# EMOTION ANALYSIS ON TURKISH TEXT

**HATİCE ERTUĞRUL GİRAZ**

A Thesis Submitted to

The Graduate School of Izmir University of Economics

Master's Program in Computer Engineering

İzmir

2020

# ABSTRACT


EMOTION ANALYSIS ON TURKISH TEXT



Ertuğrul Giraz, Hatice

Computer Engineering Master's Program

Advisor: Assoc. Prof. Dr. Senem Kumova Metin



August, 2020



Emotions govern our daily life; these are a big part of the human experience and affect our decision-making. We would like to repeat actions that make us happy, but we avoid actions that make us angry or sad. Through natural language processing, subjective information can be obtained from written sources such as suggestions, reviews, and social media publications. It also allows us to understand the emotions expressed by the author of the text and therefore act accordingly.

In this thesis, we explored the effect of different emotion lexicons and word vector representations of text in emotion detection task. We proposed two approaches to construct emotion lexicon. In addition, we represented the sentences by three different approaches based on word vector comparison to emotion lexicon words. Experiments are performed employing both unsupervised and supervised approaches in Turkish texts and the results are reported.

Keywords: Emotion Analysis, Vector Representation, Machine Learning, Turkish.

# ÖZET

TÜRKÇE METİNLERDE DUYGU ANALİZİ

Ertuğrul Giraz, Hatice

Bilgisayar Mühendisliği Yüksek Lisans Programı

Tez Danışmanı: Doç. Dr. Senem Kumova Metin

Ağustos, 2020

Duygular günlük hayatımızı yönetir; duygular insan deneyiminin büyük bir parçasıdır ve karar verme sürecimizi etkiler. Bizi mutlu eden eylemleri tekrarlamak isteriz, ancak bizi kızdıran veya üzen eylemlerden kaçınırız. Doğal dil işleme sayesinde; öznel bilgiler, öneriler, incelemeler, sosyal medya yayınları gibi yazılı kaynaklardan elde edilebilir. Ayrıca, metnin yazarı tarafından ifade edilen duyguları anlamamızı ve dolayısıyla buna göre hareket etmemizi sağlar.

Bu tezde duygu sözlüklerinin ve sözcük vektör gösterimlerinin duygu belirleme görevindeki başarısı araştırılmıştır. Duygu sözlüğü oluşturmak için iki yöntem önerilmiştir. Buna ek olarak, cümlelerin ifadesinde cümleyi oluşturan kelimelerin

duygu sözcükleri ile vektörel kıyaslanmasını temel alan üç yaklaşım önerilmiştir. Türkçe metinlerde gözetimsiz ve gözetimli yöntemler ile deneyler yürütülmüş ve deney sonuçları raporlanmıştır.

Anahtar Kelimeler: Duygu Analizi, Vektör Temsili, Makine Öğrenmesi, Türkçe.

## ACKNOWLEDGEMENT

I would like to take the opportunity to thank some people who supported me during the process of this thesis.

First of all, I would like to thank to my supervisor Assoc. Prof. Dr. Senem KUMOVA METİN who gave me the opportunity to work with her, for her excellent guidance, encouragement and patience.

I would like to thank to my mother Ayşe ERTUĞRUL, my father Aziz ERTUĞRUL and my brother Mustafa ERTUĞRUL for always being with me and motivating me with their endless support.

Lastly, I am very thanks full to my husband, Abdülkadir GİRAZ for his trust and unconditional support all the time.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

# CHAPTER 1 : INTRODUCTION

Today, emotion analysis is becoming more and more popular and a lot of research is being done on this topic. As psychological theory, an emotion is “a complex psychological condition that includes three separate components: a subjective experience, a physiological response, and a behavioral or meaningful response” (García, 2019). We don't all react to similar situations in the same way. For example, it may sound funny to someone who bothers another person. This is an example of the subjectivity of emotions. In addition, when we feel the emotion, we may have some physiological responses, such as increased heart rate, sweating and rapid breathing. Also, basically everyone feels some emotions for certain topics. At this stage, companies use emotion analysis for customer satisfaction or feedback on the product. This issue is the same among researchers. There are many studies in the literature for the English language. At the same time, it is promising in the studies for the Turkish language every day. We are part of this and in this thesis, we focused on emotion analysis for the Turkish language.

In this thesis, we studied on emotion detection in Turkish texts. Our main motivation is to speed things up and to decrease the effort used in emotion detection. Today, most data are stored and processed virtually. For example, thousands of tweets can be sent a day, hundreds of complaints can be received for a brand, or dozens of comments can be written for a product. These data can be too large and complex for a person to read and process. For example, for marketing analysis, we want to automate processes such as a customer's feedback about a product or brand, information about the product, or customer transactions, and speed up such things. In addition, we aim faster and increased customer satisfaction in this way. With the emotion analysis models used, we aim to automate the business processes of companies and to obtain meaningful information, to minimize manpower and to obtain better performance results. In addition, thanks to the automated system, it is possible to eliminate human based errors.

In this thesis, we explored the performance of lexicon-based emotion detection approaches on Turkish texts. The concept of emotion lexicon refers to a list of words (will be named as emotion words or lexicon words) and their associations with a set of emotions (such as anger, fear, surprise, sadness, joy, disgust). Simply, in lexicon-based approaches, lexicon words and text units (e.g. sentence, paragraph) are compared in several different ways and the comparison results are used in emotion identification experiments. As this identification procedure is examined, it is explicit that the performance depends on two actors. The first actor is the emotion lexicon employed in experiments. Simply, if the set of emotion words in lexicon are sufficient and competent enough to represent emotions, the decision given by comparison of these words and text units will be succeeding. The second actor in emotion identification is the comparison method. In lexicon-based approaches, comparison may be performed in several different ways. A simple string-matching algorithm may be implemented to count the number of matching words or more complex algorithms that enable multiple dimensions of similarity to different emotions may be applied.

The thesis focuses on two research goals based on two actors of lexicon-based emotion identification. The first goal is presenting new lexicon construction methods that will decrease the effort required while increasing the performance in emotion detection. We proposed the use of word vector similarity in the determination of emotion word candidates as an alternative to crowd sourcing approaches. We built up two lexicons based on proposed approach and reported the performance results of experiments. The second is proposing a new method to measure the similarity of text and emotion words. In this study, as an alternative to traditional methods that employ surface forms of words in text, both the text and emotion words is proposed to be represented by a vector. The vector/vector representation is assumed to hold the overall meaning/semantic information contained in text unit. The vector representations are obtained from an independent resource and they are used in both lexicon construction and emotion identification stages.

The thesis is structured as follows; Chapter 2 is where the concept of emotion and emotion analysis is given. Chapter 3 presents a survey of the literature. In Chapter 4, the methodology of the study and in Chapter 5 experimental results are given. Finally, in Chapter 6, we conclude our study.

