DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

LEXICON-BASED EMOTION ANALYSIS IN TURKISH

by
Mansur Alp TOÇOĞLU

July, 2018
İZMİR

LEXICON-BASED EMOTION ANALYSIS 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
Mansur Alp TOÇOĞLU

July, 2018
İZMİR

Ph.D. THESIS EXAMINATION RESULT FORM

We have read the thesis entitled "**LEXICON-BASED EMOTION ANALYSIS IN TURKISH**" completed by **MANSUR ALP TOÇOĞLU** under supervision of **ASSOC. PROF. DR. ADİL ALPKOÇAK** 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.

Assoc. Prof. Dr. Adil ALPKOÇAK

Supervisor

Assoc. Prof. Dr. Deniz KILINÇ

Thesis Committee Member

Asst. Prof. Dr. Canan ATAY

Thesis Committee Member

Prof. Dr. Ebru Sezer

Examining Committee Member

Doç. Dr. Akın Özçift

Examining Committee Member

Prof.Dr. Latif SALUM

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my supervisor, Assoc. Prof. Dr. Adil ALPKOÇAK, for his guidance, endeavor and seamless support throughout this study. My cordial thanks and appreciation also extend to my thesis committee members, Assoc. Prof. Dr. Deniz KILINÇ and Asst. Prof. Dr. Canan ATAY, for their moral support, thought provoking and invaluable comments.

I would also like to offer my thanks to my friends Okan ÖZTÜRKMENOĞLU and Ali Rıza ARIBAŞ for their seamless help and support during this study.

Finally, I would like to express my sincere gratitude to my dear wife, son, parents and sister for their support, patience and love. I could not be able to complete this thesis without their precious support.

Mansur Alp TOÇOĞLU

LEXICON-BASED EMOTION ANALYSIS IN TURKISH

ABSTRACT

This thesis presents a new dataset and a new lexicon for emotion analysis studies in Turkish text. To gather this dataset, we conducted a survey and collected 27,350 entries from 4,709 individuals. Then, we performed a validation process in which annotators validated each entry one by one by assigning a related emotion category. As a result, we obtained two datasets, one raw and the other validated. Subsequently, we generated four versions of these two datasets using two different stemming methods and then modeled them using a vector space model. Then, we ran machine learning algorithms on the models to calculate the accuracy, precision, recall and F-measure values. Based on the results we obtained, we concluded that the SVM classifier yielded the highest performance value and that the models trained with a validated dataset provide more accurate results than the models trained with a non-validated dataset.

In the second phase of the thesis, we propose a lexicon for the use of lexicon-based emotion analysis in Turkish text by using the dataset we constructed within the thesis. We explored the effects of stemming, term-weighting, lexicon enrichment and term selection approaches for lexicon-based emotion analysis. We first pre-processed the documents (entries) to obtain stems of each term using different approaches. Afterward, we proposed two different weighting schemas based on term-class frequencies and Mutual Information values. Next, we examined bi-grams and concept hierarchy for lexicon enrichment. Furthermore, we applied term selection for efficiency issues. Lastly, we evaluated the performance of the lexicon by using keyword-spotting technique on a separate Turkish dataset. The experiments showed that use of our proposed lexicon in keyword-spotting technique produces a satisfactory result in emotion analysis in Turkish Text.

Keywords: Emotion analysis, emotion extraction, Turkish language, text classification, TREMO dataset, Turkish emotion lexicon

TÜRKÇE METİNLERDE SÖZLÜK TABANLI DUYGU ANALİZİ

ÖZ

Bu tez, Türkçe metinlerde duygu analizi çalışmaları yapmak için yeni bir veri seti ve yeni bir sözlük ortaya koymaktadır. Bu veri setini oluşturmak için, 4,709 katılımcıdan 27,350 adet doküman toplandığı bir anket yürütülmüştür. Ardından, etiketleyicilerin her bir dokümanın duygu kategorisini birer birer doğruladıkları bir doğrulama süreci yürütülmüştür. Sonuç olarak, biri ham, biri de doğrulanmış olarak adlandırılan iki adet veri seti elde edilmiştir. İki adet köke indirgeme metodu kullanılarak bu iki veri setinden dört adet versiyonu elde edilmiş ve sonrasında bir uzay vektör modeli yardımıyla bu dört versiyon modellenmiştir. Doğruluk, kesinlik, hassasiyet ve F-ölçüm değerlerini hesaplamak için makine öğrenme algoritmaları çalıştırılmıştır. Elde edilen sonuçlara dayanarak; SVM sınıflandırıcısının en yüksek performans değerini sağladığı ve doğrulanmış veri seti ile çalıştırılan modellerin, doğrulanmamış veri seti ile çalıştırılan modellerden daha doğru sonuçlar verdiği tespit edilmiştir.

Tezin ikinci aşamasında, tez içinde oluşturulmuş olan veri seti kullanılarak, Türkçe metinlerde sözlük bazlı duygu analizi için bir sözlük önerilmektedir. Köke indirgeme, terim ağırlığı, sözlük zenginleştirme ve terim seçimi yaklaşımlarının etkileri araştırılmıştır. Dokümanların farklı yaklaşımlar kullanılarak işlenmesiyle her terimin kökü elde edilmiştir. Daha sonra, terim-sınıf frekanslarına ve karşılıklı bilgi değerlerine dayanan iki ağırlıklandırma şeması kullanılmıştır. Sözlük zenginleştirmesi için bi-gram ve kavram hiyerarşisi kullanılmıştır. Sonrasında, verimlilik sorunları için terim seçimi uygulanmıştır. Son olarak, sözlüğün performansı ayrı bir Türkçe veri setinde anahtar kelime tespiti tekniği kullanılarak ölçülmüştür. Yapılmış olan deneyler, önerilmiş olan sözlükteki anahtar kelime tespiti tekniğinin kullanımının Türkçe metinlerden duygu çıkarımı için tatmin edici sonuçlar verdiğini göstermiştir.

Anahtar kelimeler: Duygu analizi, duygu çıkarımı, Türkçe dili, metin sınıflandırması, TREMO veri seti, Türkçe duygu sözlüğü

CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

CHAPTER ONE
INTRODUCTION

1.1 Overview

In today's world, with the rapid evaluation of technology, many things have changed in the lives of people. These changes have provided many opportunities to humanity which can be listed as follows: transportation, the internet, social media applications, health care and so on. One of the crucial changes has been in the area of the internet. The number of internet users increased to 4.021 billion people in 2018 which are the 53% of the total human population (Kemp, 2018). Here, the developing technology plays an important role at this increase. For example, mobile phone usage is one of them and the increase of its usage is 4% in 2018. As a result, the number of people using mobile phones increased to 5.135 billion individual (Kemp, 2018). As the usage of internet spreads all over the world, it has been started to be used in many areas, such as social media, blogs, shopping sites and so on. The use of internet in these areas has become very popular and very common in today's world. If we take social media applications into consideration, they provide user-friendly platforms for people to communicate to each other from any part of the world by sharing data formats such as text, image, video and etc. As a result of these communication developments, today's people are mostly socialized over the internet. Of course, as people socialize by using social media applications, they help to generate an extremely large amount of data. For example according to (Aslam, 2018), total monthly active Twitter users, 330 million users, generate 500 million tweets per day. In addition, the numbers of users of other platforms increase day by day such as the total number of monthly active Facebook and Instagram users increased to 2.17 billion and 800 million users respectively (Kemp, 2018). Another area, shopping from e-commercial sites, gives good opportunities to trade everything over the internet. This enables people easily to purchase or pay bills online. According to the latest statistical information, the 17% of the overall human population aged 15+ makes online purchase and/or pays bills over the internet (Kemp, 2018). As a result, people who make shopping from these sites also share their comments about the product they purchase. All these actions can be varied to many examples which have a common point. It is that tremendously large data is generated as a result of using such applications and it expands day by day.

The tremendous expansion of the raw data comes with new problems to be solved such as extracting meaningful knowledge about a target entity. Here, target entity can be anything such as a product, a person, an activity and so on. Extraction of a meaningful knowledge of a target entity is required in order to answer some questions. For example, a simple question sentence might be like that “What is being said about the book that I plan to buy?” or the question can be different such as “Is this book good or bad?” or another question can be more complex when the extraction of emotions about a target entity is asked. Here, the struggle with this problem is that this raw data is in a non-structured format which requires applying very complex and expensive processes to transform the raw data to a structured format which is called unstructured document categorization. In the literature, there are many datasets created to solve this problem. One of them is the TTC-3600 benchmark dataset formed for Turkish text categorization (Kılınç et al., 2015). After the categorization of the raw data based on some given constraints, many kinds of machine learning algorithms are used to categorize textual data having similar structures and meanings. These newly created category groups can be used to categorize new non-structural text files.

Machine learning is an approach that makes inferences from existing data using mathematical and statistical methods and makes predictions for the unknown with these inferences. The machine learning algorithms, applied in text categorization problems, are used for categorizing a given text. They are categorized in three learning methods which are supervised, unsupervised and semi-supervised. These learning methods are categorized according to the interaction of the algorithms with the input data. The main difference between them is the structure of the input data. In supervised learning approach, the samples in the data are completely labeled. In other words, each instance has its own label or result. Some of the most popular supervised learning algorithms are as follows: Support vector machines, neural networks, Naïve Bayes, decision tree, Random Forest etc. On the other hand, the unsupervised approach is the opposite of the supervised learning approach which means that there are no labeled instances in the input data. K-means and Apriori algorithms can be given as some examples of unsupervised learning algorithms. In addition to these unsupervised algorithms, lexicon-based approach is another unsupervised learning method which is based on spotting keywords in a target text data. The spotting process is based on pre-defined lexicon composing of significant keywords. In the semi-supervised learning, this case is the mixture of both supervised and unsupervised approaches which means that the input data contains both labeled and unlabeled instances. Even though supervised and unsupervised approaches are popular and have many algorithms, semi-supervised learning is a hot topic in image classification problems.

Two similar approaches rise to prominence when it comes to extracting meaningful information about feelings from the raw data. The first is emotion analysis and the second is sentiment analysis also known as opinion mining. In sentiment analysis, mostly classifications of categories are positive, negative and neutral. This helps to get an overall idea of a target entity out of the raw data. When it comes to obtaining more than sentiments such as basic and complex emotions, emotion extraction process takes place. These two analyses are often referred to as emotion analysis when expressed in Turkish in the literature. However, as already mentioned, one examines a number of emotions expressed in the texts, and the other examines the feelings in the texts, i.e., the negative, positive, or neutral situations. Emotion extraction can be applied to many applications such as managing customer relations (Bougie et al., 2003), assistant robots that sense human emotions (Breazeal & Brooks, 2004; Hollinger et al., 2006), extraction of emotion from newspaper headlines (Bellegarda, 2010) and so on.

1.2 Goal and Contribution of the Thesis

The supervised machine learning approach and the lexicon-based approach stand at the center of studies based on emotion analysis. Both of these approaches have a common feature, which is that they both require a labeled dataset and a lexicon. Besides, both have their own advantages. The former approach provides higher precision value than the latter approach. On the other hand, lexicon-based approach is still competitive, as it is not sensitive to the quantity and quality of the training dataset (Hailong et al., 2014).

It is primarily necessary to have a labeled dataset or a pre-constructed lexicon for emotion analysis when supervised machine learning or lexicon-based approaches are decided to be used. In the literature, there are many dataset and lexicons for the purpose of emotion analysis. Most of these data are constructed in English. When it comes to other languages, the presence of this type of data is very few. The need of this kind of data is solved mostly by translating the data to the required language.

Turkish is one of these languages which is lack of labeled datasets and pre-constructed lexicons for emotion analysis. To fill the gap in this area, the motivation of this thesis is to create a new training dataset and a lexicon for the purpose of emotion analysis in Turkish. We decide on the Ekman's six basic emotions such as joy, fear, anger, sadness, disgust and surprise. The reason for the selection of this emotion category list is because it is one of the most used emotion list for emotion analysis studies in the literature such as (Demirci, 2014; Aman & Szpakowicz, 2008; Strapparava & Valitutti, 2004), and the low number of emotion categories eases the collection and annotation progress of the data.

This thesis addresses the following research questions:

- How to construct the first training dataset to be used at supervised machine learning methods for emotion analysis in Turkish?

- How to construct an emotive lexicon to be used at keyword-spotting approach for emotion analysis based on six basic emotion categories in Turkish?

There are two important contributions of this thesis to the literature. The first and the most important one is the Turkish Emotion Dataset (TREMO), which is labeled for Ekman’s six basic emotion categories (Ekman, 1992). To the best of our knowledge, TREMO is the first generated dataset for emotion analysis processes in Turkish, which is not a translation from other languages. To construct this dataset, we first conducted a survey and obtained 27,350 entries from 4,709 individuals. To validate this raw dataset, we performed a validation process where 48 annotators voluntarily participated. Here, a total of 92,986 individual annotations were made, and at the end, 1,361 entries were discarded from the raw dataset owing to ambiguities in the emotion categories. Then, we applied a set of machine learning algorithms to evaluate the effects of the validation process, stemming methods and term selection approaches. The second contribution of this thesis is a special lexicon, TREMO_LEX, which is required in emotion analysis studies in Natural language text using key-spotting technique. TREMO_LEX is a list of emotive keywords selected with different weighting coefficients and term-enrichment methods. It is the first Turkish lexicon constructed by using a Turkish emotive text, TREMO, which is not a translation. It is constructed in full and reduced sized for considering efficiency as well as effectiveness of emotion analysis problems.

1.3 Organization of the Thesis

This thesis is composed of five main chapters. In chapter one, the introduction of the thesis is given. In this chapter, the motivation, goal of the thesis, research questions and the contribution of the thesis are discussed. The remaining chapters of this thesis are explained briefly below;

In the second chapter categorized related works and the definitions of some basic terms within the thesis’s research area are given.

In the third chapter, the detailed explanation of the data collection and the validation phases of TREMO dataset are given. The methods of these phases are explained in details.

In the fourth chapter, the explanation of the steps taken to create the lexicon, TREMO_LEX, is shared in details.

In the fifth chapter, the supervised learning methods used for implementing emotion analysis on TREMO dataset are shared. The classification results are explained in details. In addition, to compare the performances of the newly created lexicon, a set of experiments, based on keyword-spotting method, are applied on a test dataset.

In the last chapter, the conclusion of the thesis is given sharing the overall results and the contribution of them to academic areas. Then, the future works are discussed.

CHAPTER TWO
RELATED WORKS

2.1 What is Emotion?

Nowadays the amount of text data has grown so rapidly by the use of social media applications and others. This condition makes hard to reveal valuable knowledge of a target entity because of large amount of unstructured data. There are many possible outputs that can be extracted out of this tremendous amount of text raw data. One of these is emotions in the text. Emotions are very important to study because they are undisputed parts of the lives of humans as they are a part of humans from the beginning of this world. In other words, they are innate. Some researchers claim that basic emotions are same in facial expressions no matter in which culture those humans belong to (Ekman & Friesen, 2003). However, there are other studies saying that even though there are some similarities between emotion expressions among different cultures and language, these differences make a huge effect on shaping emotions (Elfenbein & Ambady, 1994; Russell, 1994). On the other, there is a debate going on whether animals have emotions or not. There are several studies concentrated on higher mammals, canines, felines, and even some fish for this purpose (Masson, 1996; Guo et al., 2007). One of the oldest studies for the purpose of revealing emotion among humans and animals is accomplished by Charles Darwin with his book The Expressions of the Emotions in Man and Animals (Darwin, 1872). Evolutionary biologists and psychologists discovered that emotions manage to evolve in time among species in order to improve productivity as they are very crucial for species to survive. To give an example to this case, fear triggers the capability of fight or flight. As the primates and humans have more complex brain structures, they are capable of experiencing more complex emotion types such as optimism and shame other than basic emotions such as anger and surprise. In addition to emotions, mood is another fact which can be considered as a tool to evaluate how well a living being is. These two facts, emotion and mood, are expressed as affect (Scherer, 1984; Gross, 1998; Steunebrink, 2010). Even though, they seem to be similar, there are some differences between them. The duration time of mood is much longer compared to emotions. The second difference is that generally emotions are focused on one specific thing but the mood is more diffuse on the other hand (Nowlis & Nowlis, 2001; Gross, 1998; Steunebrink, 2010), (Mohammad & Turney, 2012).

Several theories were proposed about classifying emotions into taxonomies by psychologists in the literature. Generally, emotions are divided into two main groups, basic and complex. But there are some psychologists said that, emotions are divided into two groups instinctual and cognitive. The emotions are considered instinctual when we can sense and perceive them. The others classified as cognitive are those we can obtain by thinking and reasoning (Zajonc, 1984). However, this distinction is not accepted by other psychologists and they think that emotions cannot be a result of cognition (Lazarus, 1984, 2000). According to (Plutchik, 1985), there can be no end for these discussions because of not having concrete experimental proofs. He thinks that it is only a problem of definition (Mohammad & Turney, 2012).

In the literature, there are several studies proposing basic emotions (Ekman, 1992; James, 1884). In one of these studies, Ekman mentioned about six basic emotions which are joy, sadness, anger, fear, disgust, and surprise (Ekman, 1992). On the other hand, Plutchik decided on a group of eight emotions which is including the six emotions of Ekman and also adding two new emotions which are trust and anticipation (Plutchik, 1980, 1994). Plutchik used a wheel to explain the relationships between emotions. In this wheel of emotions, radius specifies the intensity of the emotions. In other words, emotions in the center are more intensified than the outer ones. According to Plutchik’s eight emotions, there are four opposite pairs in meanings which are anticipation-surprise, joy-sadness, trust-disgust, and anger-fear. These pairs are placed opposite to each other in Figure 2.1. In this Figure, it is also possible to see the combinations of emotions which are contiguous to each other. These combination emotions, which are also called primary dyads, are placed in the white spaces between the basic emotions. Even though, it seems that there are concrete boundaries between emotions in Figure 2.1, generally it is not that easy to define clear boundaries between them (Mohammad & Turney, 2012).

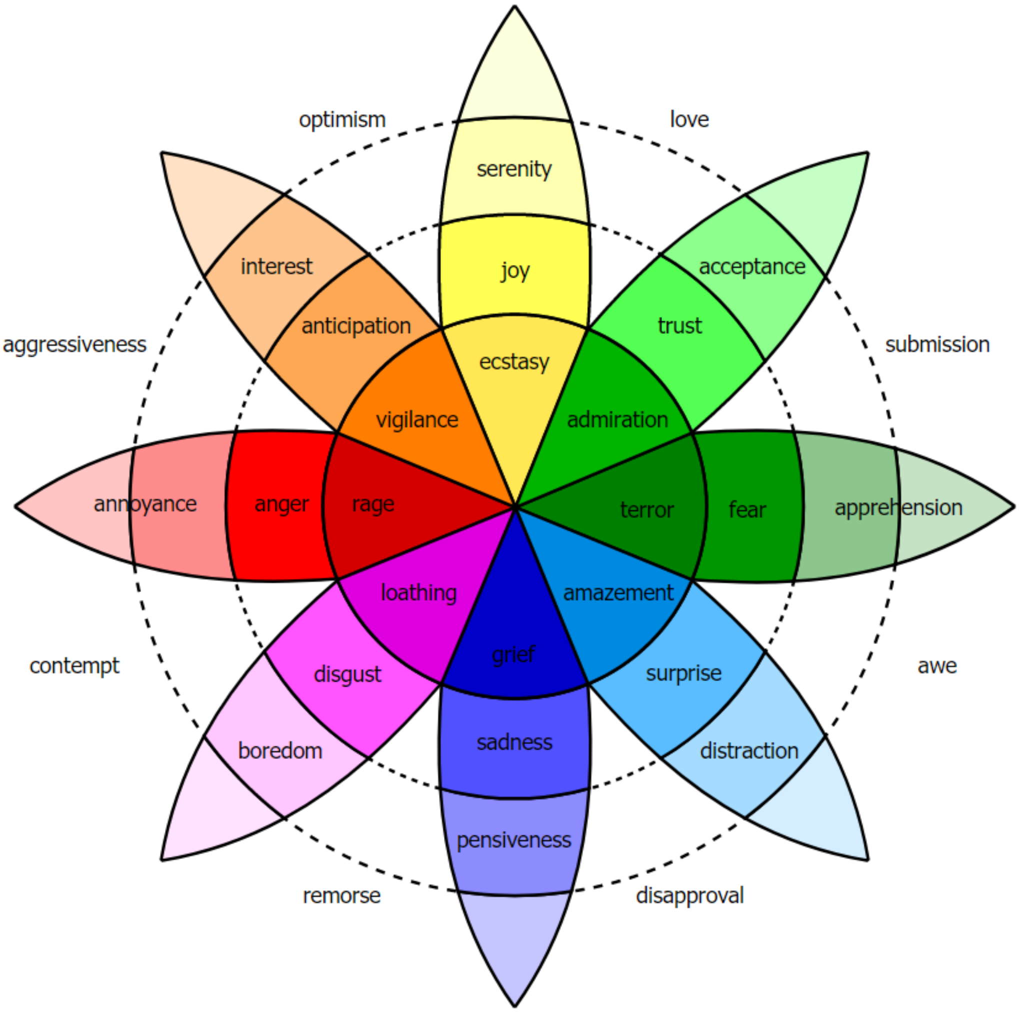


Figure 2.1 Plutchik’s wheel of emotions (Wikimedia Commons, 2018)

2.2 Sentiment

To understand the content of the enormous amounts of data it is important to answer the question “Is something good or bad being said about the target entity?” and “Is the speaker happy with, angry at, or fearful of the target?” The answer to these questions indicates the area of sentiment analysis. It focuses on observing the opinions and private states of a person about a target entity. These private states can be defined as feelings, beliefs and speculations (Wiebe, 1994). In other words, sentiment analysis is used to extract positive or negative polarities of an entity out of a word, phrase or document. Here, positive polarity means favorable sentiment towards an entity and negative polarity indicates unfavorable sentiment towards an entity (Turney & Littman, 2003; Pang & Lee, 2008). Sentiment analysis is applied on many applications nowadays. For example, managing customer relations where the system analyze the tension of the speaker and transfer him/her to a higher-level manager who can help solve the problem. In another type of application, the companies look forward to analyze the feedbacks of the customers about their products automatically on blogs, forums and social media applications.

2.3 Studies Based on Sentiment and Emotion Analysis

In the literature, there are many studies implemented in order to analyze emotion and sentiment out of text. To do so, most of the time the researchers utilized from two main techniques which are machine learning techniques and symbolic techniques. Machine learning techniques mainly composed of unsupervised, weakly supervised and fully supervised where a training dataset is required. On the other hand, for symbolic techniques, pre-defined rules and lexicons are used in (Boiy et al., 2007). To use these techniques properly, researchers utilized mostly from labeled lexicons and datasets which are pre-constructed or newly constructed by the researchers for the use of their studies. As all these studies are applied on texts, an important decision has to be made on the language. Most of the studies in the literature are concentrated on texts written in the English language. On the other hand, there are also studies which are applied on other languages such as Turkish, Chinese, Arabic and etc.

Within the scope of this section, we categorize the related works according to two main techniques which are machine learning techniques and symbolic techniques. For the sub categorizations, we categorize studies based on emotion and sentiment analyzes and we focus on whether these studies create ground-truth data which can be dataset or lexicon, within the scope of them.

2.3.1 Studies Based on Machine Learning Technique

2.3.1.1 Sentiment Analysis Studies

In the study (Alm et al., 2005), they used supervised machine learning approach with SNoW learning architecture. They utilized from a fairy tales dataset with a lexicon called WordNet Affect (Strapparava & Valitutti, 2004). Within the scope of the study, they focused on sentiment analysis classifying for positive, negative and neutral categories and also on classification results on the existence of emotions.

