası hayal kırıklığı yaşadım. Viking dönemi ile ilgili bir savaş filmi beklerken uzaylı bir yaratığın olduğu (Bilim kurgu filmlerini severim ama...)bir film seyrettim. İzlemeseniz de olur. Çok şey kaybetmezsiniz.5/10 ";**Negative;test;2**

18819;Movie Review;" bu kadar kötü bir film olamaz kesinlikle zaman kaybı kimseye tavsiye etmem ";**Negative;test;3**

21412;Movie Review;" ispanyol sineması son yıllarda iyi işler yapıyor.Açıkçası buna dayanarak Hipnoz'u izlemeye gittim.Ancak bu kez tel tel dökülen bir İspanyol filmi vardı.Filmi izlerken baya sıkıldım.Çünkü izleyiciyi çekecek herhangi bir şey yoktu.Sanki senaryo yok gibi.Açıkçası boynu bükük şekilde ayrıldım ";**Negative;test;4**

24526;Movie Review;" bu puan çok..bu kadar kötü film izlememiştim...Bole film yapmamamaları lazım yazıktır gühantır yav...(0/10) ";**Negative;test;5**

65925;Movie Review;" İnsanın hayatına yön verecek insanı kendi içine döndürecek ve çoğğu insana da ders verecek bir film.Bu film hayatın ta kendisi... ";**Positive;test;1**

68305;Movie Review;" Her ne kadar mantık hataları olsada testere bana göre yılın en iyi gerilim filmi. Katilin bütün planlarının saat gibi işlemesi falan... Yönetmeni oyuncuları çok iyi tanımasakta başarılı bir yapım.Filmin sonuna kadar ne olacağını kestiremiyorsunuz ve film şaşırtıcı bir sonla bitiyor. Final sahnesi nefes kesici. umarım devam filminide başarılı yaparlar. ";**Positive;test;2**

70599;Movie Review;" kesinlikle harika bir film izlememiş olanlar mutlaka izlesinler... çok mantıklı bır konusu var filmi izlerken kesinlikle çok zevk alıcaksınız=) ";**Positive;test;3**

73567;Movie Review;" izlerken insanı meraklandıran ve köz kırttırmayan bi film hele şükür michael douglas... ";**Positive;test;4**

76857;Movie Review;" Sinema severlerin kesinlikle izlemesi gereken bir film diye düşünüyorum.Oyunculuk on numara senaryo on numara ve tabiki de Nolan faktörünü unutmamak gerek. Böyle yaratıcı ve yetenekli bir yönetmenden bu kadar kaliteli enfes bir film çıkar. Helan sana Nolan. ";**Positive;test;5**

Twitter dataset

Positive tweets

abla cuma günü turkcell muzikte seni canli canli internettenmi dinlicez turkcell muzikte olcagini biliyorum ama nasil olcak?

turkcell in 3g si kamil kocun aptal wifi indan cok daha ii ki :)

Negative tweets

işe bak! reklam için aramış, ulaşamayıp ses kaydını sesli mesaj bırakmış. dinleme ücreti kesti benden! fiyasko!

ilk firsatta hattimi iptal iptal ettirecegim. tebrikler.