# CHAPTER 2 : THE CONCEPT OF EMOTION AND EMOTION ANALYSIS

Emotion is the mental state caused by the influence of the environment. In Turkish, thought, arousal and intuition are used inadvertently in the sense of meaning and turmoil occurs. Emotion in English is distinct and is often a means of social knowledge. According to the Oxford dictionary, it is the strong feeling that one feels about his condition, mode and relationship with others. Wikipedia defines satisfaction and dissatisfaction as a result of thoughts, behaviors and feelings (Yüce, 2019).

There are many different types of emotions that affect interact with others. The actions we make, the choices we make and the perceptions we have are affected by the emotions we experience at every moment.

According to, psychologists have also tried to identify the different types of emotions people experience (Nasilbe, 2018). Several different theories have emerged to classify and explain the emotions people feel. According to the theory of psychologist Paul Ekman in the 1970s, there exist six basic emotions, depicted by exemplary facial images in Figure 2.1. The emotions he described were happiness, fear, anger, sadness, disgust and surprise. He then expanded the list of basic emotions including some other emotions, such as pride, shame and excitement (Ekman, 1992).


Figure 2.1. Paul Ekman's six emotions (Source: Envision Your Evolution, 2019)

Psychologist Robert Plutchik (1980) came up with the idea of a "wheel of emotion" as shown in Figure 2.2. According to this theory, several different emotions can be mixed with each other to form an emotion. Just like we mix it with some colors to create some colors. Also, according to Plutchik, more basic emotions act like building blocks. More complex emotions are a mixture of these basic ones. For example, basic emotions such as joy and trust can be combined to create joy.


Figure 2.2. Adapted from Robert Plutchik's wheel of emotions (Source: Standard, 2017)

Though there exists several different approaches in emotion categorization, such as Robert Plutchik’s, in a wide range of studies, still six basic emotions proposed by Ekman are accepted: happiness, sadness, disgust, fear, surprise and anger. These six basic emotions are defined as follows, according to (Nasilbe, 2018).

- ***Happiness*****:** This emotion, unlike other emotions, is the mood that most people want to be. Happiness is often described as a pleasant emotional state characterized by feelings of joy, satisfaction, and well-being. Since the 1960s, research on happiness has increased significantly in some psychology disciplines. The feeling of happiness is sometimes expressed as:
  - Smiley-like facial expressions
  - Body language like a relaxed posture

- Optimistic, pleasant tone

For many years, people have believed that happiness and well-being are interdependent. Additionally, the idea that Happiness plays an important role in mental and physical health was supported. The feeling of happiness has been associated with peaceful marriage and long life. On the contrary, unhappiness has been shown to be the cause of many health problems. For example, emotions such as anxiety, stress, loneliness, depression, increased inflammation, reduced immunity, and reduced lifespan have been associated with illnesses.

- ***Sadness*:** This feeling is often described as a temporary emotional state. It is often associated with sensory states such as grief, apathy, frustration, and despair. Sadness, like other emotions, is a type of emotion that many people experience occasionally. As a result of certain events and situations, people may experience severe and prolonged episodes of depression. The cause, type or severity of this feeling may vary from person to person. Also, the method of dealing with these emotions can be different for everyone.

- ***Fear*:** Fear is a powerful emotion that can play an important role in survival. It is a strong emotion compared to other emotions. It plays an important role in human instinctual survival. When you encounter a danger in any situation, your whole body is quickly warned. Your heart rate and breathing increase, your muscles become tense and your mind becomes more alert. This situation stimulates our body and forces it to fight or get away from danger.

  In fact, fear is a rapid emotional response to a situation. Also, we give similar responses in our minds for future potential threats. This reaction is often called anxiety. On the other hand, some people want to experience fear of their own accord. People like this enjoy the fear of extreme sports, scary environments, or exciting games. Also, repeated exposure to a certain fear state leads to habituation. This may lead to a decrease in fear. For example, the roller coaster in the amusement does not give you the same fear every time.

- ***Disgust*****:** The feeling of disgust can be caused by a number of things, such as an unpleasant taste, sight or smell. A sense of disgust is an emotional response to situations such as an unpleasant taste, an unpleasant sight, or a foul odor. Researchers believe this emotion develops in response to food that could be fatal or harmful. Many people have the same typical reaction when they taste or smell a bad food.

- ***Anger*****:** Anger is a powerful emotion that results from feelings of frustration or hostility towards a person or object. Feelings of anger can also play a role in our body's war response. When a threat triggers anger, we often tend to protect ourselves and move away from danger. Anger is usually represented as:
  - Facial expressions such as frowning or glare
  - Loud voice tone
  - Physiological responses, such as sweating or turning red
  - Body language, such as taking a strong posture or getting away from someone
  - Aggressive behavior, such as hitting, kicking or throwing objects

  Feelings of anger can be the trigger to mobilize you, clarify your wishes in your human relationships, and find solutions to problems for you. Contrary to what is thought, this shows that the feeling of anger is not always a negative situation.

- ***Surprise*****:** Unlike other emotions, the feeling of surprise is a reaction that occurs after an unexpected situation and takes a very short time. The feeling of surprise can be shown as a result of positive, negative or neutral situations. Sometimes there can be unpleasant periods. For example, the sudden appearance of a friend in the dark may surprise you negatively. As another example, an offer you received at an unexpected moment can positively surprise you.

The six basic emotions (such as happy, fear, anger, sadness, disgust and surprise) defined by Ekman and discussed earlier are just a few of the many different types of emotions people experience. In the Ekman theory, he suggests that these six basic emotions are universal to the whole world. According to him these basic feelings are the same everywhere. However, other theories and new research continue to explore many different types of emotions and how they are classified. Ekman later expanded the list and added other emotions. However, he suggested that unlike the original six emotions, all of these could be encoded with facial expressions. Some of the emotions he described later are entertainment, satisfaction, excitement, humiliation, trouble, pride in success, crime, satisfaction and shame.

The concept of emotion analysis is commonly coupled with the sentiment analysis, in literature. According to the Oxford dictionary (2020), the definition of sentiment analysis is the process of defining and classifying views expressed in a piece of text to determine whether an author's attitude to a particular topic is positive, negative, or neutral. From this perspective, emotion analysis can be interpreted as an extension to sentiment analysis. But in emotion analysis, a more sophisticated and complex system is to be employed to run a deeper analysis of human emotions and sensitivities. In emotional analysis, different variations of human mental subjectivities must be considered.