Go et al. (2009) worked on classifying tweets as either negative or positive. They created their dataset by gathering tweets according to emoticons that indicate the differences between positive and negative emotions. They used machine learning algorithms, Naïve Bayes, Maximum Entropy and Support Vector Machines for classifying Twitter messages. The accuracy results of each of these machine learning algorithms are above 80%.

Eroğul (2009) focused on sentiment analysis in Turkish classifying the movies as negative or positive. This analysis is accomplished by using Support Vector Machine algorithm, which is one of the most effective supervised machine learning technique. To apply this algorithm a dataset is created which is composed of labeled movie reviews and rating values. In the preprocess stage, Zemberek stemming approach is used for stemming. For term selection, the author decided on a threshold value. In addition he used n-gram model where unigrams, bigrams and trigrams are examined. In overall, the calculated classification accuracy value is 85%.

Kouloumpis et al. (2011) evaluated effects of using features on Twitter sentiment analysis, focused on sentiments positive, negative and neutral, using supervised learning approaches. To achieve this goal, they used three different corpora of Twitter messages. Two of these, hashtagged (HASH) and emoticon (EMOT) datasets were used as training datasets. For testing their model, they benefited from a dataset called iSieve which was annotated manually by iSieve Corporation.

Albayrak (2011) studied on sentiment analysis in Turkish text in a different point of view. Instead of positive and negative sentiments, the author focused on different sentiments such as depressed, non-depressed, anxious and non-anxious. As these sentiment types indicate psychological states of the people, this analysis also called psychological text analysis which is named as Text Investigator for Psychological Disorders (TIPD). For constructing feature vector, the author used different feature types which are bag of words, mainly used words in each group of documents, frequency of tenses and pronouns. For morphological analyzer, the author used Zemberek. In the classification stage, Weka tool is utilized to implement a system which uses Naïve Bayes and support vector machines as classifier algorithms.

In another study, sentiment analysis is applied to Turkish political columns news. The performance of four supervised machine learning algorithms is compared between each other. Maximum Entropy and N-gram language model outperformed the algorithms SVM and Naïve Bayes (Kaya et al., 2012).

2.3.1.2 Emotion Analysis Studies

Yang et al. (2007) worked on emotion classification problems for four emotion categories, joy, happiness, sadness, and fear. They used blog posts and emoticons as training datasets and focused on comparing the results obtained by support vector machines and conditional random field classifiers.

In the study (Aman & Szpakowicz, 2007), an annotated corpus is created for the use of emotion analysis. They evaluated this annotation process by using Cohen’s kappa statistics (Cohen, 1960) to measure the agreements for emotion categories.

Danisman & Alpkocak (2008) worked on ISEAR dataset (Scherer & Wallbott, 1994) with a classification model Vector Space Model (VSM). In their model they decided to use five different emotion categories, which are anger, disgust, fear, sadness and joy. They used and compared three different classifier types, which are Naive Bayes, Support Vector Machines and Vector Space Model. In the preprocess section of the study, they stemmed and removed stopwords. As weighting schema, authors applied tf×idf weighting over the term document matrix. Additionally, they added several emotional words into training dataset from Wordnet-Affect and WPARD in order to improve the results. In the evaluation part, they used 10 folds cross validation technique. The accuracy results obtained out of this study was calculated as 70.2%.

Aman & Szpakowicz (2008) worked on a study where they applied machine learning approach on a data collected from blogs. Ekman's list of six basic emotions is used for emotion categorization. The lowest precision value is calculated as 0.318 for surprise and the highest precision value is 0.824 for fear.

In the study (Strapparava & Mihalcea, 2008), the emotion categorization for six basic emotions is applied on news headlines. To be used in this purpose, they created an annotated dataset composed of news headlines. In the evaluation process, a variation of Latent Semantic Analysis (LSA) and Naïve bayes classifier are used. The results for each emotion categories are varied according to classifiers.

Chaffar & Inkpen (2011) focused on emotion analysis of six emotion categories from a text by using a dataset which is heterogeneously composed of news headlines, fairy tales, and blogs. They adopted supervised machine learning techniques and features, such as bag-of-words and N-grams.

Boynukalin (2012) studied on the area of emotion analysis for Turkish text and applied several machine learning approach. She used two datasets, which are Turkish translation of ISEAR dataset and Turkish fairy tales for analyzing four emotion categories which are joy, sadness, anger, and fear. The author preprocessed the dataset by stemming and removing stop words. For weighting schema, they used three approaches which are Presence-nonpresence, term frequency (tf) and tf×idf. In the classification process, they used three different classification methods which are Naive Bayes, Complement Naive Bayes and Support Vector Machine. These methods were applied on the dataset by using WEKA. According to results, Complement Naive Bayes gave the best results. They also used 10-fold cross validation for evaluating the system and obtained the accuracy values of 81.34% for the ISEAR dataset with four classes, 76.83% for the Turkish fairy tales with five classes and 80.39% for the combination of the two datasets with four classes.

Calvo & Kim (2013) proposed a dimensional model of emotions which can be used for visualizing emotions in a psychologically meaningful space and for emotion detection tasks. They stated that three-dimensional space of valence, arousal and dominance can be used to represent emotions better. They compared the results of proposed model with the statistically driven techniques in four datasets which are SemEval Affective Text data (Strapparava & Mihalcea, 2007), ISEAR dataset (Scherer & Wallbott, 1994), fairy tales and USE (Unit of Study Evaluations). The results showed that there are no big differences between the proposed model and the categorical model.

Demirci (2014) studied on emotion extraction from Turkish micro-blog entries. She focused on gathering tweets for the six emotions anger, disgust, fear, joy, sadness, and surprise using the Twitter search mechanism for hashtags. Demirci defined hashtags containing the derivatives of each emotion word for each emotion category. As a result, Demirci collected 1,000 tweets for each emotion which makes 6,000 tweets in total.

In the study (Tocoglu & Alpkocak, 2014), they presented an emotion extraction system to be used in Turkish text. The system is able to recognize seven emotional states from a given text for happy, shame, guiltiness, disgust, sadness, angry and fear categories. They considered emotion extraction as a text classification problem, which requires a training set. Thus, they obtained the required training set which is collected by a survey conducted among 500 university students where they are asked to describe their most intense moments they remember for seven emotions categories (Açıcı, 2012). Then, the text describing emotional moments are preprocessed and modeled in Vector Space Model where tf×idf weighting schema is used. Then they applied Naive Bayes classifier and tested with 10-fold cross validation, in WEKA tool. They evaluated the system in terms of accuracy, precision, F-Measure and recall measures. The results they obtained are very promising where it is around 86% accuracy for all of the seven emotional classes in average.

2.3.2 Studies Based on Symbolic Techniques

Symbolic techniques are based on pre-defined rules and lexicons (Boiy et al., 2007). Most of the time, these rules and lexicons are created manually. As we focused on the creation of an emotional lexicon within a part of this thesis, we mainly shared studies based on lexicon-based approach. In addition, we also discussed some studies based on rule-based approach.

Lexicon-based approach is an unsupervised classification method to extract sentiments from a given text. So it does not require a labeled training dataset. Instead of a training dataset, the main requirement for a lexicon-based approach is a well-constructed lexicon. The quality of the lexicon plays an important role in the efficiency of this approach. A lexicon can be constructed by two ways which are manually (Stone et al. 1966; Tong, 2001) or automatically (Hatzivassiloglou & McKeown, 1997; Turney, 2002; Turney & Littman, 2003). When the lexicon is created automatically, it is enlarged by using a pre-defined list of seed words. Most of the time, adjectives are used in lexicon-based studies because they are considered as important indicators for extracting sentiments (Hatzivassiloglou & McKeown, 1997; Hu & Liu, 2004; Wiebe, 2000; Taboada et al., 2006). Instead of adjectives, there are also several studies using two-word phrases (Turney, 2002), adjective phrases (Whitelaw et al., 2005) and adjectives with adverbial modifiers (Benamara et al., 2007).

2.3.2.1 Sentiment Analysis Studies

Hu & Liu (2004) have utilized the lexicon-based approach by using the words that they have received from WordNet (Fellbaum, 1998). The purpose of this study is to summarize all the customer reviews of a product as positive or negative. The method is based on mining product features, identifying opinion sentences and calculating an opinion about a product. Given the results obtained, besides its deficiencies, this method was found to be easy and effective.

Ding et al. (2008) tried to identify positive, negative, or neutral views from the comments made on the product sales. Within the scope of the study, they proposed a holistic lexicon-based approach and created an application named Opinion Observer. The purpose of this approach is to classify the context-dependent view words. The study also deals with many specific words, sentences and language structures that influence views based on the linguistic patterns. Finally, the method they developed is supposed to collect the words from multiple contradictory views found in a sentence.

Taboada et al. (2011) used a lexicon-based approach to perform a sentiment analysis in texts. They developed an application called SO-CAL to perform these analyses. This application uses lexicons labeled with semantics (polarity and strength) and includes negativity and concentration. The SO-CAL is applied in the polar classification task, which means doing positive or negative labeling to a relevant text in order to get the main idea of a given text. They received support from the Mechanical Turk service to check the consistency and reliability of the lexicons used by the SO-CAL application. According to the results obtained, SO-CAL performed well in blog posts and video game reviews.

Nielsen (2011) has created a new lexicon, called AFINN, by using twitter data. AFINN Lexicon consists of more than 2,000 words to be used in identifying opinion-related terms. As the values of the polarity, he gave positive values between 1 and 5 for the words representing the positive information, and between -1 and -5 for the words representing the negative information.

Xie & Li (2012) focused on creating a domain-independent and corpus-related lexicon to be used in sentiment analysis problems where positive, negative and neutral sentiments are focused. The newly created lexicon is called corpus-related because it is constructed based on the related corpus. To achieve this goal, they proposed a new probabilistic modeling framework, Tag Sentiment Topic Model (TSTM), which is based on Latent Dirichlet Allocation (Blei et al., 2003). In addition, TSTM model requires two prior knowledge word sets which are positive and negative word lists. Both of the lists contain seven terms. The positive words are as follows; excellent, good, nice, positive, fortunate, correct and superior. The negative words are as follows; nasty, bad, poor, negative, unfortunate, wrong and inferior (Turney & Littman, 2003).

Akbas (2012) focused on opinion mining by extracting aspects of entities on Turkish tweets. The author utilized from a Turkish opinion word list constructed manually and proposed a word selection algorithm to automate new words with their sentiment strengths.(Sevindi, 2013) translated SentiWordNet (Baccianella et al., 2010) lexicon to Turkish and created a Turkish sentiment lexicon with a term size of 12697.

Vural et al. (2013) created a framework for unsupervised sentiment analysis in Turkish. They created their own lexicon to be used in the framework by translating the lexicon of SentiStrength sentiment analysis library (Thelwall et al., 2010).

Musto et al. (2014) have developed a new lexicon-based approach called the fine-grained approach, which performs sentiment analysis. Basically, this approach is based on dividing a given tweet into parts called micro-phrases by taking into account the division clues they obtained from the text. Punctuation marks, adverbs, and conjunctions were used as division clues. Thus, the sentiment expressed by a Tweet T is defined as the sum of polarities expressed by each of the micro-expressions that create it. In their study, they developed four different methods for the approach they described. They used four different lexicons when comparing these methods. They performed tests on SentiWordNet (Baccianella et al., 2010), WordNet-Affect (Strapparava & Valitutti, 2004), MPQA (Wiebe et al., 2005), SenticNet (Cambria et al., 2014) and two important datasets, SemEval-2013 (Nakov et al., 2013) and Stanford Twitter Sentiment (Go et al., 2009).

Dehkharghani et al. (2015) developed a new Turkish lexicon with a semi-automatic method to perform sentiment analyses in Turkish language. The newly created lexicon, SentiTurkNet, is the first Turkish polarity resource in literature by assigning three polarity values, positive, negative and neutral, to 14795 one or more synonyms (synsets) clusters found in Turkish WordNet (Bilgin et al., 2004).

Awwad & Alpkocak (2016) focused on sentence-level and document-level sentiment analysis by using lexicon-based approach. To do so, they used four pre-defined lexicons which are as follows: Harvard IV-4 Dictionary (HarvardA), MPQA subjectivity lexicon (HRMA) (Elarnaoty et al., 2012) and two different versions of MPQA. As this study focuses on Arabic language, they translated all these lexicons to Arabic. To evaluate all these four lexicons, they utilized from three datasets from different domains. The first one is PatientJo which is about health comments. The second one is Twitter data (Elarnaoty et al., 2012) and the third one is about book reviews (LABR) (Aly & Atiya, 2013). The overall results showed that there are no big differences in terms of performances between sentence-level and document-level. Although, the performance of each lexicon differs in terms of datasets. HRMA lexicon performed highest result compared the others when LABR dataset is used. On the other hand, HarvardA performed better in PatientJo dataset. They also found out that giving extra weights to first and last sentences in sentence-level approach provides improvements in accuracy results which is the case also mentioned in (Dehkharghani et al., 2015).

In another study, Ucan et al. (2016) proposed an automated translation approach to construct sentiment lexicons for new languages by using English resources. At the end of their study, they achieved to construct three different lexicons for Turkish.

2.3.2.2 Emotion Analysis Studies

Stone et al. (1966) have created a lexicon named GI by tagging 11788 words with 182 tag categories. There are positive and negative categories as semantic orientation within these 182 tagging categories. Apart from this, there are also the categories of pleasure, arousal, feeling and, pain.

Strapparava & Valitutti (2004) have created a lexicon named WAL which is abbreviated form of WordNet-Affect Lexicon. The WAL lexicon is an affective extension of the WordNet lexicon. It was created by using several hundred core-words tagged by certain emotion categories. The process at this stage is to find the synonyms of the core-words in the WordNet lexicon and assign them the emotional type of the relevant core word. In the resulting lexicon, 1536 words are linked to Ekman’s (Ekman, 1992) six emotion categories.

Katz et al. (2007) proposed two SemEval-2007 entries in their study. The second entry they proposed is about annotating the emotional content of news headlines.

They focused on emotion categories: Anger, disgust, fear, joy, sadness, and surprise. To predict these emotion categories from news headlines, they implemented a supervised system which uses a unigram model. In this system each headline is scored along seven axes which are the six emotion categories on a scale from 0 to 100 and positive/negative polarity (valence) value on a scale from -100 to 100. For the training set, they annotated 1250 news headlines. In the preprocessing stage, they lemmatized each headlines to reduce the sparseness of the data. They used CELEX2 (Baayen et al., 1996) data in this lemmatization process. Then, they constructed word-emotion mapping by scoring the emotions and valence of each word as the average of the emotions and valence of every headline. After the creation of the word-emotion mapping, they predicted the emotion and valance value of a given headline. This study gave promising results on SemEval Affective Text Task (Strapparava & Mihalcea, 2007).

Mohammad & Turney (2012) have created a new lexicon, EmoLex, which contains 14182 words in total. It was created by considering eight emotion categories which are anger, disgust, fear, expectation, joy, sorrow, surprise and confidence (Plutchik, 1980).

In another study, Mohammad (2012) generated a large dataset with 11418 words for Ekman’s six emotion categories by using data he obtained from Tweeter. At this stage, he decided to use the names of the hashtags for deciding whether the tweets have emotional contents or not. In addition to this study, he has created a new emotion-based lexicon using the same dataset.

Mohammad (2012) focused on determining whether word-emotion association lexicons yield better results than using n-gram features. In addition, he found that emotion lexicon features yield better results in new domains than using n-gram features. To achieve these steps, he used two-emotion lexicon features annotated for Ekman’s six emotions, WordNet Affect Lexicon (Strapparava & Valitutti, 2004) and NRC-10 (Mohammad & Turney, 2010). For his training dataset, he chose to use the SemEval-2007 Affective Text corpus (Strapparava & Mihalcea, 2007).

Yang et al. (2014) proposed an emotion-aware LDA (EaLDA) model to create fine-grained domain-specific lexicons for languages. EaLDA is an extended version of the Latent Dirichlet Allocation (LDA) (Blei et al., 2003). The newly constructed lexicon is used for emotion classification purposes for six emotion categories, which are joy, anger, disgust, surprise, sadness and fear. The EaLDA model requires a set of domain-independent emotion terms to be used in grouping semantically related words together. This feature enables the newly created lexicons to be more adaptive.

In the study (Mohammad & Kiritchenko, 2015), an emotion dataset named Hashtag Emotion Corpus was created by using the hashtag structure in Tweeter. Here, the process of emotion identification is based on the names of the hashtags. In the first step, this dataset started with six emotions, and later covered 585 emotions. In the next step, they created a lexicon named Hashtag Emotion Lexicon from the Hashtag Emotion Corpus dataset.

2.3.2.3 Rule-Based Approaches

Text-to-Emotion engine is developed in order to analyze and extract emotions categories, Ekman’s list, from texts generated by chatting. To do so, the author used rule-based approach. In this approach, a set of rules are defined and applied to text in order to calculate a score for analyzing an emotion category (Boucouvalas, 2003). In another study, same approach is used to obtain classification results from on-line communication environments (Neviarouskaya et al., 2011). Another rule-based approach is implemented by extracting triplets of each sentence of a given text. These triplets are subject, verb and object. To these triplets, also the existing adjectives and adverbs are attached as attributes. In addition, in the same study, a lexicon, which contains word-valence pairs, is also used to calculate a valence value of a sentence, which indicated sentiment of the corresponding sentence (Shaikh, 2008). In another study, an emotion model OCC is created. OCC stands for the initial characters of surnames of the authors which are Ortony, Clore, Collins (Ortony et al., 1988). This model can be applied to emotion categories such as sorry for, hope, fear, etc. The OCC model uses a set of pre-defined rules to evaluate the triplets which are events, agents, and objects. In this study, the authors utilized from a semantic parser to extract agents and events. The same approach is also used for the purpose of emotion extraction in another study (Yashar, 2012).

2.4 Conclusion

The existence of lexicons and datasets are the corner stones in sentiment and emotion analyses. For this reason, in the literature the researchers focused on using and creating such data to be used for both sentiment and emotion analyses. The number of datasets and lexicons are high enough in English language. When it comes to the existence of datasets and lexicons for sentiment and emotion analysis in Turkish, the number of data is low.

Most of the studies used non-Turkish datasets translated into Turkish, such as Boynukalin used a portion of the ISEAR dataset containing documents for four emotions translated into Turkish (Boynukalin, 2012). In another studies Vural et al. (2013) and Sevindi (2013) created their own lexicons by translating pre-constructed lexicons. On the other hand, there are some studies where the translated data is annotated manually. For example, Dehkharghani et al. (2015) focused on creating a new polarity resource, SentiTurkNet, by assigning three polarity values, positive, negative and objective, to 14795 synsets found in Turkish WordNet manually. In another study Ucan et al. (2016) managed to create a word-level sentiment lexicon automatically for Turkish language. To do this, they translated a well-known annotated lexicon, SentiWordNet, by using a multiple bilingual translation approach which contains three different algorithms.

In other studies, datasets are collected from social media applications such as Twitter. Demirci (2014) focused on extracting emotion from Turkish micro-blog entries. She collected tweets for the six emotions anger, disgust, fear, joy, sadness, and surprise using the Twitter search mechanism for hashtags. For each emotion category, Demirci defined hashtags containing the derivatives of each emotion word. As a result, Demirci succeeded in collecting 1,000 tweets for each emotion, 6,000 tweets in total. In another study, Akbas (2012) also utilized from Twitter data by labeling tweets manually between two sentiments positive and negative. As a result, the author managed to create a gold standard corpus composed of labeled tweets.

There are also studies created corpuses for the use of analyzing sentiments in different domains such as Turkish political columns, movies and fairy tales. Kaya et al. (2012) focused on extracting negative and positive sentiments out of political columns. To do so, they collected a total of 400 political columns from 6 different Turkish newspapers, 200 positive and 200 negative. Afterwards, they annotated these political columns by three native speakers of Turkish. Eroğul (2009) collected a set of labeled movies as negative, positive or neutral from a Turkish movie web site, where the users enter their comments about movies by stating their opinions with icons. In another study, Boynukalin (2012) collected 25 children's Turkish fairy tales from several web sites and annotated them manually.

Within the scope of this thesis, we propose a Turkish dataset and a lexicon for emotion analysis. To the best of our knowledge these two data collections are the first dataset and the lexicon for emotion analysis in Turkish. The novelty in proposed dataset is that it is collected by a survey among 4,709 participants where each individual is asked to share their memories or any experiences they would have for six emotion categories. In addition to the collection process, the validation of the collected data plays an important role in the preparation of the dataset. The novelty in the lexicon is that the use of proposed dataset as the source of the terms composing the lexicon. In addition to this, the methods which are used to generate the lexicon increase the value of it.

CHAPTER THREE
TREMO: A DATASET FOR EMOTION ANALYSYS IN TURKISH

3.1 Data Gathering

In the literature, there exist well-formed and labeled datasets for the use of classification techniques. However, the total number of these datasets is not in large numbers due to difficulties faced in their creation processes. In the creation process of these datasets there are several steps to take. First of all, the language and the categories of the dataset should be determined. In the literature, most of the datasets are in English and used basic emotion categories. In the next step, the source of the dataset must be determined. The raw data can be collected by making surveys or from social media applications such as Twitter. In the last step which is optional due to expensive and time-consuming reasons, annotation process takes place for the collected dataset by asking each instance of the dataset to people for their opinions.

Within this section of the thesis, we propose a well-structured and labeled dataset to be used in emotion analysis in Turkish. To do so, we conducted a survey with the participation of 5,000 people from different living areas and different age ranges to collect a dataset based on six emotions. In this survey, we asked participants to share their memories or any experiences they would have for the six emotion categories that Ekman described (Ekman, 1992). As a result of this process, a total of 4,709 people were approved for participation, and a total of 27,350 entries were collected. Table 3.1 shows general statistical features for TREMO dataset. Term count in entries feature shows the amount of terms within an entry. On the other hand, term character length feature provides the character size of each term in the dataset. Another feature named sentence count in entries shows the amount of sentences within an entry. The last feature called sentence character length provides the character size of each sentence in the dataset.

Table 3.1 General statistical features for TREMO dataset

Feature	**Min**	**Max**	**Average**
Term count in entries	1	68	5.9
Term character length	3	12	7.45
Sentence count in entries	1	16	1.05
Sentence character length	3	255	43.3

Participation in this survey was conducted either through a website or by manually filling in the fields for each emotion category in a given paper which is shown in Figure 3.1. In Table 3.2, the participation rate using the first method is very low compared to that of the second method. The most important reason for this is that we were not able to obtain the number of participants we had planned to collect on the web-based method. So, we also conducted the survey at high schools and universities. Of course, in this case, we had to give each participant a paper to fill out. This put an additional burden on us, as we had to enter each paper into the system. The same table also shows the female-male distributions of the participants. There were more female participants than males.

Table 3.2 Distribution of attendance types in the survey

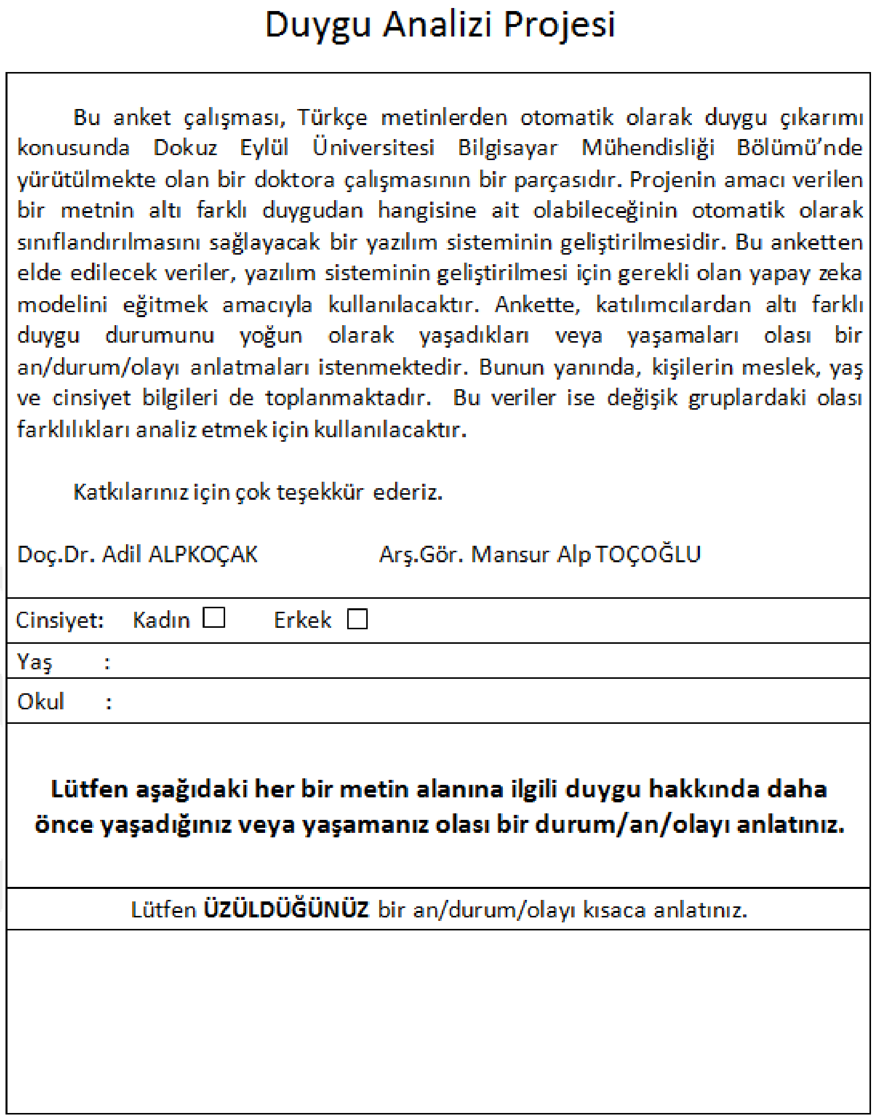
Attendance type	**Attendance number**	**Female participant**	**Male participant**
Web-based	673	392	281
Paper-based	4,036	2,378	1,658
Total	4,709	2,770	1,939

The ages of the most of the participants were between 15 and 24 as a result of having many participants from high schools and universities. Table 3.3 shows the distribution of the participants in five different age groups. Most participants were positioned in two age groups, 14-20 and 21-30, that comprise the normal age range for a student.

Table 3.3 Distribution of individuals in the survey according to age groups

Age ranges	**# of individuals**
14-20	3,854
21-30	531
31-40	157
41-50	101
51-70	60

As noted, we conducted this survey at several universities and high schools. The majority of these are educational institutions located in Izmir. There are also high schools in different cities in Turkey, including Ankara, Balıkesir, and Diyarbakır, where we conducted the survey. We obtained official permission from the Provincial Directorate of National Education to be able to go to the high schools in Izmir. While making this choice, we focused on visiting high-ranked schools. Table 3.4 shows these high schools. In addition to high schools, we also went to two state universities, Dokuz Eylul University and Katip Çelebi University. Because we went to many educational institutions, 73.88% of the participants were high school students, and 15.44% were university students. The remaining participants practiced 32 different occupations.



Duygu Analizi Projesi

Bu anket çalışması, Türkçe metinlerden otomatik olarak duygu çıkarımı konusunda Dokuz Eylül Üniversitesi Bilgisayar Mühendisliği Bölümü'nde yürütülmekte olan bir doktora çalışmasının bir parçasıdır. Projenin amacı verilen bir metnin altı farklı duygudan hangisine ait olabileceğinin otomatik olarak sınıflandırılmasını sağlayacak bir yazılım sisteminin geliştirilmesidir. Bu anketten elde edilecek veriler, yazılım sisteminin geliştirilmesi için gerekli olan yapay zeka modelini eğitmek amacıyla kullanılacaktır. Ankette, katılımcılardan altı farklı duygu durumunu yoğun olarak yaşadıkları veya yaşamaları olası bir an/durum/olayı anlatmaları istenmektedir. Bunun yanında, kişilerin meslek, yaş ve cinsiyet bilgileri de toplanmaktadır. Bu veriler ise değişik gruplardaki olası farklılıkları analiz etmek için kullanılacaktır.

Katkılarınız için çok teşekkür ederiz.

Doç.Dr. Adil ALPKOÇAK   Arş.Gör. Mansur Alp TOÇOĞLU

Cinsiyet: Kadın □ Erkek □

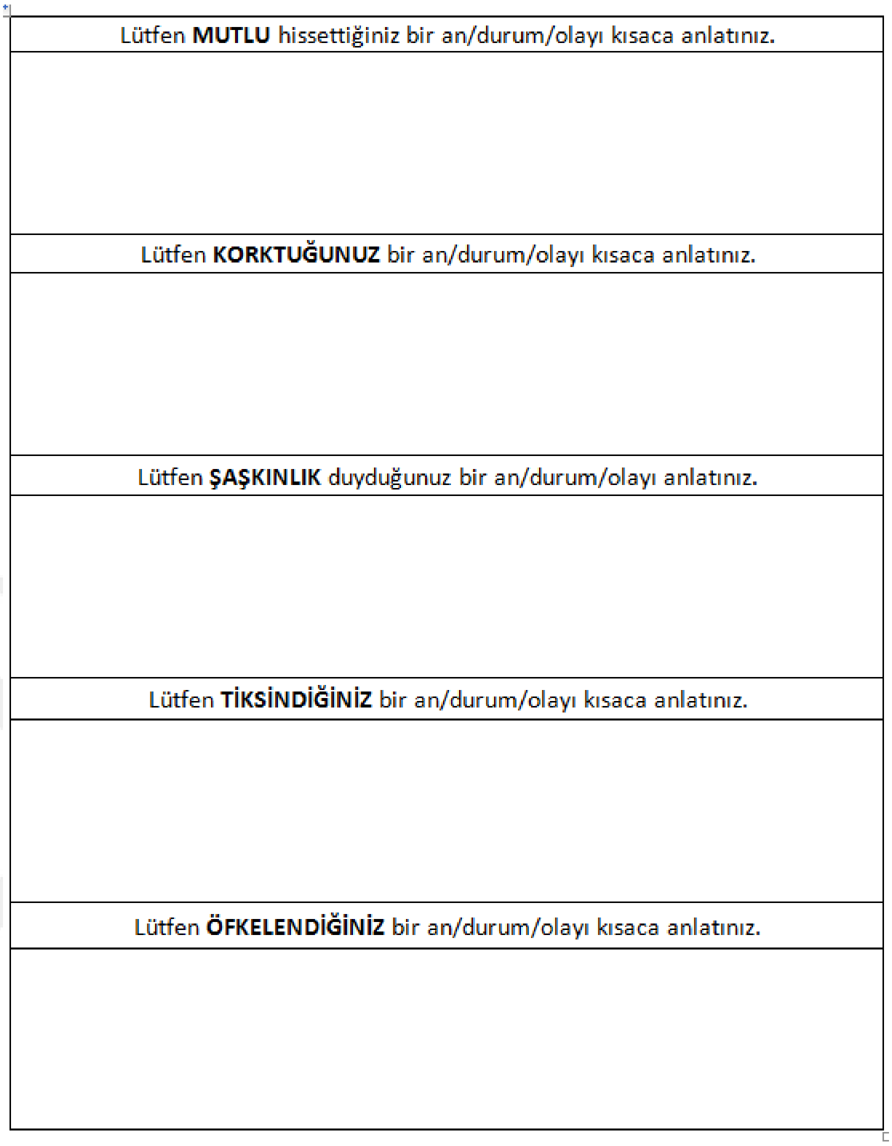
Yaş :

Okul :

Lütfen aşağıdaki her bir metin alanına ilgili duygu hakkında daha önce yaşadığınız veya yaşamanız olası bir durum/an/olayı anlatınız.

Lütfen **ÜZÜLDÜĞÜNÜZ** bir an/durum/olayı kısaca anlatınız.

Figure 3.1 First page of the survey paper



Lütfen **MUTLU** hissettiğiniz bir an/durum/olayı kısaca anlatınız.

Lütfen **KORKTUĞUNUZ** bir an/durum/olayı kısaca anlatınız.

Lütfen **ŞAŞKINLIK** duyduğunuz bir an/durum/olayı anlatınız.

Lütfen **TİKSİNDİĞİNİZ** bir an/durum/olayı kısaca anlatınız.

Lütfen **ÖFKELENDİĞİNİZ** bir an/durum/olayı kısaca anlatınız.

Figure 3.1 continues

Table 3.4 Name of the schools attended in the survey from Izmir

School name
Atatürk Lisesi
Bornova Anadolu Lisesi
İzmir Kız Lisesi
Karşıyaka Anadolu Lisesi
Karşıyaka Cihat Kora Anadolu Lisesi

Table 3.4 continues

Karşıyaka Atakent Anadolu Lisesi
60. Yıl Anadolu Lisesi
Buca Fatma Saygın Anadolu Lisesi
Buca İnci-Özer Tırnaklı Fen Lisesi

In the survey, some participants could not manage to write an entry for each emotion category. This caused differences in the distribution of the entries for each emotion. Table 3.5 shows the distribution of the entries among emotion categories.

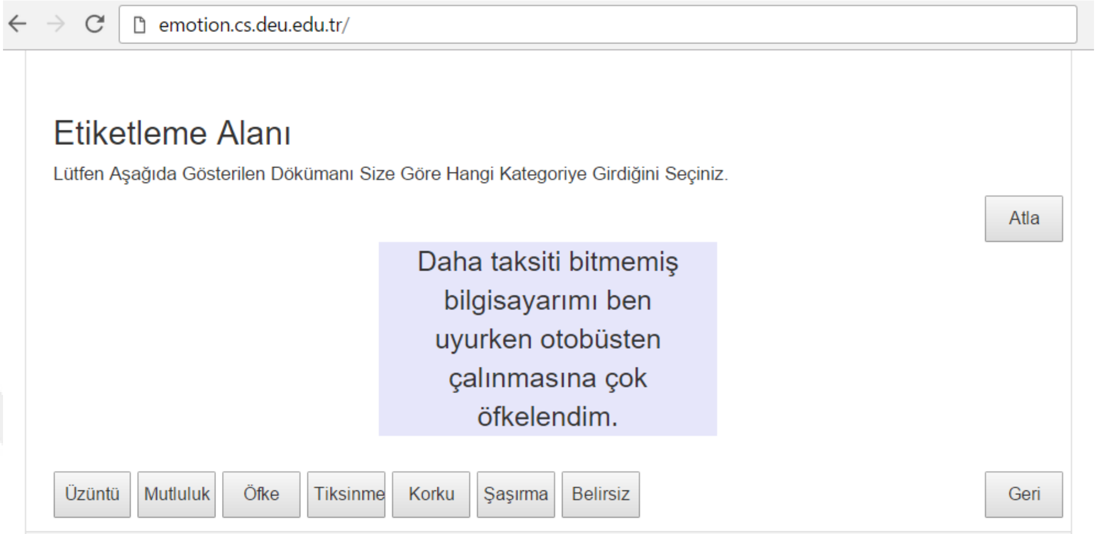
Table 3.5 Distribution of the entries for each emotion category

Happiness	**Fear**	**Anger**	**Sadness**	**Disgust**	**Surprise**
4,700	4,616	4,636	4,664	4,522	4,212

3.2 Validation of TREMO Dataset

The validation process of the TREMO dataset plays an important role to eliminate the entries which are considered as ambiguous or fake in terms of emotional categorization. If such entries are not discarded from the raw dataset, they can result in many outliers being in the training set which negatively impacts the performance of supervised learning algorithms. In the validation process, we first created an application supporting web and mobile interfaces. Annotators who want to join the validation process, first register on the system giving basic information such as name, surname, gender, occupation, age, e-mail address, and define a password, which is required by the system during annotation process. In the following stage, the annotator enters the system only after the system administrator's approval. This authorization process is designed to prevent unauthorized registrations on the system to secure the validation process. An annotator who gets confirmation can log into the system by using his or her specified mail address and password. At annotation stage, each entry is displayed in random order to annotator. The annotator simply clicks a button, representing one of the six emotion categories, to annotate the entry.

Additionally, an extra button is placed for ambiguous condition, which is used when the annotator cannot decide on suitable emotion category. Figure 3.2 shows a screen shot of the validation page in the web application.


emotion.cs.deu.edu.tr/

Etiketleme Alanı

Lütfen Aşağıda Gösterilen Dökümanı Size Göre Hangi Kategoriye Girdiğini Seçiniz.

Daha taksiti bitmemiş bilgisayarımı ben uyurken otobüsten çalınmasına çok öfkelendim.

Figure 3.2 Screen shot of the validation page of the web application

Each entry is presented to at least three different annotators. If three of them annotate the same emotion category, then we assume that the entry is validated. If not, then the entry is presented to different annotators until reaching three votes for the same emotion category. If three votes are not obtained at the end of five annotations, we remove corresponding entry from the dataset. Figure 3.3 shows basic steps in validation of an entry. In the validation process, there are three different possible conditions, which are consensus, majority-of-votes, and reject. In this process, if consensus is reached with first three annotators, system makes a decision. Majority-of-votes may have three different vote distributions (i.e., 3-1-1, 3-1 and 3-2 votes) which is also reaching three votes for the same emotion category. On the other side, reject condition has also three different vote distributions (i.e., 2-2-1 or 1-1-1-1-1 and 2-1-1-1 votes). Additionally, it is also possible to reject entries with consensus and majority-of-votes conditions since an annotator can make a choice for ambiguous condition. Table 3.6 shows the number of entries in all these conditions after the validation process. Table 3.7 shows the examples of entries translated into

English with their original and validated emotion categories, conditions and vote distributions in the validation process.

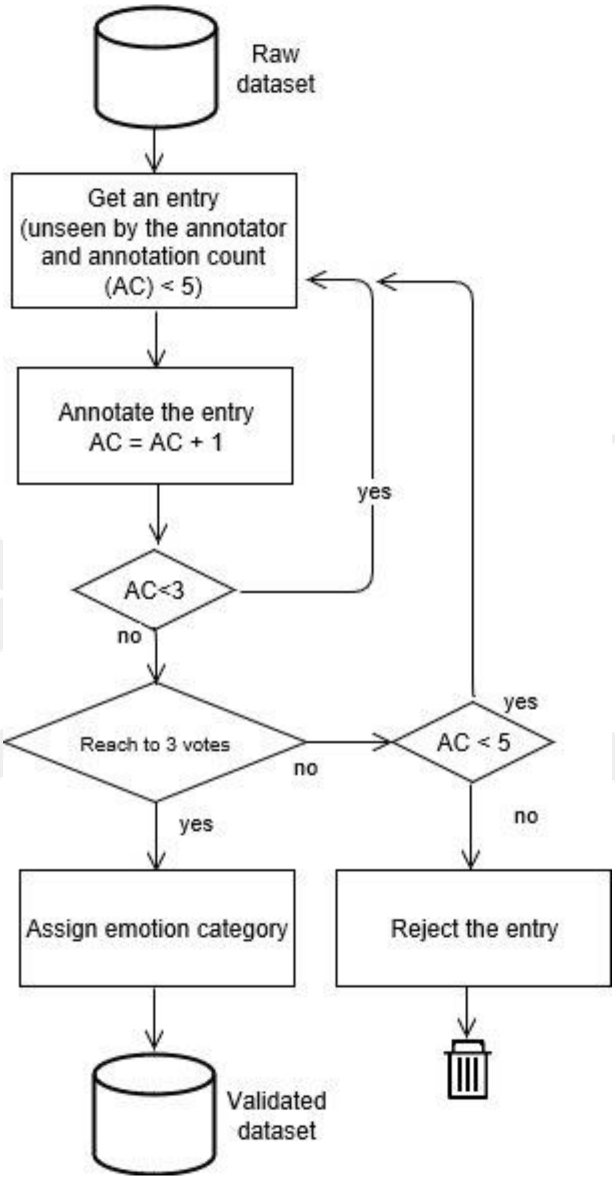


Figure 3.3 Basic steps in the validation of an entry (initial value of AC is 0)

Table 3.6 Distribution of the validation conditions of the entries at the end of the validation

Conditions	Votes	# of validated entries	# of entries validated with their original emotion	# of entries validated with different emotion
Consensus	3-0	19,462	18,154	1,308
Majority-of-votes	3-1	4,583	3,639	944
	3-1-1, 3-2	1,944	1,190	754
Reject	-	1,361	0	1,361

Table 3.7 Examples of entries translated into English with their original and validated emotion categories, conditions and vote distributions in the validation process

ID	Entry	Original Emotion	Validated Emotion	Condition	Vote Distribution
6	I am surprised to encounter a surprise that I never expected	Surprise	Surprise	Consensus (3-0)	3 Surprise
2048	My colleague's attitude is bothering me	Disgust	Anger	Consensus (3-0)	3 Anger
3254	Little gestures from someone you do not know and behaviors that he or she think and care about you	Happiness	Happiness	Majority-of-votes (3-1)	3 Happiness 1 Surprise
116	The moment I notice that I have uploaded the wrong assignment.	Fear	Fear	Majority-of-votes (3-1-1)	3 Fear 1 Sadness 1 Anger
3741	In general, I sleep by opening the television and cutting down the volume of it.	Fear	Ambiguous	Reject	2 Happiness 2 Fear 1 Ambiguous
19301	When I see a goalkeeper scored to his own goal.	Surprise	Ambiguous	Reject	1 Happiness 1 Sadness 1 Surprise 1 Ambiguous 1 Anger

In validation process, 48 volunteered annotators worked for 92,986 individual annotations. Annotators have 11 different professions where engineers, students, and academicians are among the prominent as shown in Figure 3.4. Their age distribution is between 14 and 67 years old, where only two of them are less than 20 years old. Figure 3.5 shows the number of annotators for each age group. The maximum number of individual annotations, which an annotator performed, is 10,763, and the minimum is only eight.

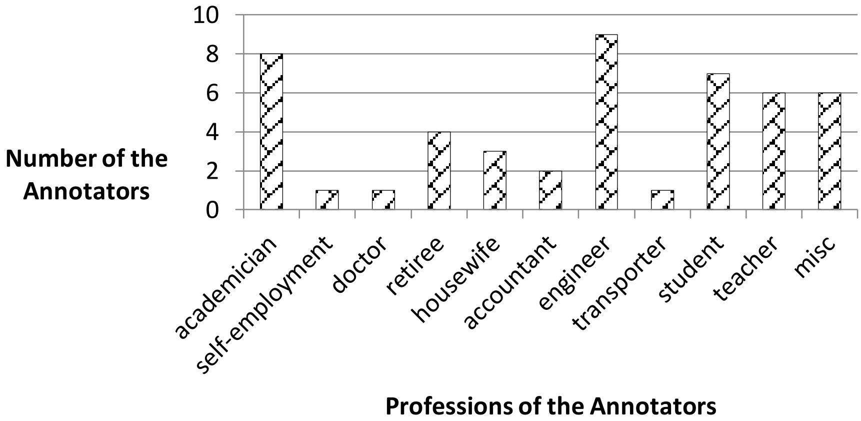
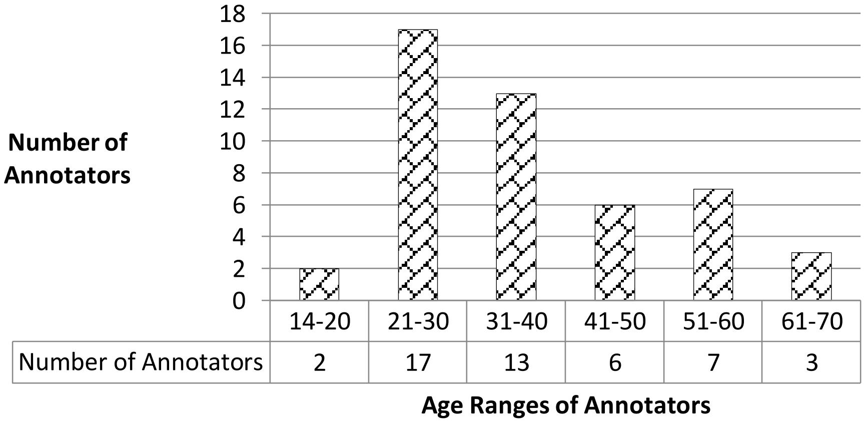


Figure 3.4 Distribution of professions of the annotators



	14-20	21-30	31-40	41-50	51-60	61-70
Number of Annotators	2	17	13	6	7	3

Figure 3.5 Number of annotators for each age group

Table 3.8 shows the total numbers of individual annotations versus age groups of annotators in the validation process. Two of these age groups, 21-30 and 31-40, are the top two dominant groups since they have the maximum number of annotators. Figure 3.6 represents the distribution of annotators' contribution to validation process, which illustrates the proportion of the total annotations that is cumulatively annotated by percentage of the annotators. For example, the top-25% of annotators has 79% of the whole individual annotations in the validation process. Furthermore, we evaluated level of agreement between annotators by Cohen's kappa (Cohen, 1960) value which is found 0.83 indicating very good level of agreement between annotators.

Table 3.8 Distribution of the individual annotations according to annotators' age ranges

Age ranges	# of annotations
14-20	790
21-30	37,103
31-40	20,446
41-50	1,317
51-60	19,119
61-70	14,211

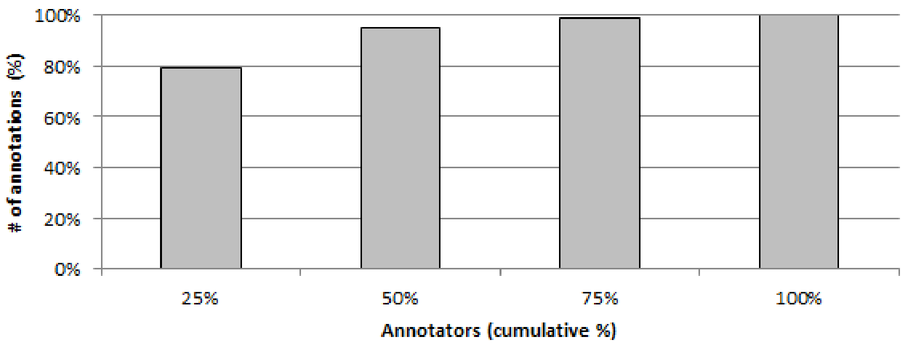


Figure 3.6 Cumulative percentages of annotations versus cumulative share of annotators

The validation process discards entries containing ambiguity in their emotion categories. As a result, we removed 1,361 entries in total, comprising 4.98% of the overall raw dataset, and called this new version as validated dataset. Table 3.9 represents the total number of entries both original and the resulted number after the validation process, where a clear difference can be easily observed. Some of the entries were annotated contradictorily to participant’s original emotional category. This indicates that the raw dataset includes some fake entries, or some of the emotions such as surprise and happiness are easy to confuse. In addition, a decrease in the number of surprise entries is clearly observable while happiness is increased.