According to Pervan (2019), emotion analysis aims to divide the text into multiple classes according to its subjective content. The way that the text is analyzed varies based on the ultimate goal to reach, in other words the expected output of the analysis. For example, in some studies, just the polarity of the text is considered similar to sentiment analysis, on the other hand multiple feelings that are to be transferred to the reader are retrieved in other studies. Some examples to this different type of emotion analysis are listed below:

- **Fine-grained analysis:** In some problems, it may not be sufficient to evaluate the text as positive or negative. In this case, by classifying fine grains in text, the outputs are diversified as very positive, positive, neutral, negative, and very negative. In this type of analysis, very positive and very negative expressions also provide an opportunity to evaluate whether the speaker is angry or happy.
- **Emotion detection:** It aims to detect emotions such as happiness, sadness, frustration, anger in text. Commonly, a list of words (named as emotion lexicon) showing the emotion of the text are employed in this type of analysis,

though the lexicon words vary depending on the judgments of that society or the content of the data set created.

- **Aspect-based emotion analysis:** In aspect-based emotion analysis, it is aimed to measure the satisfaction of people about a feature of a product rather than determining their positive or negative opinion about any product.

As Pervan (2019) points out in SisaSoft's blog, 80% of the world data is estimated to be unstructured. These data consist of texts such as emails, chats, social media comments, surveys, articles. It is obvious that retrieving information in such an unstructured data is hard for any type of systems. By analyzing this great amount of data, emotion analysis enables companies to advance in automating all kinds of processes such as product feedback, public relations, marketing analysis, product reviews, and customer service.

Emotion analysis models aim to automate business processes of companies and extract meaningful information, minimizing human power and obtaining clearer results. Thanks to an automatic system, it is possible to eliminate the human component's error. The main advantages of emotion analysis systems are

- **Scalability:** Thousands of tweets, customer support meetings, or even manually editing customer comments are difficult to imagine. Emotion analysis allows data to be scaled up efficiently and cost-effectively.
- **Real-time analysis:** During a particular situation, emotional analysis can be used in real time to identify critical information. The post of an angry customer on social media is an example where real-time analysis is required. In such cases, it is of great importance to make due diligence and respond quickly.
- **Consistent criteria:** People may not be able to make a consistent assessment when determining the emotion of a text. It is estimated that only 60-65% of different people make the same classification when they evaluate the feeling of a particular text. Emotion analysis is a subjective area that is heavily influenced by people's experiences, beliefs, and thoughts, but automated emotion analysis systems reduce errors and increase data consistency.

While emotion analysis systems provide many assets to both natural language processing systems and business processes of companies, researchers face some challenging situations due to language differences, lack of data and noisy data during emotion analysis. Pervan (2019) from SisaSoft gave some examples of these difficulties as follows.

- The basis of automatic emotion analysis is based on the use of large data sets. To use the emotion analysis system and take action, it is necessary to have a big data source and volume. When there is not enough data to feed the system, it is inevitable for the model to make an incorrect evaluation.
- Since the emotion analysis aims to extract meaning from subjective texts, the texts created can be written in everyday speaking language or without paying attention to grammar rules. In addition, texts can contain metaphorical and cynical expressions. Also, data capacity is limited in defining various human emotions such as anxiety and fear.
- Multilingual texts are also a challenge. While it is possible to reach a lot of data and research especially in English language, the number of resources in other languages is very low. Each language has its own sentence structure, emotional words and metaphoric expressions. The emotional analysis models created in this case also differ. In multilingual texts, determining the language of the text and developing a language-appropriate model can also be a step.

# CHAPTER 3 : LITERATURE REVIEW

In today's world, which is becoming more digital every day, many things are at the tip of one finger. It is very easy to shop, comment on social media platforms or express their feedbacks about a product. At this point, customer satisfaction or emotion analysis of customers has become very important for companies and foundations. Today, there are many studies on this field in the literature. Numerous studies, especially for the English language, show how deep and important the subject actually is. In addition, recent studies in Turkish language are very promising. However, there are two approaches that should not be confused with these analyzes. The first one is Emotion Analysis, and the second one is Sentiment Analysis, also known as Opinion Mining. These two important approaches can often be confused. Emotion analysis focuses on different emotion categories such as happiness, anger or fear, while sentiment analysis mainly focuses on negative, positive, and neutral categories. In the literature, there are studies both in Turkish and English languages for both sentiment and emotion analyses. However, sentiment analysis is more popular than the emotion analysis. Actually, there are two reasons for this. The first reason is that companies or foundation wants more precise analysis results. In other words, they are interested in feedbacks about a product. The second reason is that sentiment analysis is easiest process because of the diversity of emotion categories. In another saying, emotion analysis shows itself as a more detailed method compared to sentiment analysis. In addition, Sentiment analysis actually focuses on two main approaches. These are: machine-learning and lexicon-based approaches (Taboada et al., 2011). The first of these approaches, the machine learning approach, sees the emotion analysis task as a text classification problem and it requires a training process by training the classifier on a labeled text collection (Agarwal et al., 2011), (Pang, Lee, and Vaithyanathan, 2002), (Saif, He, and Alani, 2012). Another approach, the lexicon-based approach, it requires pre-constructed emotional dictionaries to analyze individual terms. These dictionaries can be created manually or automatically. Each emotional word item given is searched in the text and the weights of each identified term are summed (Taboada et al., 2011), (Ding, Liu, and Yu, 2008). When the studies for these two methods are examined, it is observed that there are many advantages and disadvantages in both of the lexicon-based and machine learning approaches. In fact, we cannot say that one of these methods performs better than the other. Because, it depends on the many factors such as dataset, subject studied or research field. Moreover, in the machine learning approach there is no need for a dictionary and in practice it produces higher classification results than the Lexicon-Based approach in terms of accuracy. Also, in the Machine Learning approach, the classifier trained on texts in one area cannot achieve the same level of accuracy in other areas (Blinov et al., 2013), (Aue, and Gamon, 2005).

As mentioned earlier, the lexicons created for Turkish literature are very insufficient compared to English. Studies on this subject continue and there are lexicons only for sentiment analysis as follows:

Dehkharghani et al. (2015) created 'SentiTurkNet', which is the first comprehensive Turkish polarity resource in the literature. They created this by assigning three polarity scores -positivity, negativity, and objectivity- to each synsets found in Turkish WordNet (Bilgin, Çetinoğlu, and Oflazer, 2004). In addition, Turkish WordNet was created by translating the synsets in WordNet. Their method is general and suitable to other languages. They concluded that translating polar sources from another language (English) or another language (Turkish) is not the best approach. Because not all terms in a language have equivalent terms in other languages, and terms that depend on culture or language often have different polarities. Their results showed that their proposed methodology was quite successful. Using only the mapping approach provides 86.11% accuracy in classifying synsets as negative, positive or objective. However, using all the features, they achieve 91.11% accuracy.