Table 3.10 represents a table of confusion in raw versus validated datasets based on emotional categories. It shows how original emotions are interpreted and annotated differently in the validation process. For example, 642 entries originally categorized as surprise in the raw dataset are annotated as happy.

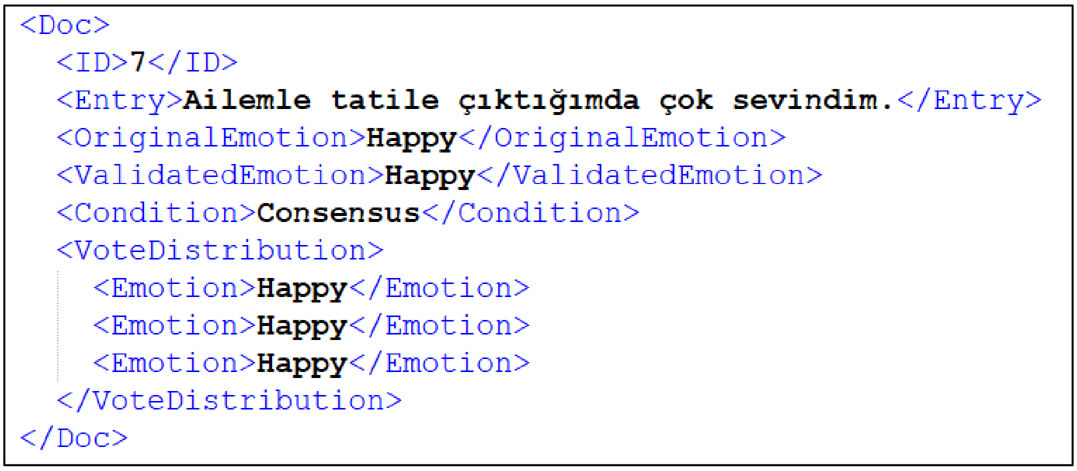
Table 3.9 Distribution of the entries after the validation process

Emotion category	**Original # of entries**	**# of entries after the validation process**
Happiness	4,700	5,229
Fear	4,616	4,393
Anger	4,636	4,723
Sadness	4,664	5,021
Disgust	4,522	3,620
Surprise	4,212	3,003
Total	27,350	25,989

Table 3.10 Table of confusion in raw versus validated datasets based on emotion categories

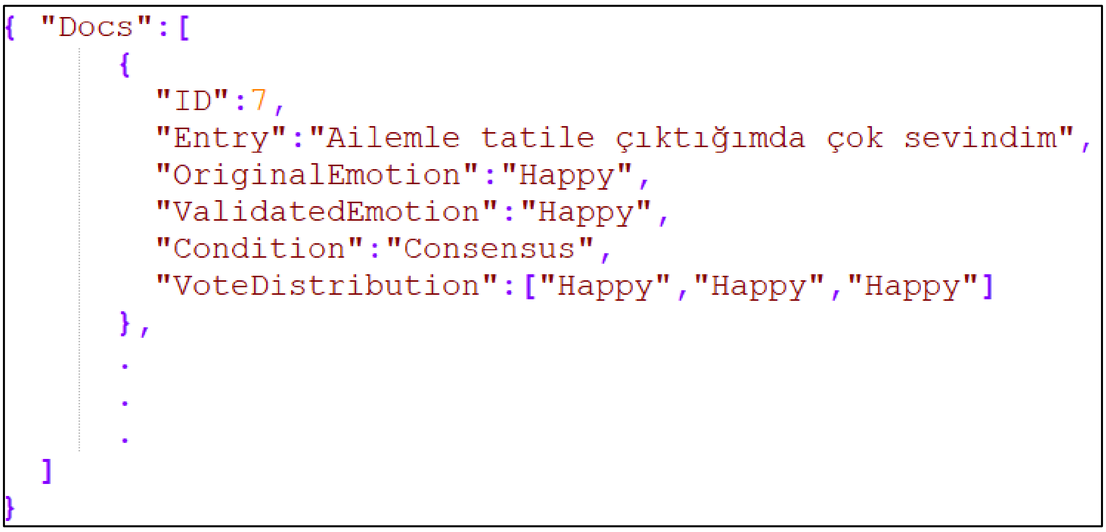
		Validated Dataset							
		Happy	Fear	Anger	Sadness	Disgust	Surprise	Reject	Total
Raw Dataset	Happy	4,513	15	2	14	1	59	96	4,700
	Fear	19	4,049	51	246	21	26	204	4,616
	Anger	19	35	3,934	357	24	35	232	4,636
	Sadness	20	95	186	4,101	11	33	218	4,664
	Disgust	16	151	421	48	3,552	16	318	4,522
	Surprise	642	48	129	255	11	2,834	293	4,212
	Total	5,229	4,393	4,723	5,021	3,620	3,003	1,361	27,350

TREMO dataset is publicly available for the use of academic researchers (Tocoglu & Alpkocak, 2018). TREMO dataset is packed into XML and JSON files where each entry is presented as it is shown in Figure 3.7 and Figure 3.8 respectively.



```
<Doc>
  <ID>7</ID>
  <Entry>Ailemle tatile çıktığımda çok sevindim.</Entry>
  <OriginalEmotion>Happy</OriginalEmotion>
  <ValidatedEmotion>Happy</ValidatedEmotion>
  <Condition>Consensus</Condition>
  <VoteDistribution>
    <Emotion>Happy</Emotion>
    <Emotion>Happy</Emotion>
    <Emotion>Happy</Emotion>
  </VoteDistribution>
</Doc>
```

Figure 3.7 XML format of an entry



```
{  "Docs":[
        {
          "ID":7,
          "Entry":"Ailemle tatile çıktığımda çok sevindim",
          "OriginalEmotion":"Happy",
          "ValidatedEmotion":"Happy",
          "Condition":"Consensus",
          "VoteDistribution":["Happy","Happy","Happy"]
        },
        .
        .
        .
   ]
}
```

Figure 3.8 JSON format of an entry

CHAPTER FOUR
LEXICON-BASED EMOTION ANALYSIS IN TURKISH

In this chapter of the thesis, we propose a Turkish emotion lexicon that can be used in emotion analysis in Turkish for six emotion categories. To the best of our knowledge, it is the first Turkish lexicon in the literature, which is generated from an original Turkish dataset, TREMO (Tocoglu & Alpkocak, 2018). To create the lexicon, we plan to examine the effects of stemming, term-weighting, lexicon enrichment methods and term selection approaches for lexicon-based emotion analysis, respectively. To evaluate the performance of the lexicon, we use the keyword-spotting technique on a different Turkish dataset. Figure 4.1 shows the stages of the creation of the lexicon.

4.1 Materials and Methods

In this section, we described the materials and methods required to create and examine the lexicon. We decided to use the TREMO dataset (Tocoglu & Alpkocak, 2018) as the material, which is used for the generation of the lexicon. Before applying any methods on TREMO, we pre-processed the dataset to remove unnecessary structures and to find the stem of each word in the dataset. After the completion of pre-processing, we weighted each stem using term-class frequencies and Mutual Information (MI) (Manning et al., 2009) values. Next, we generated the four different lexicons by analyzing each weighted stem for bi-gram, concept hierarchy and the combination of these two approaches. After the creation of the lexicons, term selection phase is applied to decrease the dimension of the corresponding lexicons for effectiveness and efficiency issues.

4.1.1 Pre-Processing

The purpose of the pre-processing is to make the TREMO dataset ready for further operations used to create a new lexicon. In the first step, we removed punctuation marks, alpha-numeric characters, and extra spaces. Next, we performed two different stemming approaches named Zemberek (Akın & Akın, 2007) and TurkLemma (Civriz, 2011) on TREMO dataset and constructed two separate datasets, DS_Z and DS _T, added suffixes of letters Z and T, which are the initials of the used stemming approaches. Then, we deleted unnecessary words from the relevant datasets. The statistical data about DS_Z and DS_T are shared in Table 4.1.

Table 4.1 Size characteristics of DS_Z and DS_T

Datasets	**# of documents**	**# of terms**	**# of unique terms**
DS_T	25,989	121,539	6,289
DS_Z	25,989	123,581	4,009

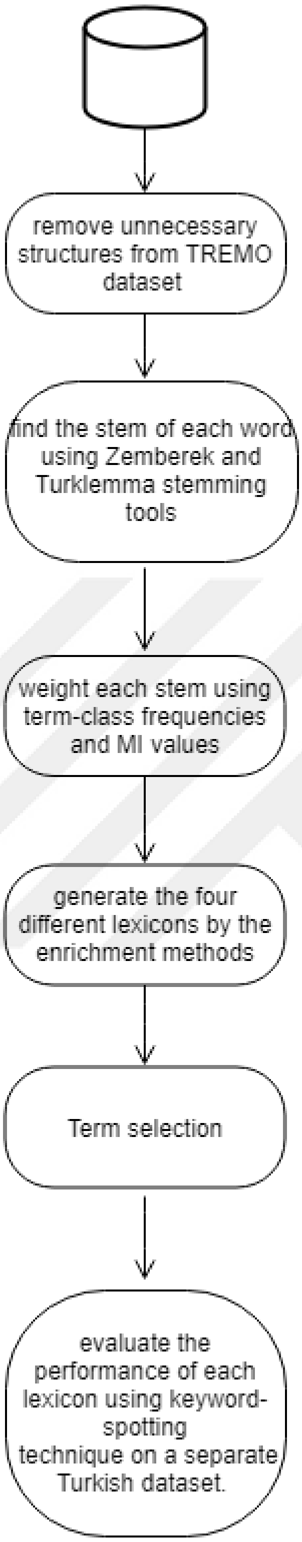

remove unnecessary structures from TREMO dataset
find the stem of each word using Zemberek and Turklemma stemming tools
weight each stem using term-class frequencies and MI values
generate the four different lexicons by the enrichment methods
Term selection
evaluate the performance of each lexicon using keyword-spotting technique on a separate Turkish dataset.

Figure 4.1 Stages of the creation of lexicon

4.1.2 Method

After the pre-processing, the next step is to generate the four lexicons. Two types of each lexicon are generated based on the datasets, DS_T and DS_Z. Firstly; we constructed the TREMO_LEX$_{\text{Basic}}$ which contains all the unique terms within the corresponding dataset. Then, we used term-class frequencies and MI values for weighting each stem in the TREMO_LEX$_{\text{Basic}}$. Table 4.2 shows the first 10 terms with the highest MI values of TREMO_LEX$_{\text{Basic}}$ for all emotion categories based on two stemming approaches, Zemberek and TurkLemma.

Table 4.2 First 10 terms with the highest MI values of TREMO_LEX$_{\text{Basic}}$ for all emotion categories based on two stemming approaches, Zemberek (Z) and TurkLemma (TL)

Happiness		**Fear**		**Anger**		**Sadness**		**Disgust**		**Surprise**	
Z	TL	Z	TL	Z	TL	Z	TL	Z	TL	Z	TL
mutlu	mutlu	kork	kork	öfkelen	öfkelen	üzül	üzül	tiksin	tiksin	şaşır	şaşır
ol	oluru	karanlık	karanlık	hak	sinirlen	vefat	vefat	tükür	koku	bekle	şaşırma
sevin	oldu	korku	korku	sinirlen	haksızlık	üz	üzüntü	koku	kusmuk	sürpriz	bekleme
kazan	sevindi	öd	korkut	yalan	yalan	öl	düşük	kusmuk	tiksindir	şaş	şaşırtı
al	kazan	gece	öd	sinir	sinir	düşük	üzer	ter	kus	görün	sürpriz
yüksek	mutluluk	film	film	yap	insan	kaybet	kay	kus	ter	şaşkın	şaş
şaşır	al	korkut	gece	söylen	haksız	üzüntü	üzüt	yemek	tükür	hayret	şaşırt
vakit	yüksek	yalnız	kovalama	konuş	izinsiz	dede	üz	koka	tükürme	ilginç	görün
birlikte	olu	kal	yalnız	söyle	al	mutlu	dedem	pis	pis	öğren	hayret
geçir	doğ	deprem	kal	insan	sınav	kork	öl	iğren	tükürülme	ummadık	ilginç

After the creation of the basic lexicon, we used it to generate 3 more lexicons, TREMO_LEX$_{\text{Bi-gram}}$, TREMO_LEX$_{\text{Concept\_hierarchy}}$ and TREMO_LEX$_{\text{Consolidated}}$, by analyzing each stem for bi-gram, constructing a concept hierarchy manually and creating the combination of these two approaches for the purpose of enrichment of lexicons. Afterwards, we applied term selection method on these lexicons.

4.1.2.1 Term Weighting

After the pre-processing of the TREMO dataset, we focused on weighting each term for each emotion category. To do this, we used term-class frequencies and MI values. We calculated the weight of each term in the lexicon by using two term weighting schemas named simple and advanced. The simple schema calculates the weight of a term by considering only the MI value. On the other hand, the second schema, calculates the weight of a term in a more detailed way in order to obtain better classification results. The formulas used in these schemas are shown in Equations (4.1) and (4.2), respectively.

We calculate the weight of the i^{th} term for c^{th} emotion category, $\mathrm{W}_{\mathrm{i}}^{\mathrm{c}}$, as follows:

$$W_i^c = MI_i^c \tag{4.1}$$

$$W_i^c = MI_i^c \times \log_2 tf_i^c \times \frac{1}{tcf} \tag{4.2}$$

The values used in these formulas are the MI value, the number of term frequency (*tf*) taken the logarithm to the base two for the corresponding emotion category, and the inverse of term class frequency (*itcf*), which indicates the number of emotion categories containing the i^{th} term.

4.1.2.2 Lexicon Enrichment Methods

In general, term selection is an important step for the sake of both text analysis accuracy and computational efficiency. However, all these trimming process reduces system performance in terms of recall and precision. Luhn defined the resolving power of words (Luhn, 1958), shown in Figure 4.2. Accordingly, the high and low frequency terms are not seen as good discriminators and the resolving power or the discrimination capability, is seen to peak at the medium frequency words (Ozkarahan, 1986). In order to do this, we included bi-grams for term phrases to decrease frequencies of high frequency terms, and construct a concept hierarchy to increase frequencies of low frequency terms.

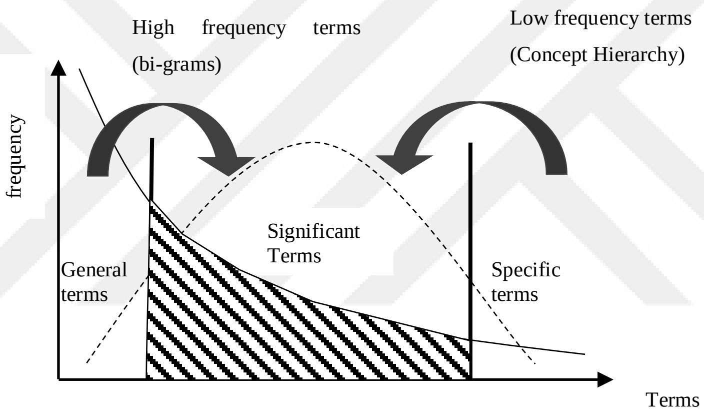


Figure 4.2 Term-frequency diagram (Luhn, 1958)

4.1.2.2.1 Bi-gram. The general idea in the bag of word model (BOW) is to consider the text as a collection of words with no regarding the sequence of words within the text. In other words, the selection of terms is not in order. This condition causes the loss of information in the text documents. To handle this problem, N-gram model can be used. It is a useful model which enables the selection of meaningful words in a sequence of n length (Fürnkranz, 1998). Using N-gram model provides opportunity to capture more contexts. Also it can be defined as effective feature selection method for word sense disambiguation (Pedersen, 2001). Each number of *n* has a name such as uni-gram stands for single word; bi-gram stands for two word phrases and so on. For example, positive oriented bi-gram examples can be found in a sentence as follows: “the best”, “I love”, “the great” and negative oriented bi-gram examples can be as “not worth”, “back to” and “returned it” (Dave et al., 2003). A list of terms of these n-gram types, for the sentence “When I see goalkeeper scored goal”, are given in Table 4.3.

Table 4.3 Example of N-gram types

N-gram Types	**Terms**
Uni-gram	(When), (I), (see), (goalkeeper), (scored), (goal)
Bi-gram	(When I), (I see), (see goalkeeper), (goalkeeper scored), (scored goal)
Tri-gram	(When I see), (I see goalkeeper), (see goalkeeper scored), (goalkeeper scored goal)

There are papers saying that n-gram model decreases the classification result, such as (Lewis, 1992), but in opposite, there are also papers such as (Fürnkranz, 1998) and (Mladenic & Grobelnik, 1998) proving that bi-gram and tri-gram models help to increase classification results.

Within the scope of this thesis, we decided to use Bi-gram model as one of the enrichment methods. The reason for using Bi-gram method is simply to decreases the high frequency individual words and increases the chance of selection of these individual words in bi-gram form. First, we concatenated each word in all documents with the one following term and then the term frequency value of each newly concatenated bi-gram term is calculated.

Table 4.4 shows the overall frequencies of two datasets after including bi-grams. The total number of bi-gram terms in each dataset is high because we added all possible bi-gram terms to the list irrespective of whether they are meaningful or meaningless. Therefore, we decided to include the first 1,000 most repeating bi-gram terms into the lexicon TREMO_LEX$_{\text{Basic}}$ and created a new one named TREMO_LEX$_{\text{Bi-gram}}$. Then, we calculated the term weights of these newly added bi-gram terms by using simple and advanced weighting schemas for six emotion categories. Table 4.5 presents the first 10 terms with the highest MI values of TREMO_LEX$_{\text{Bi-gram}}$ for all emotion categories based on two stemming approaches, Zemberek (Z) and TurkLemma (TL).

Table 4.4 Numerical information of terms for TREMO_LEX$_{\text{Bi-gram}}$ based on datasets

Datasets	**Total Terms**	**Total Bi-gram Terms**	**Total Unique Single Terms**	**Included Bi-gram terms**	**Total Term Number**
DS_Z	43,867	39,858	4,009	1,000	5,009
DS_T	54,443	48,154	6,289	1,000	7,289

Table 4.5 First 10 terms with the highest MI values of TREMO_LEX$_{\text{Bi-gram}}$ for all emotion categories based on two stemming approaches, Zemberek (Z) and TurkLemma (TL)

Happiness		**Fear**		**Anger**		**Sadness**		**Disgust**		**Surprise**	
Z	**TL**	**Z**	**TL**	**Z**	**TL**	**Z**	**TL**	**Z**	**TL**	**Z**	**TL**
mutlu	mutlu	kork	kork	öfkelen	öfkelen	üzül	üzül	tiksin	tiksin	şaşır	şaşır
mutlu ol	mutlu oluru	karanlık	karanlık	hak	sinirlen	vefat	vefat	yer tükür	koku	bekle	şaşırma
ol	mutlu olu	korku film	korku	zaman öfkelen	haksızlık	üz	öl üzül	tükür	kusmuk	görün şaşır	görün şaşır
zaman mutlu	mutlu oldu	kal kork	korkut	sinirlen	yalan	vefat et	üzüntü	koku	tiksindir	al şaşır	bekleme
mutlu et	mutlu olmuş	korku	korku film	yalan	insan öfkelen	dede vefat	düşük not	kusmuk	kus	zaman şaşır	alınca şaşır
sevin	al mutlu	kaybet kork	kal kork	yap öfkelen	sinir	düşük not	dedem vefat	insan tiksin	insan tiksin	sürpriz	al şaşır
al mutlu	oldu mutlu	zaman kork	kaybetmek kork	et öfkelen	haksızlık uğrat	öl	düşük	koku tiksin	koku tiksin	şaş	şaşırtı
kazan	mutlu edi	ol kork	öd	yalan söylen	uğrat öfkelen	zaman üzül	et üzül	ter	gör tiksin	alın şaşır	sürpriz
ol mutlu	oluru	öd	karanlık kork	ol öfkelen	sinir olu	düşük	vefat et	kus	ter	ol şaşır	bekleme olay
yüksek not	oldu	karanlık kork	film	insan öfkelen	insan	kaybet	üzer	mide bulan	yer tükür	öğren şaşır	şaş

4.1.2.2.2 Concept Hierarchy. The main purpose of constructing a concept hierarchy is to push low term-frequency valued terms into selected term set which is described in Figure 4.2. To do this, first, we fixed the terms consisting of one, two, and three words, and their representative terms that will take place of them. Tables 4.6, 4.7 and 4.8 show the terms used for the creation of concept hierarchy for all the emotion categories. After the replacement process of the terms, we recalculated the simple and advanced weights of each term in the dataset. Thus, we created a new lexicon called TREMO_LEX$_{\text{Concept\_Hierarcy}}$. Table 4.9 presents the first 10 terms with the highest MI values of TREMO_LEX$_{\text{Concept\_Hierarchy}}$ for all emotion categories based on two stemming approaches, Zemberek (Z) and TurkLemma (TL).

Table 4.6 Terms used for the creation of concept hierarchy for the emotion categories fear and sadness

Terms	**Concept**	**Terms**	**Concept**
çekinmek (hesitate) korkunç (terrible) ürkütmek (scare) ürpermek (tremble) irkilmek (recoil) ürkmek (boggle) uçurum kenarı (edge of cliff) paniğe kapılmak (panic) diz bağlarının çözülmesi (dissolving knee joints) ödü kopmak (be terrified) tedirgin olmak (worry) kaskatı kesilmek (stiffen) kabus görmek (have a nightmare) gece vakti (night time) ödü patlamak (frightened to death) ölümü çağrıştırmak (conjure death) yüreği ağzına gelmek (have one's heart in one's mouth) gözleri yuvalarından fırlamak (somebody's eyes are out on stalks)	korkmak (fear)	kahretmek (confound) kahrolmak (be grieved) burukluk (sourness) keder (sorrow) hüzün (sadness) içerlemek (resent) içi acımak (hurt so bad in one's heart) canı yakmak (get hurt) acı olay (tragic event) keyfi kaçırmak (upset) morali bozulmak (be demoralized) üzüntü duymak (feel sorry) kalbi parçalanmak (shatter the heart of someone) acı vermek (grieve) rahmetli olmak (pass away) acı söz (harsh words) vefat etmek (pass away)	üzülmek (sadness)

Table 4.7 Terms used for the creation of concept hierarchy for the emotion categories anger and disgust

Terms	**Concept**	**Terms**	**Concept**
usanmak (be weary of) bıkmak (be sickened with) ifrit olmak (fly into a fury) sinir olmak (get angry) uyuz olmak (become irritated) çileden çıkmak (lose one’s temper) deliye dönmek (get mad) sinir etmek (get one’s nerves) sinirden çıldırmak (have a fit) ayar olmak (be pissed of) iftira atmak (defame) kanın beyne sıçraması (get one's blood up) ciddi manada köpürmek (rage) gıcık (killjoy)	öfkelenmek (anger)	kusmuk (vomit) iğrenmek (be disgusted) kurtlanmak (get wormy) ıslak ekmek (wet bread) ter kokmak (stink of sweat) ter kokusu (smell of sweat) sokağa tükürmek(spit in the street) yere tükürmek(spit on the floor) içi kalkmak(feel queasy)	tiksinmek (disgust)

Table 4.8 Terms used for the creation of concept hierarchy for the emotion categories happiness and surprise

Terms	**Concept**	**Terms**	**Concept**
hayaline ulaşmak (achieve a dream) gurur duymak(be proud) ağzı kulaklarına varmak (to grin from ear to ear) küçük jest (a little gesture) bayılmak (be fond of) hayata dönmek (revive) havalara uçmak (leap for joy) içi içine sığmamak (be like a kid in a candy store)	mutluluk (happiness)	afallamak (be bewildered) şaşkınlık yaşamak (astonishment) dili tutulmak(dumbstruck) şaşa kalmak (astonishment) hayret etmek(be astonished) nutku tutulmak (dumbstruck) baka kalmak(to stand in astonishment) hayretler içinde kalmak(lost in amazement) ağzı açık kalmak(gape with astonishment)	şaşırmak (surprise)

Table 4.9 First 10 terms with the highest MI values of TREMO_LEX$_{\text{Concept\_Hierarcy}}$ for all emotion categories based on two stemming approaches, Zemberek (Z) and TurkLemma (TL)

Happiness		**Fear**		**Anger**		**Sadness**		**Disgust**		**Surprise**	
Z	TL	Z	TL	Z	TL	Z	TL	Z	TL	Z	TL
mutlu	mutlu	kork	kork	öfkelen	öfkelen	üzül	üzül	tiksin	tiksin	şaşır	şaşır
ol	oluru	karanlık	karanlık	hak	sinirlen	üz	vefat	koku	tiksindir	bekle	şaşırma
sevin	oldu	korku	korku	sinirlen	haksızlık	vefat	üzüntü	kus	koku	sürpriz	bekleme
kazan	sevindi	film	korkut	yalan	yalan	öl	düşük	yemek	kus	şaş	şaşırtı
al	kazan	gece	film	yap	insan	düşük	üzer	pis	pis	görün	sürpriz
yüksek	mutluluk	korkut	gece	söylen	haksız	kaybet	kay	ağız	ağız	şaşkın	şaşırt
şaşır	al	yalnız	kovalama	konuş	izinsiz	dede	üzüt	tükür	böcek	ilginç	şaş
vakit	yüksek	kal	kal	söyle	al	mutlu	üz	böcek	kusma	hayret	görün
birlikte	olu	deprem	yalnız	insan	sınav	üzüntü	dedem	şapır	tuvalet	öğren	ilginç
geçir	vakit	baş	yükseklik	ol	söz	kork	öl	koka	bulandır	mutlu	hayret

4.1.2.2.3 Consolidation. TREMO_LEX$_{\text{Consolidated}}$ is the last lexicon that is created by combining the two lexicons, TREMO_LEX$_{\text{Bi-gram}}$ and TREMO_LEX$_{\text{Concept Hierarcy}}$, together. Table 4.10 shows the first 10 terms with the highest MI values of TREMO_LEX$_{\text{Consolidated}}$ for all emotion categories based on two stemming approaches, Zemberek and TurkLemma.