In another study, Akbaş (2012) constructed a system for extracting aspect-based sentiment summaries on Turkish tweets. The author used a Turkish opinion word list constructed manually and proposed a word selection algorithm to automate finding new words with their sentiment strengths. Her algorithms are tested on Turkish tweet data collected over time via Twitter API. She investigated the problem of mining opinions by extracting aspects of entities/topics on collection of short texts and also, she focused on Turkish tweets that contain informal short messages. Her approach is to help develop sentiment analysis for other languages where resources as rich as the English language are not available. After a set of preprocessing stages, she grouped the tweets based on the extracted aspects. In addition to her manually constructed Turkish opinion word list, an automated generation of the words with their sentiment strengths is proposed using a word selection algorithm. Then, she offered a new representation of the text according to sentiment strength of the words, which we refer to as sentiment-based text representation. She adapted machine learning methods to generate classifiers based on the multivariant scale feature vectors to detect mixture of positive and negative sentiments and to test their performance on Turkish tweets. To build sentiment summaries, the aspects are extracted, and the tweets are grouped according to these aspects. Manually extracted aspect list are expanded with algorithms to assigned tweets to groups. After clustering the tweets, their opinion polarity is determined as their sentiment strength. The performance evaluation illustrates significant improvements over the methods adapted from the literature.

Kaynar et al. (2016) according to the content of the movie comments, they conducted emotion-thought analysis using classification algorithms such as Naive Bayes, Center Based Classifier, Multilayer Detection and Support Vector Machines. The data set concentrates on distinguishing positive and negative emotions and consists of a total of 2000 data, 1000 in the positive class and 1000 in the negative class. When the performance criteria were examined, the highest values were obtained in Artificial Neural Networks and Support vector Machines. In the training dataset, Artificial Neural Networks performed better than Support vector Machines (84.07%) with an accurate classification rate of 89.73%, while both classifiers showed almost the same performance with the correct classification rate of around 75% for the test data set.

Uçan et al. (2016) proposed an automatic translation approach to create a sentiment lexicon for a new language from available English resources. In this approach, an automatic mapping is generated from a sense-level resource to a word- level by applying a triple unification process. This process produces a single polarity score for each term by incorporating all sense polarities. The major idea is to deal with the sense ambiguity during the lexicon transfer and provide a general sentiment lexicon for languages like Turkish, which do not have a freely available machine-readable lexicon. Actually, the objective is to automatically create a sentiment lexicon in a short time, which can work as much as the supervised methods and also ones that are created manually. According to this, three sentiment lexicons, named ‘TSDs’, ‘TSDp’ and ‘TSDh’, have been proposed for the Turkish language from SentiWord-Net. These lexicons have been achieved by an automatic translation approach along with a unification process. According to their experimental results, the sentiment lexicon achieved by the parallel translation approach (TSDp) performs better than the others in the sentiment classification task along with AvgSubScores. The obtained results from the three Turkish lexicons demonstrate that the translation approach performs well over the positive terms and their SubScores are more reliable than the negative ones.

As we know, there are very few studies in Turkish focused on generating resources for emotion analysis in the literature. There are also several datasets created for this field as follows:

First dataset was generated by Demirci (2014) who focused on extracting emotion from Turkish microblog entries. She collected tweets for the six emotions, joy, fear, anger, sadness, disgust and surprise using the Twitter search mechanism for hashtags. Demirci defined hashtags containing derivatives of words for each emotion of each emotion category. As a result, Demirci managed to collect 6000 tweets in total, 1000 tweets for each emotion. In fact, her study has mainly consisted of five phases. Firstly, Since Demirci could not find the appropriate data set for the Turkish language, she collected the data from Twitter and prepared a list of words and stages in six emotional classes. In the next phase, she applied some preprocessing to dataset. In another phase, she applied Feature Vector Construction and Feature Selection. At the last stage, Demirci applied Naive Bayes, Support Vector Machine, Complement Naive Bayes, and K-Nearest Neighbor controlled supervised machine learning algorithms to classify the data. According to her research results, the SVM performed better than the others, achieving a classification accuracy of 69.92%.

Boynukalın (2012), generated two datasets for analyzing four emotion categories, joy, sadness, anger, and fear. Actually, her system mainly consisted of five stages. In the first stage, involves translating ISEAR dataset (Geneva, 2013) to Turkish, collecting data from Turkish fairy tales from several websites, annotating fairy tales' data and combining two datasets. As the second stage, she applied Preprocessing to the data sets put together and Feature selection in the other stage. Boynukalın made Feature Vector Construction in the fourth stage. In the last stage, using Weka (2014), she made classification used Complement Naive Bayes, Support Vector Machine and

Naive Bayes. According to her results, Complement Naive Bayes gave the best results which obtained the accuracy values of 81.34% for the ISEAR dataset with four classes. Also, 76.83% for the Turkish fairy tales with five classes. Moreover, 80.39% for the combination of the two datasets with four classes.

As mentioned before, there are many studies and resources related to English language compared to Turkish. These studies can be easily accessed and used. In addition to developing for lexicons for sentiment analysis in English, there are also available lexicons for emotion analysis as follows:

Strapparava and Valitutti (2004), present a linguistic resource for a lexical representation of affective knowledge. This resource was developed starting from WORDNET, through the selection and labeling of the synsets representing affective concepts. Strapparava and Valitutti created a lexicon called ‘WAL’ for use in their study. The lexicon called WAL created by them is the shorted form of WordNet-affect lexicon. The WAL lexicon was created by using several hundred core words tagged by specific 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 suitable core word. In the resulting lexicon, 1536 words are linked to Ekman’s (1992) six emotion categories (joy, sadness, anger, fear, disgust, and surprise).

Mohammad and Turney (2012) described a project aimed at creating a large lexicon of term-emotion associations. In this study, Mohammad and Turney show how they compiled a large English term-emotion association lexicon by manual annotation through Amazon’s Mechanical Turk service. They created a new lexicon named ‘EmoLex’. This lexicon contains 14,182 words in total. Also, this dataset is an order of magnitude larger than the WordNet Affect Lexicon. EmoLex is focus on the eight emotions types (such as joy, sadness, anger, fear, trust, disgust, surprise, and anticipation) (Mohammad, 2010). The terms in EmoLex have been carefully selected to include some of the most frequent English verbs, nouns, adjectives and adverbs. In their approach, first step is, four different words given and asked which word is closest in meaning to the target. Three of the four options are unrelated distractors. The remaining option is a synonym for one of the senses of the target word. Through this single question they convey the word sense for which annotations are to be provided, without actually providing annotators with long definitions. Further, if an annotator is not familiar with the target word and still attempts to answer questions pertaining to the target, or is randomly clicking options in their questionnaire, then there is a 75% chance that they will get the answer to this question wrong, and can discard all responses pertaining to this target term by the annotator.

In another study of Mohammad and Kiritchenko (2015), an emotion dataset named hashtag emotion corpus was created by using the hashtag structure in Twitter. Here, the process of emotion identification is based on the names of the hashtags. In this paper, they showed that emotion-word hashtags are good manual labels of emotions in tweets. They also proposed a method to generate a large lexicon of word labeled tweet corpus. This is the first lexicon with real-valued word–emotion association scores. In the first step, this dataset started with six emotions (such as joy, sadness, anger, fear, disgust, and surprise). Also, they defined to this dataset as the 'Hashtag Emotion Corpus'. After, it became a structure that covers 585 emotions. In the second step, Mohammad and Kiritchenko created a lexicon named 'hashtag emotion lexicon' from the emotion corpus dataset. In addition to word valued score indicating the degree of association between the word and the emotion. Higher scores indicate greater association. They first showed how the word–emotion lexicon helps in emotion detection, and then use it to improve personality detection from text. This experiment was especially telling since it showed that self-labeled emotion hashtags correspond well with annotations of trained human judges.

Mohammad (2012) created a large dataset from Twitter called 'Twitter Emotion Corpus', collecting more than 20,000 emotion-tagged tweets with hashtags. He gathered together Ekman's (1992) six emotions (joy, sadness, anger, fear, disgust, and surprise) argued to be the most basic. His aim is in this paper is to determine if he can successfully use emotion-word hashtags as emotion labels. According to Mohammad, *'hashtags serve many purposes, but most notably they are used to indicate the topic'*. To create 'Twitter Emotion Corpus' firstly, he collected tweets with hashtags corresponding to the six Ekman emotions (such as #anger, #disgust, #fear, #happy, #sadness, and #surprise). 'Twitter Emotion Corpus' (TEC) has about 21,000 tweets from about 19,000 different people. In addition, he created binary classifiers for each of the six emotions using Weka (2014). After, Mohammad chose two machine learning algorithms. These are, Support Vector Machines (SVM) and Sequential Minimal

Optimization (SMO) because of its successful application in various research problems. Then, he created a new emotion-based lexicon called 'Twitter Emotion Corpus' (TEC), containing 11,418 words types, by using the Twitter emotion corpus dataset. Finally, he showed that the lexicon leads to significantly better results than that achieved using the manually crafted 'WordNet Affect' lexicon in an emotion classification task. Some promising studies for Turkish, although slightly less than studies for the English language, are as follows:

Toçoğlu et al. (2019) compared the classification results of different machine learning algorithms on the TREMO (Toçoğlu, and Alpkoçak, 2005) data set used in the field of emotion extraction from Turkish texts. They are dealt with emotion analysis as a text classification problem. In addition, they studied four approaches: Artificial Neural Networks, Support Vector Machines, Random Forest, and K-Nearest Neighbor algorithms. The categories of happiness, fear, anger, sadness, disgust and surprise, which are the six emotions provided by the dataset, were used as emotion categories. After preprocessing stage for the dataset, the words are stemmed, and the data set is modeled with the Vector Space Model. Moreover, the most important first 500 words for each emotion were determined by Mutual Information method. According to their experimental results, The Artificial Neural Networks algorithm gave the best results. Support Vector Machines, Random Forest and K-Nearest Neighbor algorithms, on the other hand, showed reduced performance in this order.

Toçoğlu and Alpkoçak (2019) proposed a lexicon named "Turkish Emotion Lexicon" for analysis of Turkish emotions in six emotional categories (happiness, fear, anger, sadness, disgust and surprise) in their study. They also explored the effects of a lemmatizer and a stemmer, two term weighting schemes, four dictionary enrichment methods, and a term choosing approach for dictionary building. To do this, they created a Turkish emotion lexicon from the TREMO (Toçoğlu, and Alpkoçak, 2005) dataset, which contains 25,989 documents. After the preprocessing stages, they proposed two different weighting schemes for six emotion categories in which term frequency, term class frequency and mutual knowledge values were taken into account. They then enriched the lexicon using the bigram and concept hierarchy methods and made a term selection for productivity problems. Next, they compared the performance of the lexicon-based approach with the Support Vector Machine, Random Forest, and Naive Bayes supervised machine learning-based approach using the proposed lexicon. When the results of the experiments were examined, it was shown that the efficient use of the suggested lexicon yielded comparable results in the sentiment analysis in the Turkish text.

Actually, so many studies are focus on developing lexicons for emotion and sentiment analysis in English. Some studies based on developing a lexicon for the former analysis as follows:

WordNet (Fellbaum, 1998) is a popular English electronic dataset containing data for adverbs, adjectives, noun and verbs. The data in WordNet is organized as synonym sets which are called synsets (senses). Each synset is composed of synonymous words referring a common semantic meaning. This dataset used in many studies in literature.

Nielsen (2011) generated a new lexicon that can be used for the sentiment analysis. He made use of Twitter data in the creation of his lexicon. For the values of polarity, he gave negative values between -1 and -5 to the words representing the negative information and gave positive values between 1 and 5 to the words representing the positive information.

Baccianella et al. (2010) created 'SentiWordNet 3.0'. SentiWordNet 3.0 is a lexical resource explicitly devised for supporting opinion mining and sentiment classification applications. SentiWordNet 3.0 is an improved version of SentiWordNet 1.0. Also, it is a lexical resource publicly available for research aims, now currently licensed to more than 300 research groups and used in a variety of research projects worldwide. It provides an annotation based on three numerical sentiment scores for each WordNet synset. Given that this lexical resource ensures a synset-based sentiment representation, different senses of the same term may have different sentiment scores. According to their results, SentiWordNet 3.0 is essentially more accurate than SentiWordNet 1.0, with a 21.96% improvement for the ranking by negativity and a 19.48% relative improvement for the ranking by positivity.

In another example, Wiebe et al. (2005) created a lexicon named MPQA. This lexicon contains 8222 terms in total. Each term in the lexicon is labeled with polarity values (positive, negative, and neutral). In addition, each of them has their intensity values that are weak and strong.

Thelwall et al. (2010) developed a library named ‘SentiStrength’ for lexicon-based sentiment analysis in English. SentiStrength library produces sentiment scores for each word of a given text. To do so, it uses several word lists which are sentimental, idiom, booster, negation, and emoticon lists. They applied to 'MySpace' comments and with a lookup table of term sentiment strengths optimised by machine learning, SentiStrength is able to predict negative emotion with 72.8% accuracy and positive emotion with 60.6% accuracy, both based upon strength scales of 1-5.

# CHAPTER 4 : METHODOLOGY

In this thesis, we followed a lexicon-based approach where different word and sentence representations are employed to detect the emotions in sentences. In our experiments, we run both unsupervised and supervised methods assuming that each sentence must be assigned to one of 6 emotion categories (Happy, Anger, Fear, Sadness, Disgust, Surprise).