Table 4.10 First 10 terms with the highest MI values of TREMO_LEX$_{\text{Consolidated}}$ for all emotion categories based on two stemming approaches, Zemberek (Z) and TurkLemma (TL)

Happiness		**Fear**		**Anger**		**Sadness**		**Disgust**		**Surprise**	
Z	TL	Z	TL	Z	TL	Z	TL	Z	TL	Z	TL
mutlu	mutlu	kork	kork	öfkelen	öfkelen	üzül	üzül	tiksin	tiksin	şaşır	şaşır
mutlu ol	mutlu oluru	karanlık	karanlık	hak	sinirlen	vefat	vefat	tükür	koku	bekle	şaşırma
ol	mutlu olu	korku film	korku	zaman öfkelen	haksızlık	üz	öl üzül	koku	tiksindir	görün şaşır	görün şaşır
zaman mutlu	mutlu oldu	korku	korkut	sinirlen	yalan	dede vefat	üzüntü	insan tiksin	kus	al şaşır	bekleme
mutlu et	mutlu olmuş	kal kork	korku film	yalan	insan öfkelen	düşük not	düşük not	koku tiksin	insan tiksin	zaman şaşır	alınca şaşır
sevin	al mutlu	kaybet kork	kal kork	yap öfkelen	sinir	öl	dedem vefat	ter	koku tiksin	sürpriz	al şaşır
al mutlu	oluru	zaman kork	kaybetmek kork	et öfkelen	haksızlık uğrat	zaman üzül	düşük	kus	gör tiksin	şaş	şaşırtı
kazan	oldu mutlu	ol kork	öd	yalan söylen	uğrat öfkelen	düşük	et üzül	mide bulan	ter	alın şaşır	sürpriz
ol mutlu	mutlu edi	öd	karanlık kork	ol öfkelen	insan	kaybet	üzer	yemek	tükür	ol şaşır	bekleme olay
yüksek not	oldu	karanlık kork	film	insan öfkelen	el şaka	üzüntü	kay	gör tiksin	böcek tiksin	öğren şaşır	şaş

As a result of applying bi-gram and concept hierarchy enrichment methods on the TREMO_LEX$_{\text{Basic}}$, we achieved to obtain three more new lexicons. Table 4.11 shows the term numbers of these lexicons based on DS_Z and DS_T datasets.

Table 4.11 Term numbers of the four lexicons based on DS_Z and DS_T datasets

Lexicon Sets	**DS_T**	**DS_Z**
TREMO_LEX$_{\text{Basic}}$	6,289	4,009
TREMO_LEX$_{\text{Bi-gram}}$	7,289	5,009
TREMO_LEX$_{\text{Concept Hierarcy}}$	6,244	3,977
TREMO_LEX$_{\text{Consolidated}}$	7,235	4,967

4.1.2.3 Keyword-Spotting Technique

In the content of this part of the thesis, emotion analysis is planned to be performed by using lexicon-based approach, which is an unsupervised learning algorithm. This approach is based on keyword-spotting technique. Keyword-spotting is used commonly in emotion analysis. The main idea behind this technique is to spot emotion category of a document by searching for pre-defined emotional words within that document. These emotional words can be categorized for emotion categories such as joy, anger, fear and so on.

We assume that the dataset, *D*, has as set of documents

$\mathrm{D} = \{d_1, d_2, ..., d_n\}$

where an arbitrary document d_{i} is represented with a set of terms,

$d_i = \{t_1, t_2, ..., t_k\}$

In the corresponding dataset, the emotion categories, *K*, has six emotion categories,

$\mathrm{K} = \{c_{1,} c_2, c_3, c_4, c_5, c_6\}$

Keyword-spotting technique is a function, which is presented as follows in Equation (4.3);

$$E(d_i) = argmax(\sum_{c=1}^{6} \sum_{d_{ij} \in L^c} L_j^c\, w) \tag{4.3}$$

where d_i indicates the document to be classified using keyword-spotting technique, dij stands for the j^{th} term of the d_i document, L^c indicates the lexicon of the c^{th} emotion category and L_j^c w indicates weight value of the j^{th} term for c^{th} emotion category in the lexicon L.

Within the scope of keyword-spotting technique, we compared the terms of each document with the terms defined in the corresponding lexicon. If there is a match between the terms, we collected the weights of the matched terms in the lexicon separately for each emotion category, and then we calculated the overall scores of each emotion category for each document. Thus, we achieved to determine the new emotion category of the corresponding document by assigning the emotion category with the highest weight value as the new emotion category. Here, if the original emotion category of the relevant document is the same as the newly identified category, no change is made, but if it is different, the emotion category of the related document is replaced by the new one. We redefined the emotion categories of all documents in this way. The ratio of documents whose original emotion category is not changed in the classification process provides the accuracy value.

4.1.2.4 Term Selection

We applied term selection to choose significant terms for inclusion to lexicon to increase the accuracy. First we focused on deciding which lexicon, stemming approach and dataset, to be used. For this purpose, we performed an evaluation with a test set proposed by (Açıcı, 2012). The test set is created by a survey which is carried out among 500 university students from different departments for gathering a dataset for seven emotion categories. They are asked to write about a moment of their lives for each emotion categories. To do this, she prepared a survey with seven fields, and participants are asked to annotate their emotional states with a few sentences. As a result of this data compilation process, they collected 3,206 documents which contain 3,238 unique defined terms in total. There are missing documents within the dataset. This is because some of the participants left blank some of the emotional categories. Table 4.12 shows the distribution of documents for each emotion categories.

Table 4.12 Distribution of documents for each emotion categories

	happy	**fear**	**angry**	**sadness**	**disgust**	**shame**	**guilty**
Document Number	493	472	476	468	463	416	418

The test set includes only four emotion categories in common with TREMO, which are happy, fear, anger and sadness. We applied two stemming approaches, Zemberek and Turklemma, and named resulting test sets as TestSet_Z and TestSet_T, respectively. Then, we performed keyword-spotting technique on both sets. In evaluation experiment, we used four different lexicons with two different weighting schemas to evaluate their performance in terms of accuracy measures. Then, we used these results to select an appropriate threshold value to cut lexicons for efficiency issues, based on weighting values for each emotion categories individually.

Figure 4.3 presents the performance results of keyword-spotting results using simple and advanced weighting schemas for four lexicons on Zemberek and TurkLemma, respectively. The results clearly show that using advanced weighting schema gives higher accuracy values than using simple weighting schema in all cases no matter which test dataset is used. To compare Zemberek and TurkLemma, we calculated average accuracy difference between simple and advance weighting schema, where scores are 0.818 and 2.045, respectively. In other words, lexicons prepared with TurkLemma generally give higher results. Lastly, The TREMO_LEX$_{\text{Consolidated}}$ is the best lexicon for both test sets compared to others.

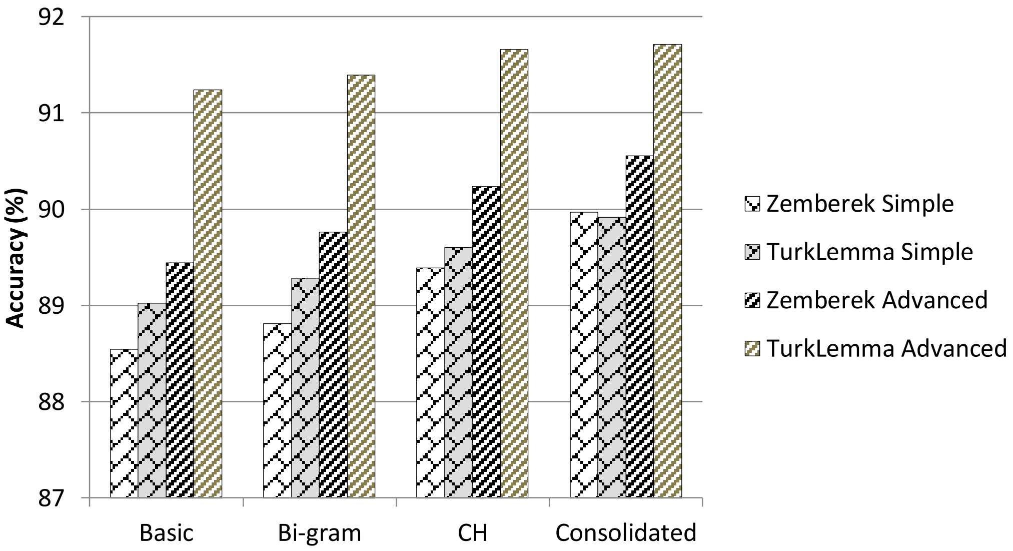


Figure 4.3 Comparison of stemming approaches in terms of accuracy for different weighing schemas over lexicons

Based on the results obtained so far, we fixed to continue our experiments with TestSet_T and TREMO_LEX$_{\text{Consolidated}}$. At this stage, we weighted the lexicon TREMO_LEX$_{\text{Consolidated}}$ by using advanced weighting schema. We used this schema for ranking the most significant terms for each emotion category. We reordered the terms of the corresponding lexicon from the highest to the lowest value and then choose the top *n* terms as interim lexicons, for each emotion category. Then, we ran keyword-spotting using these lexicons, and observed the accuracy values. Figure 4.4 shows the results of this experiment, where the best accuracy value is obtained for *n*=250. Thenceforth, we empirically selected the cut-off value as 0.00091329, which is the weight value of the 250th term of the happy emotion category. As we applied this value to other five emotion categories, the term selection value for each category varies as shown in Table 4.13.

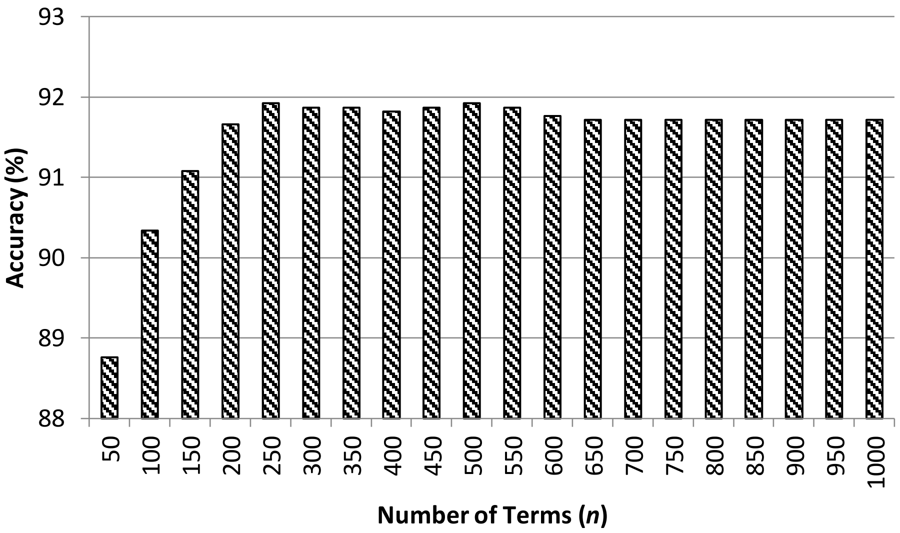
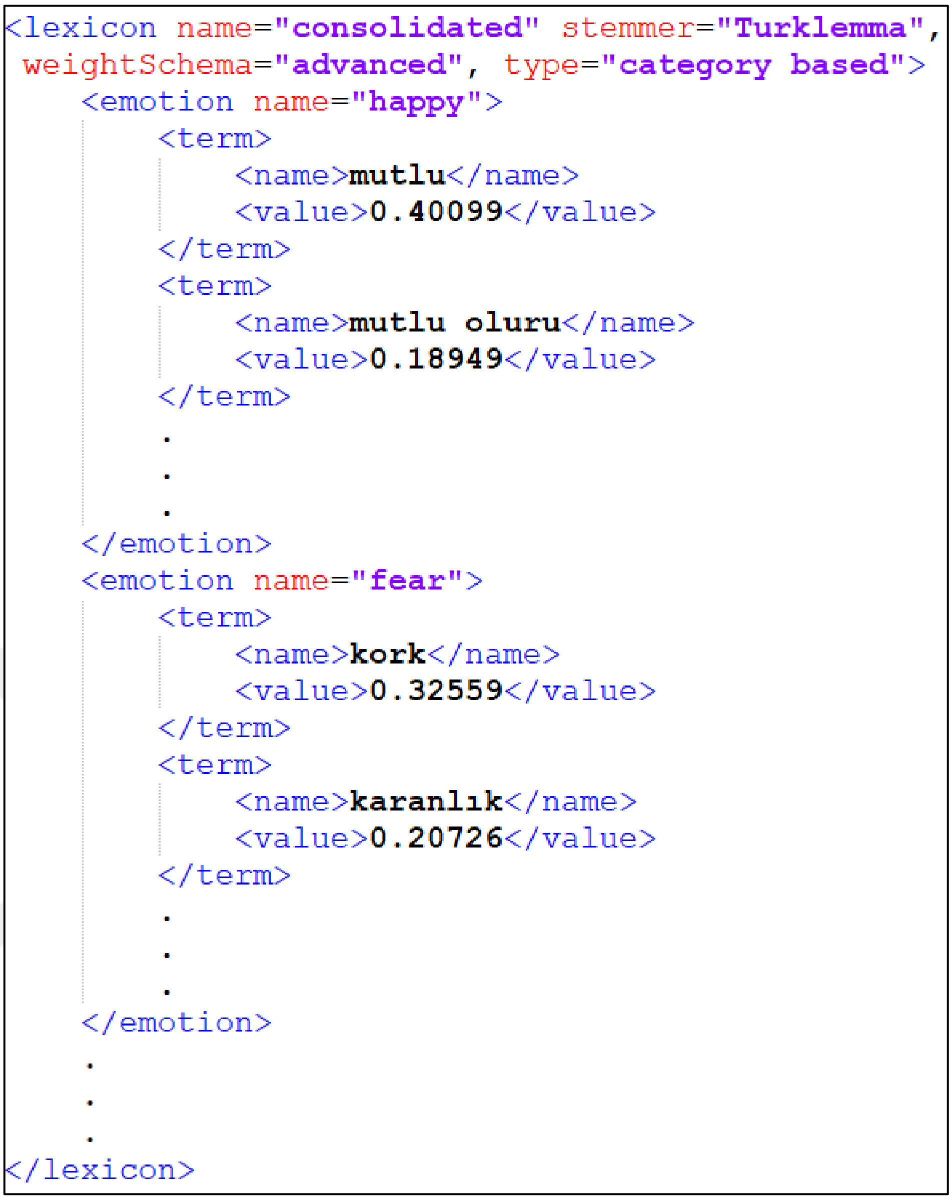


Figure 4.4 Accuracy values versus number of terms in TREMO_LEX$_{\text{Consolidated}}$

Table 4.13 Term counts of TREMO_LEX$_{\text{Consolidated}}$ after the term selection process

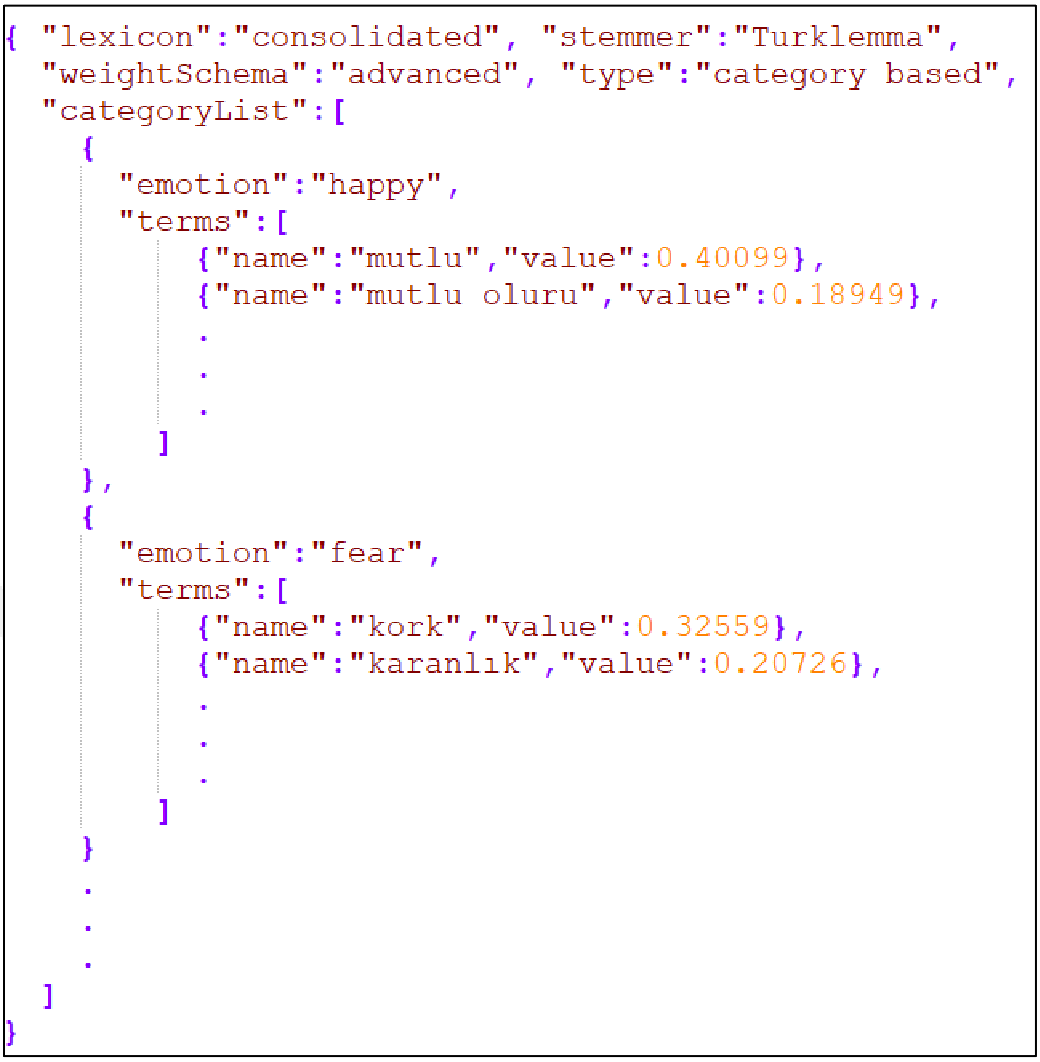
Number of terms	**Happy**	**Fear**	**Anger**	**Sadness**	**Disgust**	**Surprise**
All terms	7,235	7,235	7,235	7,235	7,235	7,235
Selected terms	250	214	261	212	228	155

TREMO_LEX lexicon is publicly available for the use of academic researchers (Tocoglu & Alpkocak, 2018). We share the advanced weighted TREMO_LEX$_{\text{Consolidated}}$ lexicon in two formats, category-based and emotion vector-based. They are both shared in XML and JSON text format in Figures 4.5, 4.6, 4.7 and 4.8. A brief example of emotion vector-based format is shown in Appendix-1.