Firstly, three lexicon sets named as "Lexicon1", "Lexicon2" and "Lexicon3" are built. The procedure of lexicon construction and regarding statistical information on sets are given below:

- ***Lexicon1:*** A survey was conducted with five university students (18-25 years old) to create the Lexicon1. Two words were given to the students for each of the six emotions (Happiness, Fear, Anger, Sadness, Disgust and Surprise). The couples of pivot words and emotion are given in Appendix A4.

  Following, it was voted which of the given words expresses that emotion better. As a result of this survey, one pivot word for six emotions was determined based on the majority of votes [A4]. Then, for each emotion, ten emotion words closest to the pivot word were automatically selected. The CBOW word vectors and cosine similarity method are employed to decide the closest words. The details on building word vectors will be given in following sections. Final Lexicon1 consists of a total of 11 words for each of 6 emotions.

- ***Lexicon2:*** A survey was conducted with five university students (18-25 years old). They were asked to choose the ten emotion words closest to the pivot word from the nearest twenty words to the pivot word [A5]. Similar to Lexicon1 similarity is defined as the cosine similarity of word vectors. As a result of the survey, the majority of votes were created [A6] pivot for each emotion and the list of the 10 closest emotion words [A7]. In addition, this set consists of 66 words in total, similar to Lexicon1.

- ***Lexicon3:*** This set is the one proposed by Toçoğlu and Alpkoçak's article in (2019). It contains the emotion words together with their weights. The weights are stated to be measure mutual information (MI) values of emotion words in (Toçoğlu, and Alpkoçak, 2019) Lexicon3 contains 1320 words together with their weights [A8].

The lexicons are employed both in unsupervised and supervised approaches. We followed up an unsupervised approach that will be named as keyword spotting from now on. In keyword spotting, the words in given sentence is compared to the lexicon words in two folds. In first group of keyword spotting experiments the exact matches are considered. Secondly, cosine distances of sentence words to lexicon words are measured and employed in emotion identification task. The details on keyword spotting will be given in section 4.3.

In supervised approach, the sentence words are compared to lexicon words similar to keyword spotting and the comparison results are recorded as vectors. These vectors are given as input to supervised machine learning models such as Naïve Bayes, SMO and J.48.

Following, datasets that are utilized in experiments (Section 3.1), the approach to build word vectors (Section 3.2), keyword spotting approach (Section 3.3), supervised learning methods (Section 3.4) will be given in detail.

### *4.1. Datasets*

In this thesis, two datasets are employed in the experiments. The first is TREMO dataset (Toçoğlu, and Alpkoçak, 2005) that involves emotion labeled sentences. Following preprocessing, the set is both used in unsupervised and supervised learning experiments. The second dataset, Wikipedia (2018), is utilized to construct word vectors/word representations. Both data resources are subjected to required preprocessing operations to obtain the computable inputs to the experiments. In below sections details on data sets and regarding preprocessing operations are given.

### *4.1.1. TREMO Dataset*

TREMO (Toçoğlu, and Alpkoçak, 2005), (Toçoğlu, and Alpkoçak, 2019) dataset is built by a survey that was conducted with 5,000 participants. The set involves 6 basic emotions. These emotions are determined as Happiness, Fear, Anger, Sadness, Disgust and Surprise. The participants of the survey were asked to share their memories and experiences as text for 6 emotional categories. At the end of this process, the documents of 4,709 participants were approved. In addition, 27.350 documents were collected from the participants, showing distribution according to the emotions in Table 4.1.

Table 4.1. Distribution of documents by emotion categories (Source: Toçoğlu et al., 2019)

| | Happiness | Fear | Anger | Sadness | Disgust | Surprise | Total |
|---|---|---|---|---|---|---|---|
| Number of sentences | 4.700 | 4.616 | 4.636 | 4.664 | 4.522 | 4.212 | 27.350 |

The Tremo data set is a diverse set of data, as it consists of people's current reactions or past events. Due to the large number of participants and the inputs in the form of text, a verification process has been implemented for this data set. In order to reduce the impact of negative documents during the verification process, each document was presented to at least 3 and at most 5 different users, and the emotion category of the document was decided by unanimous or majority vote. 48 volunteer participants participated in this process and these participants gave a total of 92.986 votes for all documents.

When the sentences in the data set are examined, it is observed that there are different situations for some sentences. Many emotions can occur at the same time in daily life or in some events that occur in our memories of the past, and the emotion of this event may not be decided clearly.

For example, it can contain several emotions at once in the sentence *''hayatın acımasızlığına hayret ediyorum''* in the data set. This sentence may contain feelings of anger, sadness, or surprise meaning there may not be a single emotion. Or the sentence *‘‘en yakın arkadaşımdan bu hareketi hiç beklemezdim’’* in the dataset is ambiguous in terms of emotion. Perhaps a participant that wrote this sentence may have felt happiness as a result of any situation or felt anger in a bad situation. Sentences that do not have a precise emotion like these sentences are marked. In order to avoid any confusion in the future studies, these sentences have been labeled 'Ambigious'.

On the other hand, the sentences like *‘‘korku nedir bilmem’’* or *‘‘trafikte bir kadının araba sürdüğünü görmem’’* are in the TREMO dataset unlike other examples, instead of having a few emotions, sentences that do not express any emotion. In order to prevent such sentiments in the data set from being confused, 'Ambigious' labels were given.

A total of 1,361 (4.98% of the data set) documents, whose emotion category was ambigious, was removed from the system as a result of voting and the data set was obtained. The distribution of this data was as shown in Table 4.2. The dataset contains a sample sentence for the most 'Happiness' feeling, while there is a minimum of 'Surprise' for the sentence.

Table 4.2. Distribution of documents according to emotion classes after verification (Source: Toçoğlu et al., 2019)

| Emotion | Number of Original Documents | Number of Documents After Verification |
|---|---|---|
| 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** |

A staged procedure is applied in preprocessing of TREMO dataset. Firstly, the data set was printed line by line. All the letters are converted to lower case; punctuation marks, numeric characters and extra spaces were removed. Following, the number of unique and total words in the dataset was found. As the final stage, the data were kept in the form of list of lists to match the Word2Vec format. This form we get is called *Surface Form*. As the next step, Porter Stemmer (Snowball Stemmer (2002)) was applied to find the stems of the words in the generated surface form and the stems of the words were found. The results obtained were called *Stemmed Form*. After all, we removed stop words [A1], stemmed stop words [A2] and single letters from Stemmed Form. Finally, the words in TREMO dataset is searched in the Wikipedia word vectors, matching vectors are found. The labeled sentences together with word vectors are named as TREMO model. Below in Figure 4.1, this preprocessing procedure is given.

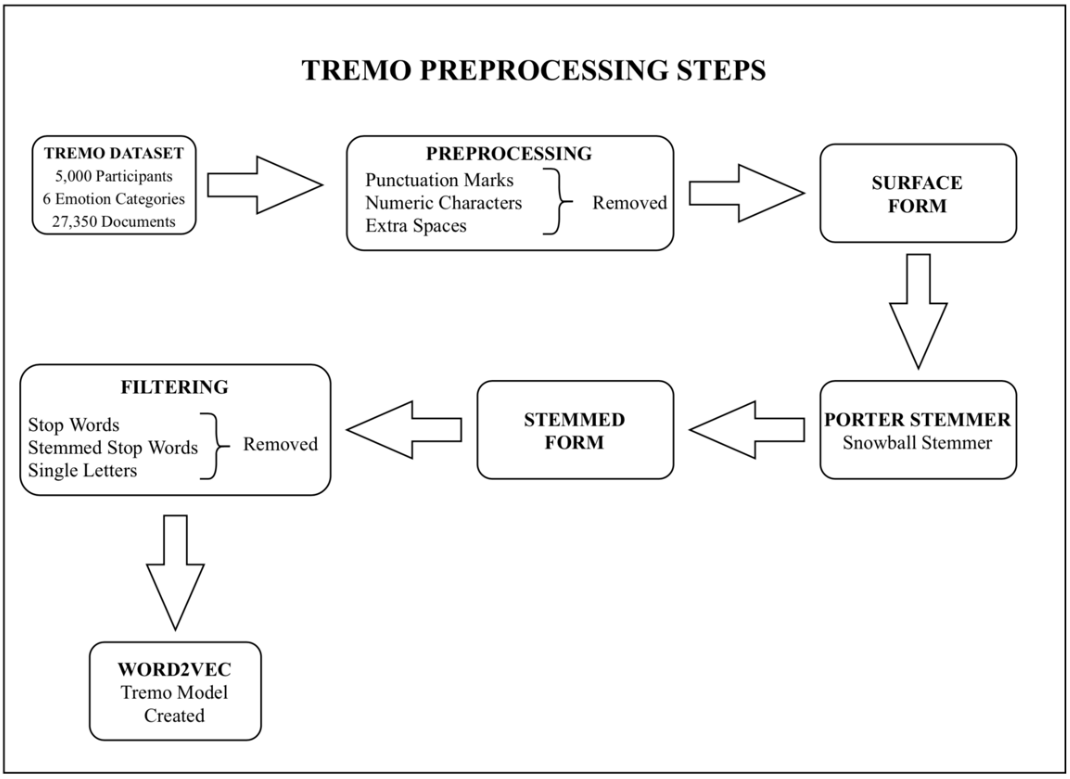



Figure 4.1. TREMO processing steps

### *4.1.2. WIKIPEDIA Dataset*

In this study, Wikipedia was chosen to build word vectors (Wikipedia, 2018). Wikipedia contains 4,184,516 sentences. The reason we use a large data set such as Wikipedia is that it contains a wide variety of words and sentences. This diversity will make a significant contribution in the future stages of experiments.

Wikipedia is subjected to the same preprocessing steps as the TREMO dataset (Toçoğlu, and Alpkoçak, 2005) as given in Figure 4.2. Briefly, punctuation marks, numerical characters, extra spaces and non-Turkish characters [A3] were deleted. All the letters were converted to lower case. This form of dataset was named as *Surface Form*. Then Porter stemmer (2002) was applied and *Stemmed form* was obtained. Stop words [A1], stemmed stop words [A2] and single letters have been deleted from this form. Finally, Word2Vec was to obtain word vectors, these word vectors are used as vectors/ word representations for matching words in TREMO.

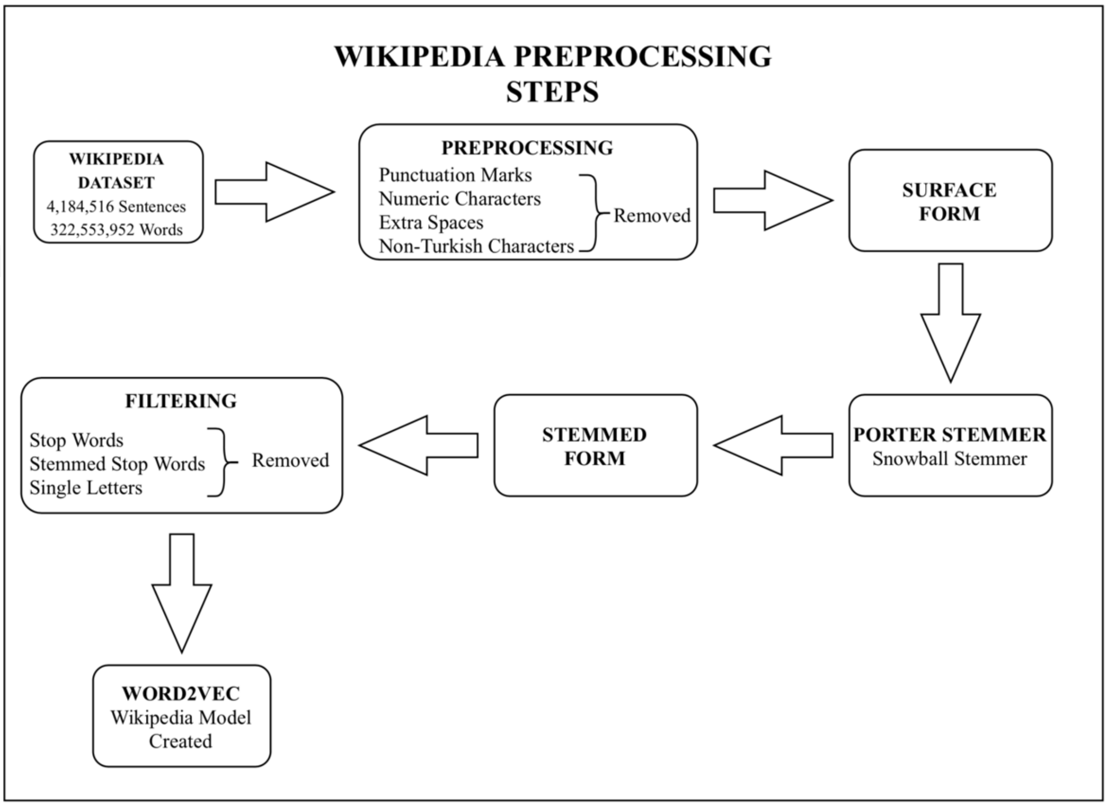



Figure 4.2. Wikipedia processing steps

### *4.2. Word Vectors*

In this thesis, to represent the word as a vector of numbers we employed the well-known and widely used word2vec method.