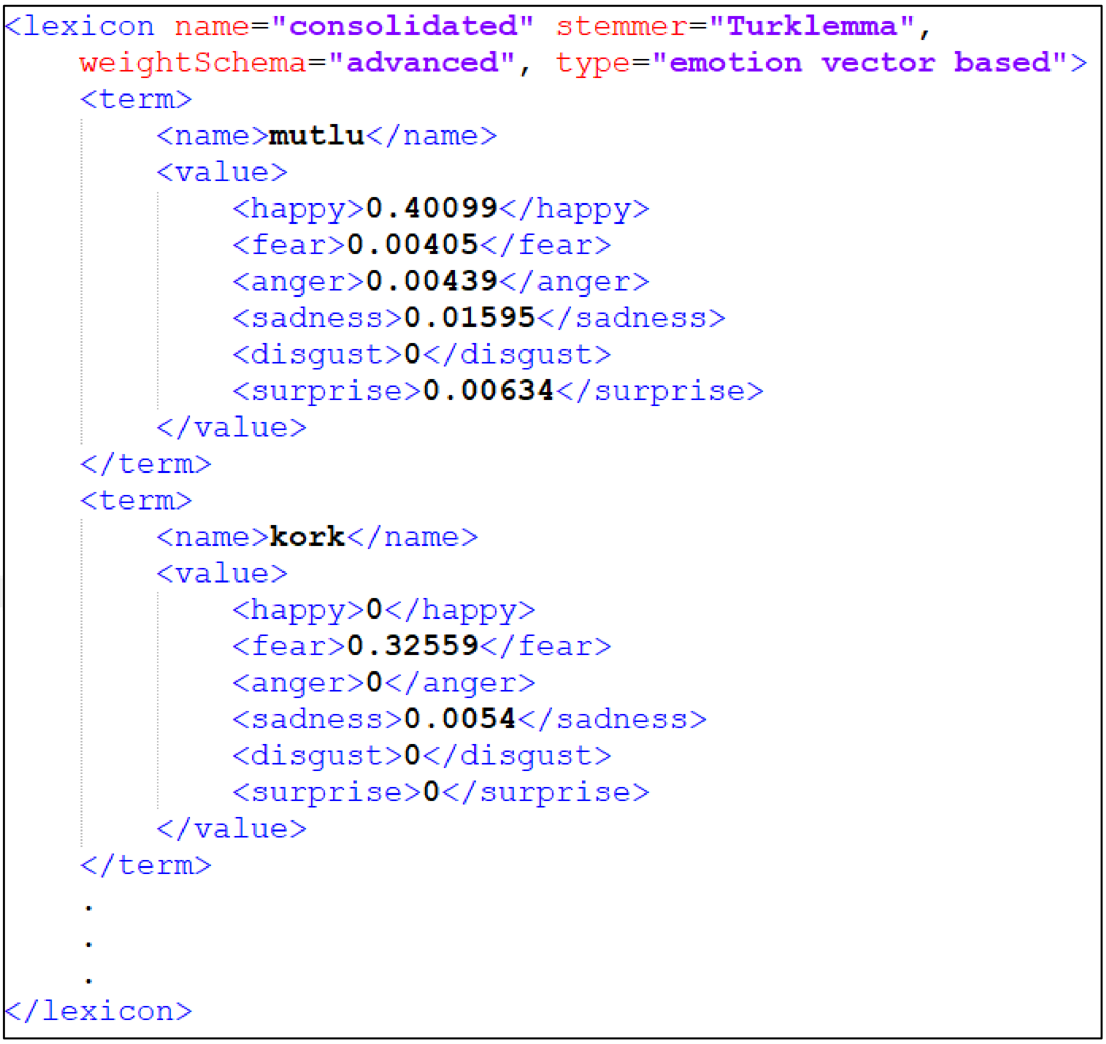
```
<lexicon name="consolidated" stemmer="Turklemma",
 weightSchema="advanced", type="category based">
    <emotion name="happy">
        <term>
            <name>mutlu</name>
            <value>0.40099</value>
        </term>
        <term>
            <name>mutlu oluru</name>
            <value>0.18949</value>
        </term>
        .
        .
        .
    </emotion>
    <emotion name="fear">
        <term>
            <name>kork</name>
            <value>0.32559</value>
        </term>
        <term>
            <name>karanlık</name>
            <value>0.20726</value>
        </term>
        .
        .
        .
    </emotion>
    .
    .
    .
</lexicon>
```

Figure 4.5 Category-based format of the lexicon in XML



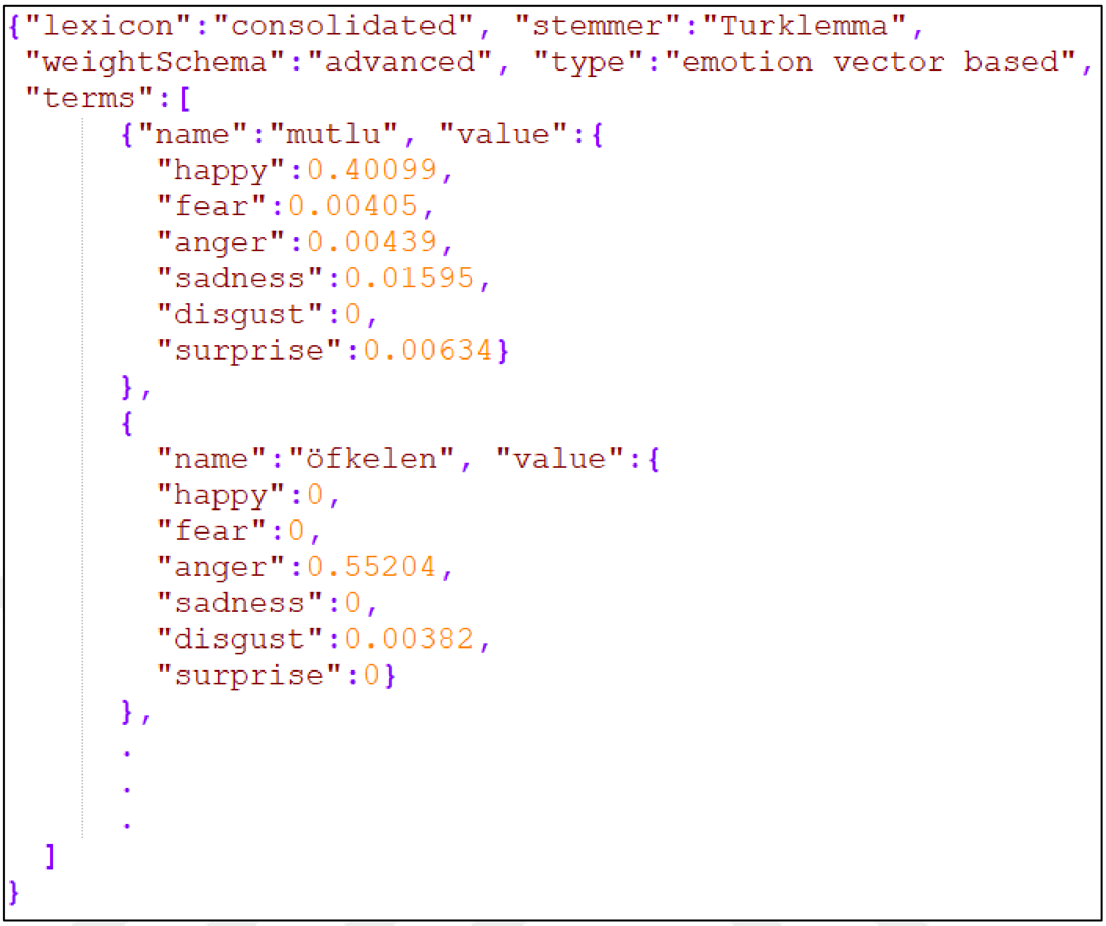
```
{ "lexicon":"consolidated", "stemmer":"Turklemma",
  "weightSchema":"advanced", "type":"category based",
  "categoryList":[
    {
      "emotion":"happy",
      "terms":[
          {"name":"mutlu","value":0.40099},
          {"name":"mutlu oluru","value":0.18949},
          .
          .
          .
        ]
    },
    {
      "emotion":"fear",
      "terms":[
          {"name":"kork","value":0.32559},
          {"name":"karanlık","value":0.20726},
          .
          .
          .
        ]
    }
    .
    .
    .
  ]
}
```

Figure 4.6 Category-based format of the lexicon in JSON



```
<lexicon name="consolidated" stemmer="Turklemma",
    weightSchema="advanced", type="emotion vector based">
    <term>
        <name>mutlu</name>
        <value>
            <happy>0.40099</happy>
            <fear>0.00405</fear>
            <anger>0.00439</anger>
            <sadness>0.01595</sadness>
            <disgust>0</disgust>
            <surprise>0.00634</surprise>
        </value>
    </term>
    <term>
        <name>kork</name>
        <value>
            <happy>0</happy>
            <fear>0.32559</fear>
            <anger>0</anger>
            <sadness>0.0054</sadness>
            <disgust>0</disgust>
            <surprise>0</surprise>
        </value>
    </term>
    .
    .
    .
</lexicon>
```

Figure 4.7 Emotion vector-based format of the lexicon in XML



```
{"lexicon":"consolidated", "stemmer":"Turklemma",
 "weightSchema":"advanced", "type":"emotion vector based",
 "terms":[
      {"name":"mutlu", "value":{
        "happy":0.40099,
        "fear":0.00405,
        "anger":0.00439,
        "sadness":0.01595,
        "disgust":0,
        "surprise":0.00634}
      },
      {
        "name":"öfkelen", "value":{
        "happy":0,
        "fear":0,
        "anger":0.55204,
        "sadness":0,
        "disgust":0.00382,
        "surprise":0}
      },
      .
      .
      .
  ]
}
```

Figure 4.8 Emotion vector-based format of the lexicon in JSON

CHAPTER FIVE
EXPERIMENTATION AND RESULTS

In this chapter, we share the experiment results of machine learning algorithms applied on TREMO dataset and lexicon-based approach results applied on newly created lexicons.

5.1 Experiments of Machine Learning Algorithms Applied on TREMO Dataset

In this section, we describe classification experiments that we performed on TREMO, including both raw and validated datasets. The difference between these two datasets is the elimination of the 1,361 ambiguous entries defined in the validation process. Before the experiments are conducted, we preprocessed datasets to remove unnecessary structures and prepared them to be used as training datasets. In the first step of the pre-processing, four dataset versions are generated from the two relevant datasets (raw and validated) using two stemming methods. Next, ineffective terms and numerical values are deleted. After the completion of pre-processing, term selection is performed to eliminate insignificant terms and minimize the dimensions of the dataset versions. Then, we transformed these versions to vector space models, where document term matrices (DTM) are generated. In the last sub-section, we present classification results we obtained using four machine learning techniques. Figure 5.1 shows the all required steps taken to complete evaluation process of machine learning experiments of the TREMO dataset.

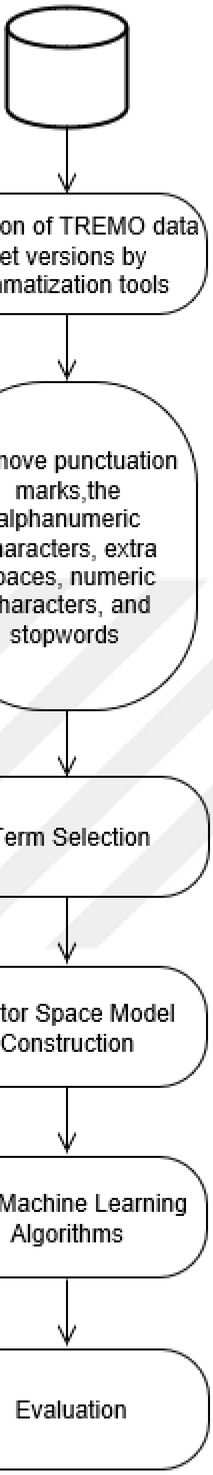

Creation of TREMO data set versions by lemmatization tools
Remove punctuation marks,the alphanumeric characters, extra spaces, numeric characters, and stopwords
Term Selection
Vector Space Model Construction
Run Machine Learning Algorithms
Evaluation

Figure 5.1 Stages of machine learning experiments

5.1.1 Pre-Processing

The goal of pre-processing is to prepare the dataset to experimentations. To do so, first we checked the spelling of all the entries manually. The second step is to find the stems of the terms in each entry, then, we deleted the punctuation marks, the extra spaces, the numeric characters, and the fluff terms from these datasets. For this purpose, we used two different stemming methods, fixed prefix stemming (FPS) (Can et al., 2008) and a directory-based Turkish stemmer named Zemberek (Z) (Akin & Akin, 2007). FPS simply gets the first n characters of a term, trims out the rest. At this point in the study, we chose n as five to represent the first-five characters (F5). We chose F5, instead of F4 or F7 since it has been shown that it has optimum performance in terms of effectiveness measures among others (Can et al., 2008). We performed these two stemming methods upon the raw and validated datasets, and created four different dataset versions. Table 5.1 shows some numeric properties of all four versions of TREMO, where V character in the name of dataset versions represents validated dataset while the others are for raw dataset.

Table 5.1 Numerical properties of the four dataset versions of TREMO

Dataset	Dataset versions	Total entry	Total terms	Unique terms
Raw	F5	27,350	132,485	6,489
	Z	27,348	129,267	4,142
Validated	F5_V	25,989	126,593	6,280
	Z_V	25,989	123,581	4,009

5.1.2 Term Selection

After the creation of four dataset versions based on two stemming methods, we used term selection to remove insignificant terms and to reduce the dimensionality of the term sets (Quinlan, 1993). It is not mandatory but it is a step for improving the classification results. In this part of this thesis, applying a term selection method to the datasets plays an important role because of the high number of terms within the datasets that prevent running some classification algorithms. For ranking the most significant terms for each emotion category, we decided to use mutual information (MI) (Manning et al., 2009). The MI is a feature selection method, which measures how much information the presence/absence of a term contributes to making the correct classification decision on *c*. We calculated the weights of each term, using MI values for six emotion categories.

U in Equation (5.1), is a random variable that takes the values e_t =1 (the document contains the word *t*) and e_t = 0 (the document do not contain the word *t*). The random variable *C* in the same formula takes the values e_c = 1 (the document is classified as category *c*) and e_c = 0 (the document is not classified as category *c*) (Manning et al., 2009).

$$I(U;C) = \sum_{e_t \in \{1,0\}} \sum_{e_c \in \{1,0\}} P(U = e_t, C = e_c) \log_2 \frac{P(U=e_t, C=e_c)}{P(U=e_t)P(C=e_c)} \qquad (5.1)$$

The N_s specified in Equation (5.2) is the sum of the document numbers. In this stage, each *N* indicates the e_t and e_c values by two subscripts. For example, N_{10} indicates the documents that contain the word *t* (e_t = 1) and are not in the category *c* (e_c = 0). $N_{1x} = N_{10} + N_{11}$ formula gives all the documents containing the word *t* (e_t =1). $N = N_{00} + N_{01} + N_{10} + N_{11}$ formula computes all the documents in the dataset (Manning et al., 2009).

$$I(U;C) = \frac{N_{11}}{N}\log_2\frac{NN_{11}}{N_{1x}N_{x1}} + \frac{N_{01}}{N}\log_2\frac{NN_{01}}{N_{0x}N_{x1}} + \frac{N_{10}}{N}\log_2\frac{NN_{10}}{N_{1x}N_{x1}} + \frac{N_{00}}{N}\log_2\frac{NN_{00}}{N_{0x}N_{x0}} \quad (5.2)$$

We reordered the terms of these four dataset versions from the highest to the lowest MI value and then, selected the most valuable terms for each emotion category based on two approaches. The first is the selection of the first 500 terms, and the other one is to make the selection for a given threshold. Here we decided on selecting the first 500 terms based on classification results obtained by using complement naive Bayes (CNB) (Rennie et al., 2003) as the classifier and F5 as the training dataset. Figure 5.2 shows these classification results where first 500 terms provide the highest accuracy value among others. On the other hand, we empirically chose a threshold value of MI value of the five hundredth term of the happiness emotion category. This is because happiness emotion category provided one of the highest classification results among other categories. Table 5.2 shows the threshold values for the four dataset versions and the number of terms fit into the given threshold value for each emotion category. To avoid the repetition of terms within each emotion category, we took intersection of the terms. The intersected term numbers according to each dataset version are shown in Table 5.3.

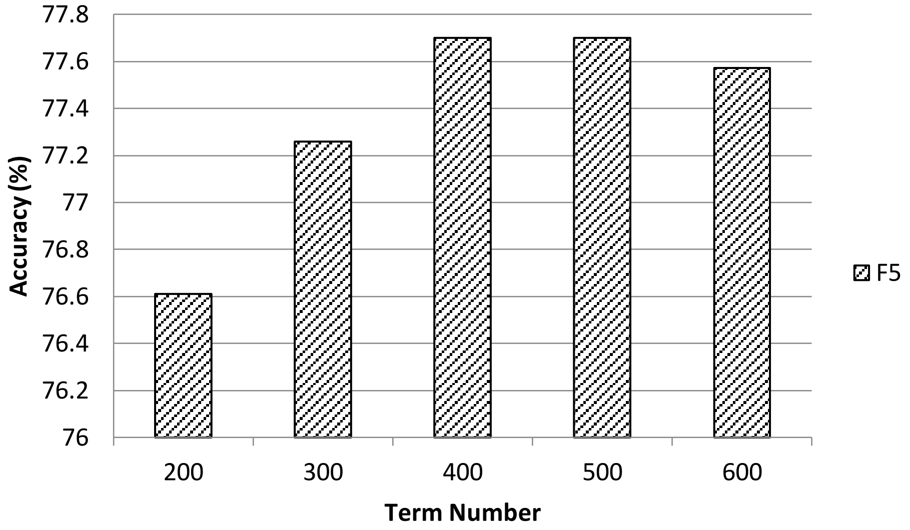


Figure 5.2 Comparison of classification results based on term numbers

Table 5.2 Number of selected terms for each emotion category and threshold values for each version

Emotion types/Threshold	**F5**	**F5_V**	**Z**	**Z_V**
Happiness	500	517	535	500
Fear	509	498	504	450
Anger	522	527	484	477
Sadness	399	409	408	397
Disgust	646	562	618	547
Surprise	371	292	409	310
Threshold value	**0.000234**	**0.0002663**	**0.000185**	**0.00021**

Table 5.3 Number of intersected terms for each term selection approaches for four dataset versions

Dataset versions	**Term selection approaches**	
	First 500 terms	**Threshold**
F5	1,439	1,395
F5_V	1,397	1,336
Z	1,336	1,386
Z_V	1,316	1,192

Table 5.4 presents numeric information of the ten highest MI-valued terms of the Z dataset version for the happiness category. The term column includes English translations in parenthesis. In addition to MI value, it also shows the overall frequency values, the ratio of the frequency value to the total number of entries and the value indicating the number of entries having the related term in their contents.

Table 5.4 Numerical information of the ten highest MI-valued terms of the Z dataset version for the happiness emotion category. English translation is given in parenthesis. N is the total number of entries

Term	**MI**	**Frequency**	**Frequency /N**	**# of entries containing the term**
mutlu (happy)	0.1965	2,095	0.0766	2,017
ol (happen)	0.0584	5,232	0.1913	4,702
kork (fear)	0.0215	2,135	0.0781	2,102
sevin (glad)	0.0177	235	0.0086	233
şaşır (surprise)	0.0173	1,753	0.0641	1,740
kazan (win)	0.0169	543	0.0199	534
üzül (sad)	0.0162	1,695	0.0620	1,641
tiksin (disgust)	0.0146	1,474	0.0539	1,464
öfkelen (anger)	0.0146	1,437	0.0525	1,427
birlikte (together)	0.0097	183	0.0067	182

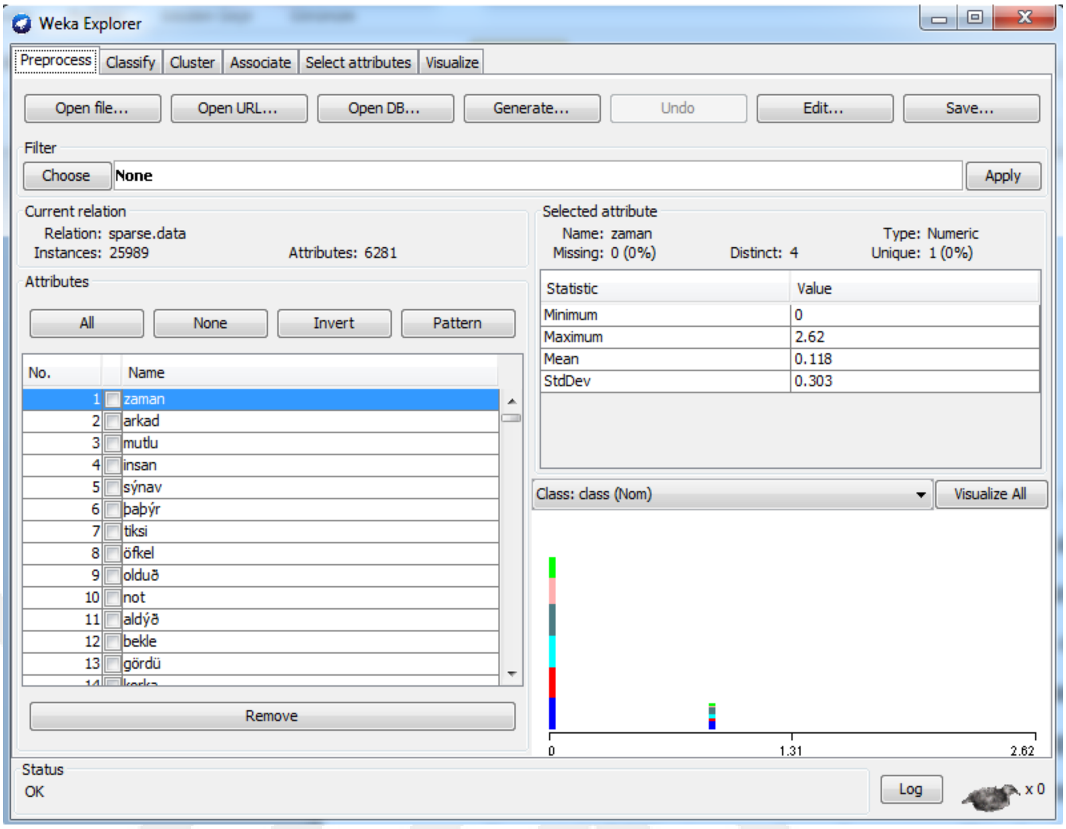
After finishing pre-processing and term selection phases, the next step is to transform these dataset versions into a vector space model (Kılınç et al., 2016). In this model, each entry is represented as a vector in DTM. In DTM, each row is a vector where columns represent terms. In other words, the columns represent the terms of the dataset. Each cell of a vector stands for a value assigned to a term. There are three well known methods for assigning value to terms which are presence-absence, term frequency (*tf*) and term frequency-inverse document frequency (*tf-idf*) methods. First one, presence-absence (Pang et al., 2002), is based on assigning 0, when the related term does not exist in the related document or 1 when the condition is vice versa. In the second method *tf*, this time the values indicate the total count of a term (*t*) within the related document (*d*). In the formula, this is denoted as $tf_{t,d}$. The third method (*tf-idf*) is more complex compared to others (Manning et al., 2009). The goal of this method is to assign higher values to terms those exist rarely. In fact, as it is displayed in Equation (5.3), it is the multiplication of two major equations $tf_{t,d}$ and inverse document frequency (idf_t). As it is displayed in Equation (5.4), idf_t is calculated by taking the logarithm of the ratio N, the number of the total documents within the dataset, divided by the document frequency value (df_t). Here, df_t stands for the number the documents containing the term t. In general, idf_t value is expected to be higher for the terms exist in small number of documents. In this part of this thesis, we used TF × IDF weighting scheme in vector space model.

$$tf\text{-}idf_{\mathrm{t,d}} = tf_{\mathrm{t,d}} \times idf_{\mathrm{t}} \tag{5.3}$$

$$idf_{\mathrm{t}} = \log(\mathrm{N}/df_{\mathrm{t}}) \tag{5.4}$$

We set up a series of experimentations to determine whether the raw dataset yields better classification results after passing through the validation process. For this reason, we subjected four dataset versions, F5, F5_V, Z, and Z_V, to evaluate the performance of different classification algorithms, which are CNB, random forest (RF) (Xu et al., 2012), decision tree C4.5 (J48) (Quinlan, 1993), and an updated version of support vector machines (SVM) (Platt, 1998). We implemented 10-fold cross validation in evaluation of the performance of each classifier, where 90% of dataset is used as a training set, and the remainder is used for testing. We compared classification results in terms of accuracy, precision, recall, and F-measure. Furthermore, we also examined the reflections of these results by emotion categories.

All the experiments are implemented in WEKA version 3.6.14 (Witten & Frank, 2005). Waikato Environment for Knowledge Analysis (WEKA) is a project presenting a platform which provides many machine learning algorithms with a well-designed graphical user interface (GUI). This GUI, displayed in Figure 5.3, is composed of six main parts which are preprocess, classification, clustering, creation of association rules, term selection and data visualization. It is developed in Java programming language and is capable of being imported as WEKA API into programs being implemented. It takes input dataset by four document types which are CSV, LibSVM's format, C4.5 or Attribute Relationship File Format (ARFF). Among these formats, arff is the WEKA's own file format. It is has got two main parts. First one is the header part which includes the relation name and all attribute names with their value types. The second part of this file format consists of the data itself.



Filter: None

Relation: sparse.data
Instances: 25989
Attributes: 6281

Name: zaman
Missing: 0 (0%)
Distinct: 4
Type: Numeric
Unique: 1 (0%)

Statistic	Value
Minimum	0
Maximum	2.62
Mean	0.118
StdDev	0.303

No.	Name
1	zaman
2	arkad
3	mutlu
4	insan
5	sýnav
6	baþýr
7	tiksi
8	öfkel
9	olduð
10	not
11	aldýð
12	bekle
13	gördü

Class: class (Nom)

Status: OK

Figure 5.3 GUI of WEKA

The evaluation process for this section is done by using several evaluation parameters such as accuracy, precision, recall and F-Measure. To start with accuracy, it shows the quality of the obtained classification results. The accuracy of a classification is computed by dividing the total number of correct cases by the number of all cases multiplied by 100, which is shown in Equation (5.5).

$$\text{Accuracy} = \text{correct\_cases} \,/\, \text{all\_cases} \times 100\ \% \qquad (5.5)$$

On the other hand, precision is calculated by dividing retrieved relevant items by retrieved all items. This computation can be clarified more clearly with the contingency table where retrieved relevant items presented as true positives and retrieved all items presented with the summation of true positives and false positives which is displayed in Equation (5.6).

$$\text{Precision} = \text{tp} / (\text{tp} + \text{fp}) \tag{5.6}$$

Another evaluation parameter is recall, which is computed by dividing relevant items retrieved by all the relevant items. This computation also has equivalent equation arranged by the contingency table which is true positives is divided by the summation of true positives and false negatives displayed in Equation (5.7).

$$\text{Recall} = \text{tp} / (\text{tp} + \text{fn}) \tag{5.7}$$

F-Measure is another evaluation parameter we used in this study. It is the harmonic mean of recall and precision. The computation of F-Measure is displayed in Equation(5.8).

$$\text{FMeasure} = 2 \times \frac{precision \times recall}{precision + recall} \tag{5.8}$$

Figure 5.4 shows the general classification results obtained for each dataset versions according to the three classification algorithms without using any term selection approach. Here, we did not include RF algorithm because of the high dimensionality of datasets. Despite minor differences, the figure shows that the two stemming techniques yield similar results. However, it appears that the best classification results are obtained when using SVM as the classification algorithm. Besides, the deletion of ambiguous entries detected in the validation process from the raw dataset has positive effects on the overall average accuracy values by 5.7% of F5 stemmed datasets and by 5.6% of Z stemmed datasets. Apart from that, the CNB and SVM techniques yield slightly better results than the J48 algorithm.

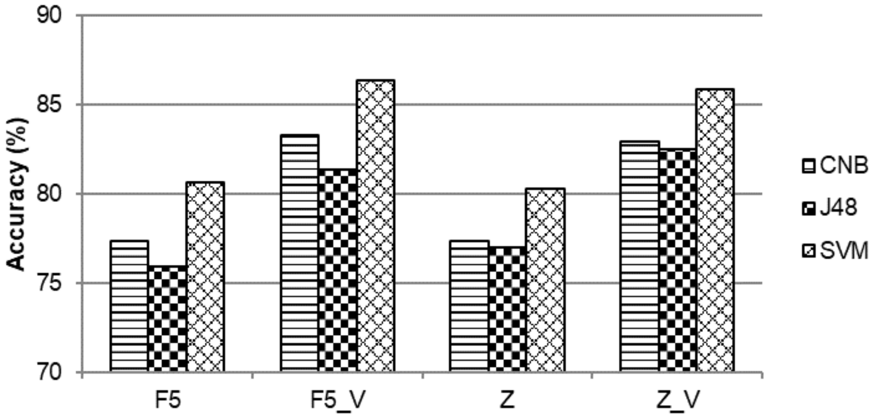


Figure 5.4 General accuracy results obtained for each dataset version according to three different classifiers

Figure 5.5 shows overall accuracy values of the raw and validated dataset versions subjected to the first 500 term selection approach. We compared the classification accuracy results, for four classification algorithms. All the comparison results are very close to each other, so there is no clear-cut winner. CNB and SVM produced higher scores for F5 and F5_V than Z and Z_V dataset versions. On the other hand, J48 and RF have better results on Z and Z_V dataset versions. These results explain that both stemming methods have similar results. These results also repeat in threshold term selection approach, which is shown in Figure 5.6.

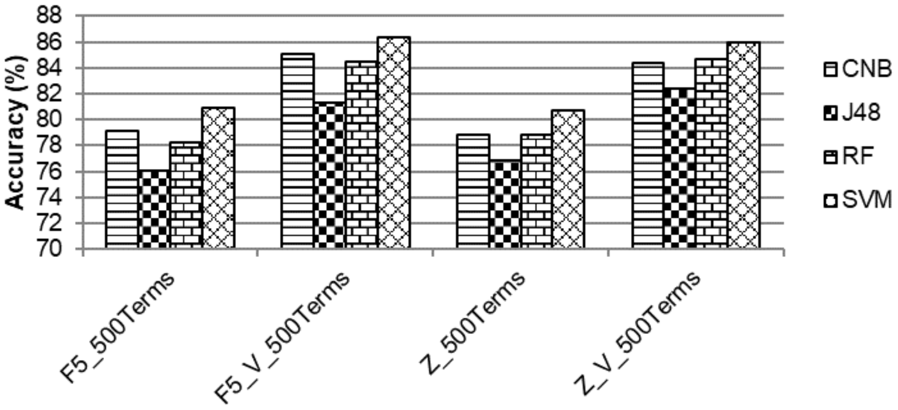


Figure 5.5 Overall accuracy values of the dataset versions subjected to the first 500 term selection approach

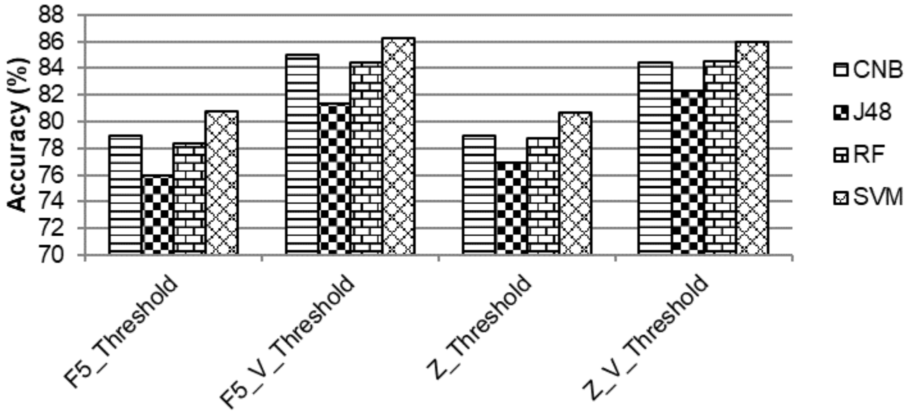


Figure 5.6 Overall accuracy values of the dataset versions subjected to the threshold term selection approach

In another experiment, we aim to determine which term selection approach yields higher classification results for the dataset versions F5 and F5_V. The results we obtained from the experiment are shown in Figure 5.7. It is obvious that there are no major differences in the results between term selection approaches. The first 500 term selection approach provides slightly higher results compared to threshold approach for the classifiers CNB, J48 and SVM applied on F5 dataset version. On the other hand, when the dataset version F5_V is used, there is a tie between the two term selection approaches. The first 500 term selection approach provided better results for the classifiers CNB and SVM, and the threshold approach achieved higher results by using J48 and RF algorithms.

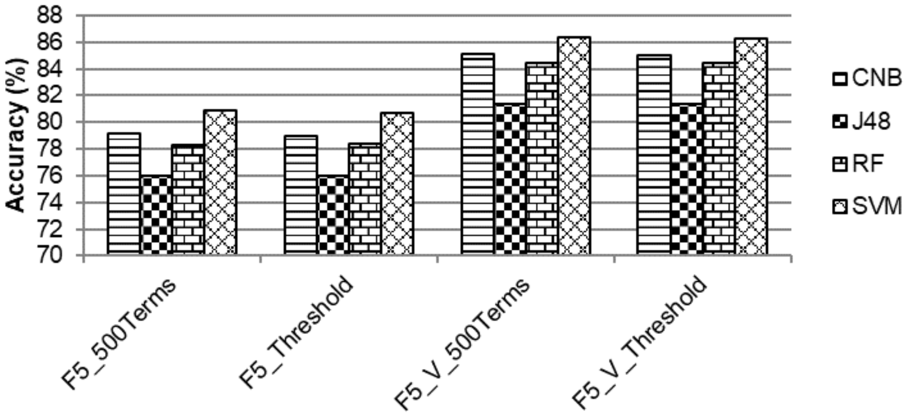


Figure 5.7 Overall accuracy values of the dataset versions F5 and F5_V subjected to the two term selection approaches

In Figure 5.8, overall accuracy values of the three F5_V dataset versions are compared to each other. One of them, F5_V, is not subject to any term selection approach, and the other two versions are subject to both term selection approaches. As a result, we concluded that term selection approaches applied to F5_V dataset version produce slightly better in accuracy values for CNB. On the other hand, the accuracy result of J48 is slightly better for the dataset version F5_V that is not subject to any selection approach.

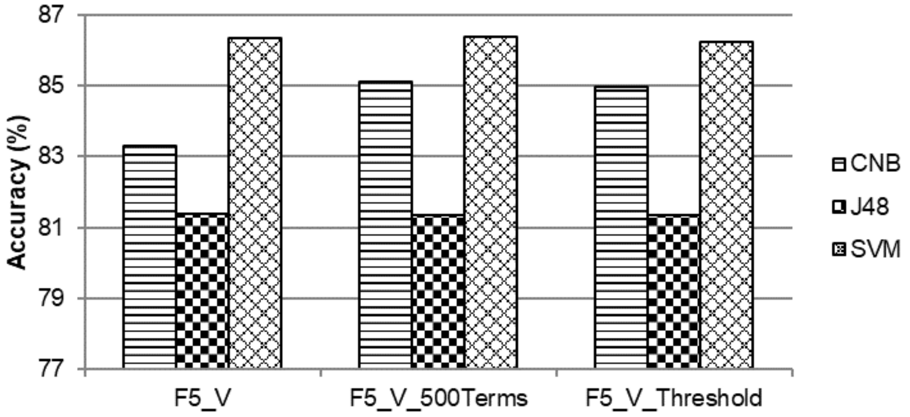


Figure 5.8 Overall accuracy values of all F5_V dataset versions

Tables 5.5 and 5.6 show the confusion matrix results of the models trained using the F5 and F5_V dataset versions for SVM classifier. In both tables, the rows represent the ground truth data, and the columns represent the classifier results. The last column of these tables provides the individual accuracy values of each emotion category. In Table 5.5, the accuracy value of the happiness emotion category is the highest. On the other hand, in Table 5.6, the disgust emotion category receives the highest score. Furthermore, we marked the most confused emotion classification results in boldface. For example, in Table 5.5, 275 happiness ground-truth entries are classified as entries indicating the surprise emotion category. The confusion matrix results of other dataset versions are shown in Appendix-2.

Table 5.5 Confusion matrix of SVM algorithm on F5 dataset version

	Happiness	**Fear**	**Anger**	**Sadness**	**Disgust**	**Surprise**	**Accuracy**
Happiness	4,052	99	117	100	57	**275**	86.21
Fear	164	3,758	164	**335**	124	71	81.41
Anger	155	151	3,653	**305**	227	145	78.80
Sadness	266	186	**328**	3,630	78	176	77.83
Disgust	92	132	**311**	90	3,832	65	84.74
Surprise	**419**	126	235	222	85	3,125	74.19

Table 5.6 Confusion matrix of SVM algorithm on F5_V dataset version

	Happiness	**Fear**	**Anger**	**Sadness**	**Disgust**	**Surprise**	**Accuracy**
Happiness	4,660	91	134	**160**	25	159	89.12
Fear	136	3,858	91	**234**	47	27	87.82
Anger	175	104	4,107	**212**	70	55	86.96
Sadness	**430**	162	235	4,072	30	92	81.10
Disgust	51	80	**148**	33	3,289	19	90.86
Surprise	**270**	62	101	93	18	2,459	81.88

Tables 5.7 and 5.8 show the precision, recall, and F-measures of each emotion categories. These results are obtained by using the F5 and F5_V dataset versions for SVM classifier. We obtained the highest values for the emotions disgust and happiness categories. The reason for this might be that these two categories also have the highest accuracy values, as shown in the confusion matrices in Tables 5.5 and 5.6. The precision, recall and F-measure values of other dataset versions are shown in Appendix-3.

Table 5.7 Precision, recall, and F-measure of SVM on F5 dataset version

	Precision	**Recall**	**F-Measure**
Happiness	0.787	0.862	0.823
Fear	0.844	0.814	0.829
Anger	0.76	0.788	0.774
Sadness	0.775	0.778	0.777
Disgust	0.87	0.847	0.859
Surprise	0.81	0.742	0.775
Average	0.807	0.806	0.806

Table 5.8 Precision, recall, and F-measure results of SVM on F5_V dataset version

	Precision	**Recall**	**F-Measure**
Happiness	0.814	0.891	0.851
Fear	0.885	0.878	0.882
Anger	0.853	0.87	0.861
Sadness	0.848	0.811	0.829
Disgust	0.945	0.909	0.927
Surprise	0.875	0,819	0.846
Average	0.865	0.864	0.864

5.2 Lexicon-Based Approach Experimental Results

After the completion of term selection process, to compare the performances based on the calculated cut-off value and all terms of the lexicon TREMO_LEX$_{\text{Consolidated}}$ in emotion analysis, we ran a set of experiments on the TestSet_T dataset. The purpose of running these experiments is to show the positive effects of term selection process. First, we focused on the comparison of the overall keyword-spotting results and then shared a confusion matrix showing the distribution of results based on four emotion categories. Figure 5.9 shows the comparison of overall keyword-spotting results of the four lexicons for all and selected terms. It is clear that, the term selection process increased the overall accuracy results.

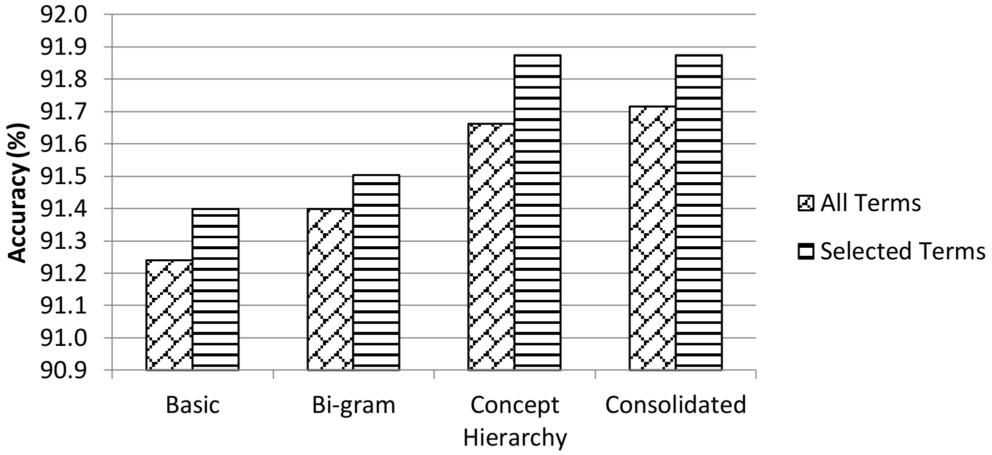


Figure 5.9 Comparisons of overall keyword-spotting results of four lexicons for all and selected terms

Table 5.9 shows confusion matrices, which compares the emotion categories, based on distribution numbers of the documents between keyword-spotting results and documents' original emotion categories. Each column in the matrix specifies the distribution results obtained for four emotion categories using keyword-spotting technique. The lines, however, represent the original document amounts for the corresponding emotion categories. Table 5.9 shows the keyword-spotting results of the lexicon, TREMO_LEX$_{\text{Consolidated}}$, for all terms and selected terms for each emotion categories. The results for each emotion category are shown in two columns where left is for all terms and right is for selected terms.

Table 5.9 Confusion matrix keyword-spotting results using TREMO_LEX$_{\text{Consolidated}}$. The results for each emotion category are shown in two columns where left is for all terms and right is for selected terms

		Keyword-spotting								
		Happiness		**Fear**		**Anger**		**Sadness**		**Total**
Training Set	**Happiness**	474	480	7	4	4	2	3	2	488
	Fear	9	15	446	444	5	4	11	8	471
	Anger	17	20	42	39	405	406	7	6	471
	Sadness	22	27	16	16	14	11	413	411	465

CHAPTER SIX
CONCLUSION

The main purpose of the thesis is to prepare a dataset, TREMO, and an emotive lexicon, TREMO_LEX, to be used in emotion analysis with six emotion categories in Turkish texts. To the best of our knowledge, they are both the first dataset and the lexicon to be used in emotion analysis in Turkish.

As the creation of TREMO_LEX emotive lexicon is based on TREMO dataset, firstly we focused on the construction of TREMO dataset in the thesis. To do this, first we conducted a survey and obtained 27,350 entries from 4,709 individuals. To validate this raw dataset, we conducted a validation process in which 48 people voluntarily participated. Here, a total of 92,986 validations were made, and at the end, 1,361 entries were discarded from the raw dataset owing to ambiguity in the emotion category.

After the validation process, two versions of datasets were formed, raw and validated datasets. Then we applied two stemmers, F5 and Zemberek NLP toolkit, on these two datasets to obtain four versions, F5, F5_V, Z, and Z_V. In the next step, we used MI for term selection on these four dataset versions. After this selection process, we modeled these versions to vector space model. Then, we used four supervised machine learning algorithms, CNB, J48, RF, and SVM, to compare the validation, stemming, and term selection effects of the dataset versions. In the classification process, we evaluated the performance in terms of accuracy, precision, recall, and F-measure. The results showed that for cases in which the validated datasets were used as the training datasets, classification results were higher than those using the raw sets by 5.7% of the F5 stemmed datasets and by 5.6% of the Zemberek NLP toolkit stemmed datasets. This shows that the validation process is succeeded for eliminating misleading declarations of emotion definitions by the users. On the other hand, the stemmers, F5 and Zemberek, showed similar performances. For term selection, we can say that both of the term selection approaches, selecting the first n terms or using a threshold, yield similar results, and that term selection in general decreased overall classification results. Comparing the classification algorithms to each other, the best algorithm among the four was SVM. The CNB and RF classifiers obtained results close to each other.

After the creation of TREMO dataset, we focused on the creation of TREMO_LEX lexicon set for lexicon-based emotion analysis in Turkish text for six emotion categories, namely happy, fear, anger, sadness, disgust and surprise. We used TREMO dataset, as the source of the newly created lexicon. The weight of each term is calculated based on term-class frequencies and MI values. To improve the term quality of lexicon we included bi-grams for high frequency terms and constructed concept hierarchy for low-frequency terms. We investigated the effects of stemming, term-weighting, lexicon enrichment methods and term selection approaches for lexicon-based emotion analysis.

For the evaluation of TREMO_LEX lexicon set, we performed a set of experiments to find out the best conditions. To do this, we first analyzed performance of two stemming approach: Zemberek and TurkLemma. Out of these results, we came to a conclusion that TurkLemma always outperformed Zemberek. Additionally, we observed that advanced schema produced higher accuracy values over simple schema. For lexicon construction, including bi-grams and concept hierarchies provides an improvement. Furthermore, we applied term selection for efficiency issues. In the lights of experiments we conducted, we selected a cut-off value as 0.00091329, which is the weight value of the 250th term of the happy emotion category. Then, we selected the top term for each emotion categories using the same cut-off value. As a result, we found that the term selection process improved the overall accuracy results for lexicon-based emotion analysis.

At the end of the evaluation processes, we managed to reach our goals by creating the lexicons in most effective and efficient ways. They are effective because we found out the best conditions to obtain the highest performances on calculating keyword-spotting results. On the other hand, they are efficient because we achieved to decrease the number of terms used within the corresponding lexicon in large proportions for each emotion category and still performances higher than full set of lexicons. We only used 3.46 percentage of the overall lexicon to perform keyword-spotting technique, without performance loss. Decreasing the number of terms in each lexicon supports to perform faster in the classification processes.

For future work, we have plans for both of TREMO and TREMO_LEX individually. We plan to use the TREMO dataset with different classifiers using different term selection approaches. In addition, we plan to find the optimum value for selecting the first *n* terms to obtain the highest accuracy value. On the other hand, we plan to perform a lexicon-based emotion analysis on different datasets, collected from social media applications, by using the lexicon TREMO_LEX. Then, we look forward to studying on automatic construction of concept hierarchy which is manually proposed within this thesis. In addition, word sense of the terms can be studied for extracting different meanings of each term within documents, to improve the quality of TREMO_LEX. Lastly, we intend to compare the classification results of the lexicon-based approach and machine learning algorithms where TREMO and TREMO_LEX are used together.

REFERENCES

Açıcı, E. (2012). *Emotion extraction from Turkish text*. BSc Thesis, Dokuz Eylul University, Izmir.

Akbas, E. (2012). *Aspect based opinion mining on Turkish tweets*. MSc Thesis, Bilkent University, Ankara.

Akın, A. A., & Akın, M. D. (2007). Zemberek, an open source NLP framework for Turkic Languages.

Albayrak N. B. (2011). *Opinion and sentiment analysis using natural language processing techniques*. MSc Thesis, Fatih University, İstanbul.

Alm, C., Roth, D., & Sproat, R. (2005). Emotions from text: machine learning for text-based emotion prediction. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 579-586.

Aly, M., & Atiya, A. (2013). LABR: A large scale Arabic book reviews dataset. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 494-498.

Aman, S., & Szpakowicz, S. (2007). Identifying Expressions of Emotion in Text. *International Conference on Text, Speech and Dialogue*, 196-205.

Aman, S., & Szpakowicz, S. (2008). Using Roget's Thesaurus for Fine-grained Emotion Recognition. *Proceedings of the Third International Joint Conference on Natural Language Processing*, *1*, 312-318.

Aslam S. (2018). *Twitter by the numbers: Stats, demographics and fun facts*. Retrieved June 1, 2018, from https://www.omnicoreagency.com/twitter-statistics.

Awwad, H., & Alpkocak, A. (2016). Performance Comparison of Different Lexicons for Sentiment Analysis in Arabic. *Network Intelligence Conference (ENIC)*, 127-133.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1996). CELEX2 LDC96L14, Linguistic Data Consortium.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *Proceedings of the 7th International Conference on Language Resources and Evaluation*, 2200-2204.

Bellegarda, J. (2010). Emotion analysis using latent affective folding and embedding. *Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, 1-9.

Benamara, F., Cesarano, C., Picariello, A., Reforgiato, D., & Subrahmanian, V. S. (2007). Sentiment analysis: Adjectives and adverbs are better than adjectives alone. *Proceedings of International Conference on Weblogs and Social Media (ICWSM).*

Bilgin, O., Çetinoğlu, Ö., & Oflazer, K. (2004). Building a wordnet for Turkish. *Romanian Journal of Information Science and Technology*, *7*(1-2), 163-172.

Blei, D. M., Ng, A.Y., & Jordan, M. I. (2003). Latent dirichlet allocation. The *Journal of Machine Learning Research*, *3*, 993–1022.

Boiy, E., Hens, P., Deschacht, K., & Moens, M. F. (2007). Automatic sentiment analysis of on-line text. *11th International Conference on Electronic Publishing*, 349-360.

Bougie, R., Pieters, R., & Zeelenberg, M. (2003). Angry customers don’t come back, they get back: The experience and behavioral implications of anger and dissatisfaction in services. *Journal of the Academy of Marketing Science*, *31*(4), 377-393.

Boucouvalas, A. C. (2003). *Real time text-to-emotion engine for expressive Internet communications*. Ios Press, 306-318.

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

Breazeal, C., & Brooks, R. (2004). *Robot emotions: A functional perspective.* In Who Needs Emotions. Oxford University Press.

Calvo, R. A., & Kim, S. M. (2013). Emotions in text: dimensional and categorical models. *Journal of Computational Intelligence*, *29*(3), 527-543.

Cambria, E., Olsher, D., & Rajagopal, D. (2014). Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 1515-1521.

Can, F., Kocberber, S., Balcik, E., Kaynak, C., Ocalan, H. C., & Vursavas, O. M. (2008). Information retrieval on Turkish texts. *Journal of the American Society for Information Science and Technology*, *59*(3), 407-421.

Chaffar, S. & Inkpen, D. (2011). Using a heterogeneous dataset for emotion analysis in text. *Proceedings of the 24th Canadian Conference on Advances in Artificial Intelligence*, 62-67.

Civriz, M. (2011). *Dictionary-based effective and efficient Turkish lemmatizer*. MSc Thesis, Dokuz Eylül University, İzmir.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37-46.

Danisman, T., & Alpkocak, A. (2008). Feeler: Emotion classification of text using vector space model. *Proceedings of the AISB 2008 Convention, Communication, Interaction and Social Intelligence*, 53-59.

Darwin, C. (1872). The Expressions of the Emotions in Man and Animals. *John Murray*.

Dave, K., Lawrence, S., & Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *Proceedings of WWW-03, 12th International Conference on the World Wide Web*, 519–528.

Dehkharghani, R., Yanikoglu, B., Saygin, Y., & Oflazer, K. (2015). Sentiment analysis in Turkish: Towards a complete framework. *Natural Language Engineering*.

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.

Demirci, S. (2014). *Emotion analysis on Turkish tweets*. MSc Thesis, Middle East Technical University, Ankara.

Ding, X., Liu, B., & Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. *Proceedings of the 2008 International Conference on Web Search and Data Mining*, 231-240.

Ekman, P. (1992). An argument for basic emotions. *Cognition and emotion*, *6*(3), 169-200.

Ekman, P., & Friesen, W. V. (2003). *Unmasking the Face: A Guide to Recognizing Emotions from Facial Expressions*. Malor Books: Cambridge.

Elarnaoty, M., AbdelRahman, S., & Fahmy, A. (2012). A machine learning approach for opinion holder extraction in Arabic language. arXiv preprint arXiv:1206.1011.

Elfenbein, H. A., & Ambady, N. (1994). Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological Bulletin*, *115*, 102–141.

Eroğul, U. (2009). *Sentiment analysis in Turkish*. MSc Thesis, Middle East Technical University, Ankara.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. MIT Press.

Fürnkranz, J. (1998). A study using n-gram features for text categorization. *Austrian Research Institute for Artifical Intelligence, 3,* 1-10.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report*, *1,* 1-12.

Gross, J. J. (1998). The emerging field of emotion regulation: An integrative review. *Review of General Psychology, 2*(3), 271–299.

Guo, K., Hall, C., Hall, S., Meints, K. & Mills, D. (2007). Left gaze bias in human infants, rhesus monkeys, and domestic dogs. *Animal Cognition*, *12*, 409–418.

Hailong, Z., Wenyan, G., Bo, J. (2014). Machine learning and lexicon based methods for sentiment classification: A survey. *Proceedings of the 11th Web Information System and Application Conference*, 262–265.

Hatzivassiloglou, V., & Mckeown, K. R. (1997). Predicting the semantic orientation of adjectives. *Proceedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics*, 174–181.

Hollinger, G., Georgiev Y., Manfredi A., Maxwell, B. A., Pezzementi Z. A., & Mitchell B. (2006). Design of a social mobile robot using emotion-based decision mechanisms. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 3093–3098.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 168-177.

James, W. (1884). What is an emotion? *Mind*, *9*, 188–205.

Katz, P., Singleton, M., & Wicentowski, R. (2007). SWAT-MP: The SemEval-2007 systems for task 5 and task 14. *Proceedings of the 4th International Workshop on Semantic Evaluations*, 308-313.

Kaya, M., Fidan, G., & Toroslu, I. H. (2012). Sentiment analysis of turkish political news. *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01*, 174–180.

Kemp, S. (2018). *Global Digital Report 2018: global overview.* Retrieved June 10, 2018, from https://wearesocial.com/uk/blog/2018/01/global-digital-report-2018

Kılınç, D., Özçift, A., Bozyigit, F., Yıldırım, P., Yücalar, F., & Borandag, E. (2015). TTC-3600: A new benchmark dataset for Turkish text categorization. *Journal of Information Science*, *43*(2), 174-185.

Kılınç, D., Yücalar, F., Borandağ, E., & Aslan, E. (2016). Multi-level reranking approach for bug localization. *Expert Systems, 33*(3), 286-294.

Kouloumpis, E, Wilson, T. & Moore, J. (2011). Twitter sentiment analysis: The good the bad and the omg! *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*, 538-541.

Lazarus, R. S. (2000). The cognition-emotion debate: A bit of history. In *Handbook of Cognition and Emotion* (1–20). New York: Guilford Press.

Lazarus, R. S. (1984). On the primacy of cognition. *American Psychologist*, *39*(2), 124–129.

Lewis, D. D. (1992). *Representation and Learning in Information Retrieval*. Phd Thesis, University of Massachusetts, Massachusetts.

Luhn, H. P. (1958) The automatic creation of literature abstracts. *IBM Journal of Research and Development*, *2*(2):159–165.

Manning, C. D, Raghavan, P., & Schütze, H. (2009). *An Introduction to information retrieval book.* Cambridge University Press.

Masson, J. M. (1996). *When Elephants Weep: The Emotional Lives of Animals*. New York: Delta.

Mladenic, D. & Grobelnik, M. (1998). Word sequences as features in text learning. *Proceedings of the 17th Electrotechnical and Computer Science Conference,* 145-148.

Mohammad, S.M. (2012). #Emotional tweets. *First Joint Conference on Lexical and Computational Semantics*, 246-255.

Mohammad S. M., & Kiritchenko, S. (2015). Using hashtags to capture fine emotion categories from tweets, *Computational Intelligence*, *31*(2), 301-326.

Mohammad, S. M. & Turney P. D. (2012). Crowdsourcing a word-emotion association lexicon, *Computational Intelligence, 29*(3), 436-465.

Mohammad, S. M. & Turney, P. D. (2010). Emotions evoked by common words and phrases: Using mechanical Turk to create an emotion lexicon. *Proceedings of Workshop on Computational Approaches to Analysis and Generation of Emotion in Text (CAAGET)*, 26-34.

Mohammad, S. (2012) Portable features for classifying emotional text. *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 587-591.

Musto, C., Semeraro, G., & Polignano, M. (2014). A comparison of Lexicon-based approaches for Sentiment Analysis of microblog posts. *International Workshop on Information Filtering and Retrieval*, 59-68.

Nakov, P., Kozareva, Z., Ritter, A., Rosenthal, S., Stoyanov, V., & Wilson, T. (2013). Semeval-2013 task 2: Sentiment analysis in twitter. *International Workshop on Semantic Evaluation*.

Neviarouskaya, A., Predinger, H. & M. Ishizuka. (2011). Affect analysis model: Novel rule-based approach to affect sensing from text. *Natural Language Engineering*, *17*, 95-135.

Nielsen, F. A. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *Workshop on making sense of microposts*, 93-98.

Nowlis, V., & Nowlis, H. H. (2001). The description and analysis of mood. *Annals of the New York Academy of Sciences, 65*(4), 345–355.

Ortony, A., Clore, G. L., & Collins, A. (1988). *The cognitive structure of emotions,* Cambridge University Press.

Ozkarahan, E. (1986). *Database machines and database management*. New Jersey: Prentice-Hall.

Pang, B., & Lee, L., (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1–2), 1–135.

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

Pedersen, T. (2001). A decision tree of bigrams is an accurate predictor of word sense. *Proceedings of the Second Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, 79–86.

Platt, J. C. (1998). *Sequential minimal optimization: A fast algorithm for training support vector machines.* Technical Report MSR-TR-98-14, Microsoft Research.

Plutchik, R. (1994). *The Psychology and Biology of Emotion*. New York:Harper Collins.

Plutchik, R., (1985). On emotion: The chicken-and-egg problem revisited. *Motivation and Emotion*, *9*(2), 197– 200.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. *Emotion: Theory, Research, and Experience*, *1*(3), 3–33.

Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. *Machine Learning*, *16*(3), 235-240.

Rennie, J. D. M., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of naïve bayes text classifiers. *Proceedings of the 20th International Conference on Machine Learning*, 616-623.

Russell, J. A. (1994). Is there universal recognition of emotion from facial expression?A review of the crosscultural studies. *Psychological Bulletin, 115,* 102–141.

Scherer, K. R. (1984). Emotion as a multicomponent process: A model and some cross-cultural data. *Review of Personality and Social Psychology*, *5*, 37–63.

Scherer, K. R., & Wallbott, H. G. (1994). Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology*, *66*(2), 310-328.

Sevindi, B. I. (2013). *Comparison of supervised and dictionary based sentiment analysis approaches on Turkish text*. MSc Thesis, Gazi University, Ankara.

Shaikh, M.A.M. (2008). *An analytical approach for affect sensing from text.* PhD Thesis, University of Tokyo, Tokyo.

Steunebrink, B. R. (2010). *The logical structure of emotions*. PhD Thesis, Utrecht University, Utrecht.

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). *The general inquirer: A Computer Approach to Content Analysis*. MIT Press.

Strapparava, C., & Mihalcea, R. (2008). Learning to identify emotions in text. *Proceedings of the 2008 ACM Symposium on Applied Computing,* 1556-1560.

Strapparava, C. & Mihalcea, R. (2007). SemEval-2007 Task 14: Affective text. *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007),* 70-74.

Strapparava, C., & Valitutti A. (2004). Wordnet-affect: an affective extension of wordnet. *Proceedings of the 4th International Conference on Language Resources and Evaluation*, 1083-1086.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Journal Computational Linguistics*, *37*(2), 267-307.

Taboada, M., Caroline, A. & Kimberly, V., (2006). Creating semantic orientation dictionaries. *Proceedings of 5th International Conference on Language Resources and Evaluation (LREC)*, 427–432.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas A. (2010). Sentiment strength detection in short informal text. *Journal of the Association for Information Science and Technology*, *61*(12), 2544–2558.

Tocoglu, M. A., & Alpkocak, A. (2018). TREMO: A dataset for emotion analysis in Turkish. *Journal of Information Science,* 1-18.

Tocoglu, M. A., & Alpkocak, A. (2014). Emotion extraction from Turkish text. *Network Intelligence Conference (ENIC)*, 130–133.

Tocoglu, M. A., & Alpkocak, A. (2018). *TREMO dataset*. Retrieved May 5, 2018, from https://demir.cs.deu.edu.tr/tremo

Tocoglu, M. A., & Alpkocak, A. (2018). *TREMO_Lex lexicon*. Retrieved May 5, 2018, from https://demir.cs.deu.edu.tr/tremolex

Tong, R. M. (2001). An operational system for detecting and tracking opinions in on-line discussions. *In Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification*, 1–6.

Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics*, 417–424.

Turney, P. D., & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems, 21*(4), 315-346.

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

Vural, A. G., Cambazoglu, B. B., Senkul, P., & Tokgoz, Z.O. (2013). A Framework for sentiment analysis in Turkish: Application to polarity detection of movie reviews in Turkish. *Computer and Information Sciences III*, 437-445.

Whitelaw, C., Garg N., & Argamon, S. (2005) Using appraisal groups for sentiment analysis. *Proceedings of ACM SIGIR Conference on Information and Knowledge Management (CIKM2005)*, 625–631.

Wiebe, J. M. (1994). Tracking point of view in narrative. *Computational Linguistics, 20*(2), 233–287.

Wiebe, J. (2000). Learning subjective adjectives from corpora. *Proceedings of 17th National Conference on Artificial Intelligence(AAAI)*, 735–740.

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language resources and evaluation, 39*(2-3), 165-210.

Wikimedia Commons (2018). *Plutchik-wheel*. Retrieved June 16, 2018, from https://commons.wikimedia.org

Witten, I. H. & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco: Morgan Kaufman.

Xie, R., & Li, C. (2012). Lexicon construction: A topic model approach. *In Systems and Informatics (ICSAI), 2012 International Conference*, 2299-2303.

Xu, B, Guo, X, Ye, Y., & Cheng, J. (2012). An improved random forest classifier for text categorization. *Journal of Computers*, *7*(12), 2913-2920.

Yang, C, Lin, K.H. & Chen, H. (2007). Emotion classification using web blog corpora. *IEEE/WIC/ACM International Conference on Web Intelligence*, 275-278.

Yang, M, Peng, B., Chen, Z., Zhu, D., & Chow, K.P. (2014). A topic model for building fine-grained domain-specific emotion lexicon. *The 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, 421-426.

Yashar, M. (2012). *Role of emotion in information retrieval*. PhD Thesis, University of Glasgow, Glasgow.

Zajonc, R. B. (1984). On the primacy of affect. *American Psychologist*, *39*(2), 117–123.

APPENDICES

Appendix-1: Sample of Emotion Vector based Lexicon

Table 1 A sample of Emotion vectors of advanced weighted TREMO_LEX$_{\text{Consolidated}}$ lexicon

Term	Happy	Fear	Anger	Sadness	Disgust	Surprise
mutlu	0.40099	0.00405	0.00439	0.01595	0	0.00634
mutlu olu	0.14151	0	0	0	0	0
kork	0	0.32559	0	0.0054	0	0
öfkelen	0	0	0.55204	0	0.00382	0
üzül	0	0.00478	0.00826	0.3078	0	0.00369
vefat	0	0	0	0.09571	0	0.00069
tiksin	0	0	0	0	2.15937	0
koku	0.00107	0	0	0	0.08808	0
şaşır	0.00618	0	0	0	0	0.34658
sürpriz	0.00255	0	0	0	0	0.03914

Appendix-2: Confusion Matrix of SVM Algorithm for Different Dataset Versions

Table 1 Confusion matrix of SVM algorithm on F5_500Terms dataset version

	Happiness	**Fear**	**Anger**	**Sadness**	**Disgust**	**Surprise**	**Accuracy**
Happiness	4,061	107	138	88	57	249	86.40
Fear	169	3,778	167	311	117	74	81.85
Anger	176	127	3,696	271	230	136	79.72
Sadness	290	175	319	3,612	81	187	77.44
Disgust	90	124	319	65	3,863	61	85.43
Surprise	431	120	237	205	101	3,118	74.03

Table 2 Confusion matrix of SVM algorithm on F5_Threshold dataset version

	Happiness	**Fear**	**Anger**	**Sadness**	**Disgust**	**Surprise**	**Accuracy**
Happiness	4,051	112	131	89	54	263	86.19
Fear	164	3,784	165	318	113	72	81.98
Anger	182	130	3,699	269	219	137	79.79
Sadness	299	172	330	3,594	79	190	77.06
Disgust	93	133	320	70	3,844	62	85.01
Surprise	439	127	235	210	95	3,106	73.74

Table 3 Confusion matrix of SVM algorithm on F5_V_500Terms dataset version

	Happiness	**Fear**	**Anger**	**Sadness**	**Disgust**	**Surprise**	**Accuracy**
Happiness	4,701	95	128	146	15	144	89.90
Fear	171	3,826	100	210	59	27	87.09
Anger	210	81	4,134	200	52	46	87.53
Sadness	480	129	234	4,060	24	94	80.86
Disgust	91	65	143	28	3,279	14	90.58
Surprise	296	48	106	91	14	2,448	81.52

Table 4 Confusion matrix of SVM algorithm on F5_V_Threshold dataset version

	Happiness	**Fear**	**Anger**	**Sadness**	**Disgust**	**Surprise**	**Accuracy**
Happiness	4,710	91	128	142	16	142	90.07
Fear	167	3,829	97	211	58	31	87.16
Anger	228	76	4,109	205	53	52	87.00
Sadness	493	126	235	4,046	26	95	80.58
Disgust	91	65	139	32	3,277	16	90.52
Surprise	304	51	108	85	15	2,440	81.25

Table 5 Confusion matrix of SVM algorithm on Z dataset version

	Happiness	**Fear**	**Anger**	**Sadness**	**Disgust**	**Surprise**	**Accuracy**
Happiness	4,025	112	138	104	70	251	85.64
Fear	175	3,750	166	314	123	88	81.24
Anger	208	112	3,665	316	197	137	79.07
Sadness	325	180	310	3,591	75	183	76.99
Disgust	125	127	311	83	3,811	64	84.30
Surprise	441	133	239	206	77	3,116	73.98

Table 6 Confusion matrix of SVM algorithm on Z_500Terms dataset version

	Happiness	**Fear**	**Anger**	**Sadness**	**Disgust**	**Surprise**	**Accuracy**
Happiness	4,058	107	138	79	77	241	86.34
Fear	184	3,767	173	301	114	77	81.61
Anger	220	109	3,705	282	193	126	79.94
Sadness	335	176	297	3,599	79	178	77.17
Disgust	121	132	326	66	3,814	62	84.36
Surprise	449	142	214	186	88	3,133	74.38

Table 7 Confusion matrix of SVM algorithm on Z_Threshold dataset version

	Happiness	**Fear**	**Anger**	**Sadness**	**Disgust**	**Surprise**	**Accuracy**
Happiness	4,056	112	136	86	74	236	86.30
Fear	192	3,757	170	308	112	77	81.39
Anger	221	111	3,703	285	189	126	79.89
Sadness	338	175	304	3,596	75	176	77.10
Disgust	122	130	317	66	3,827	59	84.65
Surprise	452	143	223	183	82	3,129	74.29

Table 8 Confusion matrix of SVM algorithm on Z_V dataset version

	Happiness	**Fear**	**Anger**	**Sadness**	**Disgust**	**Surprise**	**Accuracy**
Happiness	4,646	82	144	181	31	145	88.85
Fear	167	3,822	93	235	46	30	87.00
Anger	213	84	4,102	219	55	50	86.85
Sadness	506	148	237	4,020	25	85	80.06
Disgust	81	79	123	45	3,278	14	90.55
Surprise	285	62	98	91	12	2,455	81.75

Table 9 Confusion matrix of SVM algorithm on Z_V_500Terms dataset version

	Happiness	**Fear**	**Anger**	**Sadness**	**Disgust**	**Surprise**	**Accuracy**
Happiness	4,671	80	149	159	32	138	89.33
Fear	188	3,812	97	219	50	27	86.77
Anger	234	68	4,128	192	56	45	87.40
Sadness	543	131	234	4,010	20	83	79.86
Disgust	89	74	129	40	3,272	16	90.39
Surprise	312	62	95	70	14	2,450	81.59

Table 10 Confusion matrix of SVM algorithm on Z_V_Threshold dataset version

	Happiness	**Fear**	**Anger**	**Sadness**	**Disgust**	**Surprise**	**Accuracy**
Happiness	4,690	79	149	143	30	138	89.69
Fear	191	3,805	99	221	51	26	86.62
Anger	235	73	4,119	196	57	43	87.21
Sadness	552	127	236	4,007	19	80	79.80
Disgust	90	75	129	41	3,271	14	90.36
Surprise	316	55	97	76	13	2,446	81.45

Appendix-3: Precision, Recall, and F-measure Values of SVM Algorithm

Table 1 Precision, recall, and F-measure of SVM on F5_500 Terms dataset version

	Precision	Recall	F-Measure
Happiness	0.778	0.864	0.819
Fear	0.853	0.818	0.835
Anger	0.758	0.797	0.777
Sadness	0.793	0.774	0.784
Disgust	0.868	0.854	0.861
Surprise	0.815	0.74	0.776
Average	0.811	0. 809	0.809

Table 2 Precision, recall, and F-measure of SVM on F5_Threshold dataset version

	Precision	Recall	F-Measure
Happiness	0.775	0.862	0.816
Fear	0.849	0.82	0.834
Anger	0.758	0.798	0.777
Sadness	0.79	0.771	0.78
Disgust	0.873	0.85	0.861
Surprise	0.811	0.737	0.772
Average	0.809	0.807	0.807

Table 3 Precision, recall, and F-measure of SVM on F5_V_500 Terms dataset version

	Precision	Recall	F-Measure
Happiness	0.79	0.899	0.841
Fear	0.902	0.871	0.886
Anger	0.853	0.875	0.864
Sadness	0.857	0.809	0.832
Disgust	0.952	0.906	0.929
Surprise	0.883	0.815	0.848
Average	0.867	0.864	0.864

Table 4 Precision, recall, and F-measure of SVM on F5_V_Threshold dataset version

	Precision	Recall	F-Measure
Happiness	0.786	0.901	0.839
Fear	0.903	0.872	0.887
Anger	0.853	0.87	0.862
Sadness	0.857	0.806	0.831
Disgust	0.951	0.905	0.928
Surprise	0.879	0.813	0.844
Average	0.866	0.862	0.863

Table 5 Precision, recall, and F-measure of SVM on Z dataset version

	Precision	Recall	F-Measure
Happiness	0.76	0.856	0.805
Fear	0.85	0.812	0.831
Anger	0.759	0.791	0.775
Sadness	0.778	0.77	0.774
Disgust	0.875	0.843	0.859
Surprise	0.812	0.74	0.774
Average	0.805	0.803	0.803

Table 6 Precision, recall, and F-measure of SVM on Z_500 Terms dataset version

	Precision	Recall	F-Measure
Happiness	0.756	0.863	0.806
Fear	0.85	0.816	0.833
Anger	0.763	0.799	0.781
Sadness	0.797	0.772	0.784
Disgust	0.874	0.844	0.858
Surprise	0.821	0.744	0.78
Average	0.81	0.807	0.807

Table 7 Precision, recall, and F-measure of SVM on Z_Threshold dataset version

	Precision	**Recall**	**F-Measure**
Happiness	0.754	0.863	0.805
Fear	0.848	0.814	0.831
Anger	0.763	0.799	0.781
Sadness	0.795	0.771	0.783
Disgust	0.878	0.846	0.862
Surprise	0.823	0.743	0.781
Average	0.809	0.807	0.807

Table 8 Precision, recall, and F-measure results of SVM on Z_V dataset version

	Precision	**Recall**	**F-Measure**
Happiness	0.788	0.889	0.835
Fear	0.894	0.87	0.882
Anger	0.855	0.869	0.862
Sadness	0.839	0.801	0.819
Disgust	0.951	0.906	0.928
Surprise	0.883	0,818	0.849
Average	0.862	0.859	0.859

Table 9 Precision, recall, and F-measure of SVM on Z_V_500 Terms dataset version

	Precision	**Recall**	**F-Measure**
Happiness	0.774	0.893	0.829
Fear	0.902	0.868	0.884
Anger	0.854	0.874	0.864
Sadness	0.855	0.799	0.826
Disgust	0.95	0.904	0.926
Surprise	0.888	0.816	0.85
Average	0.863	0.86	0.86

Table 10 Precision, recall, and F-measure of SVM on Z_V_Threshold dataset version

	Precision	**Recall**	**F-Measure**
Happiness	0.772	0.897	0.83
Fear	0.903	0.866	0.884
Anger	0.853	0.872	0.862
Sadness	0.855	0.798	0.826
Disgust	0.951	0.904	0.926
Surprise	0.89	0.815	0.851
Average	0.864	0.86	0.86