Word2Vec (Mikolov et al., 2013) is a prediction-based and unsupervised (no labels) model that tries to express words in vector space. For Word2Vec method two models are presented: Skip-Gram and CBOW (Continuous Bag of Words). These two models are similar in terms of their algorithms except that in the Skip-Gram architecture, the model predicts source context-words from the target words. On the other hand, in the CBOW architecture, the model does the opposite and predicts target words from source context words (Büyükkınacı, 2018).

Word2Vec model consists of three different layers named as hidden layer, input layer and output layer. In Figure 4.3, architecture of Word2Vec is exemplified basically. The model takes on a very large input vector, compressing it until it forms a smaller dense vector, giving the possibilities of the target words.

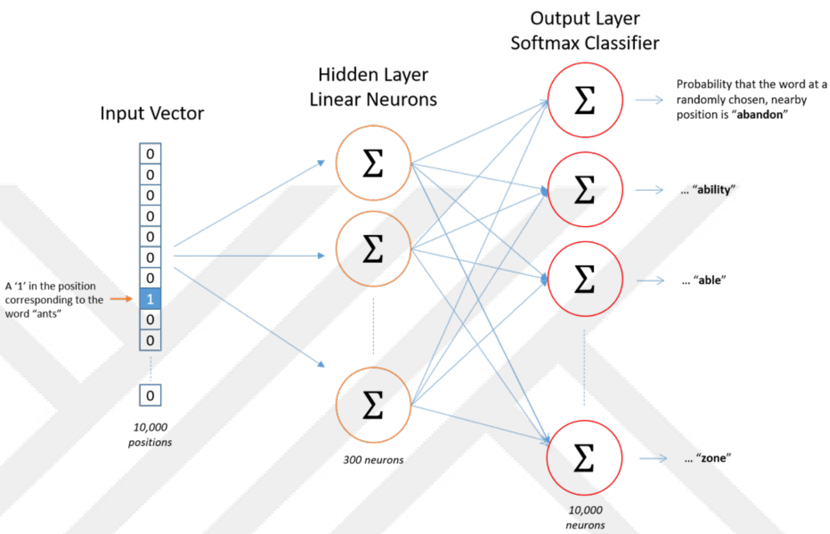


Figure 4.3. Architecture of Word2Vec (Source: Büyükkınacı, 2018)

In other saying, Word2vec's input layer is one-hot vector which has the same size with the vocabulary. In Figure 4.4 shows that to generate one-hot vector, it is filled with zeros outside the index of the input word.

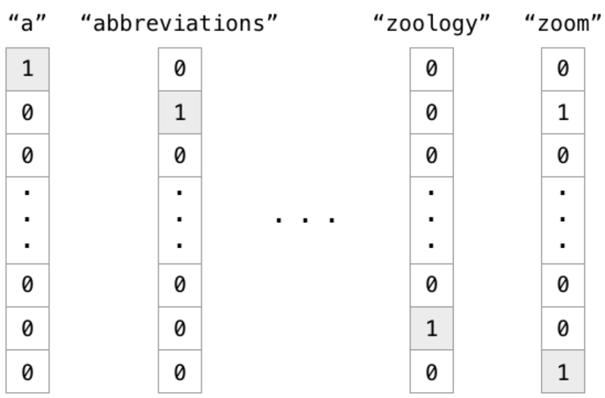


Figure 4.4. Example of One-hot encoding (Source: Büyükkınacı, 2018)

In the CBOW model, the words that are not in the center of the window size are taken as input and the words that are in the center of the window are tried to be estimated as output. This process continues until the end of the sentence. These operations applied to a sentence are applied to all sentences, so that we have mapping to the unlabeled data initially available and are ready to train. In Figure 4.5 the neural network representation is given.

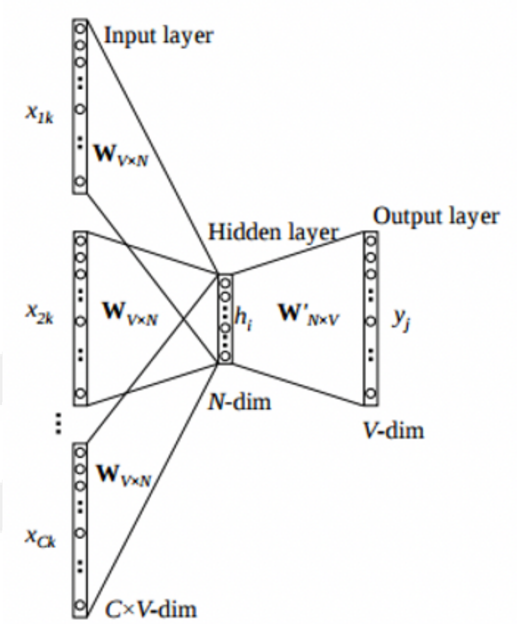



Figure 4.5. Demostration of CBOW on the neural (Source: Büyükkınacı, 2018)

To exemplify, assume that "bir yandan da hiç konuşmak istemiyor" is our sentence as given in Figure 4.6. The CBOW model, which takes this sentence as an input and has window_size = 1. If window size = 1 then it means that each target word will be predicted by preceding and following word. This model works as follows: First, it puts the word “bir” in the center of the window, then it is predicted by its following word "yandan". Then the algorithm shifts to second word of sentence and the words "bir" and "da" inputs to the neural network to predict the second word "yandan" and it continuous till the end of the sentence.

| | CBOW WORD2VEC WITH WINDOW_SİZE =1 | | | | | | Training Samples | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | Input | Output |
| 1) | bir | yandan | da | hiç | konuşmak | istemiyor | yandan | bir |
| 2) | bir | yandan | da | hiç | konuşmak | istemiyor | bir | yandan |
| | | | | | | | da | yandan |
| 3) | bir | yandan | da | hiç | konuşmak | istemiyor | yandan | da |
| | | | | | | | hiç | da |
| 4) | bir | yandan | da | hiç | konuşmak | istemiyor | da | hiç |
| | | | | | | | konuşmak | hiç |
| 5) | bir | yandan | da | hiç | konuşmak | istemiyor | hiç | konuşmak |
| | | | | | | | istemiyor | konuşmak |
| 6) | bir | yandan | da | hiç | konuşmak | istemiyor | konuşmak | istemiyor |

Figure 4.6. CBOW Word2Vec model. green represents the input, orange represents the output (Source: Büyükkınacı, 2018)

In the Skip-Gram model, the word in the center is taken as input and the words that are not in the center are tried to be predicted as output. It is largely the same with CBoW except that the input one hot vector which is center word. In Figure 4.7 the neural network representation is given.

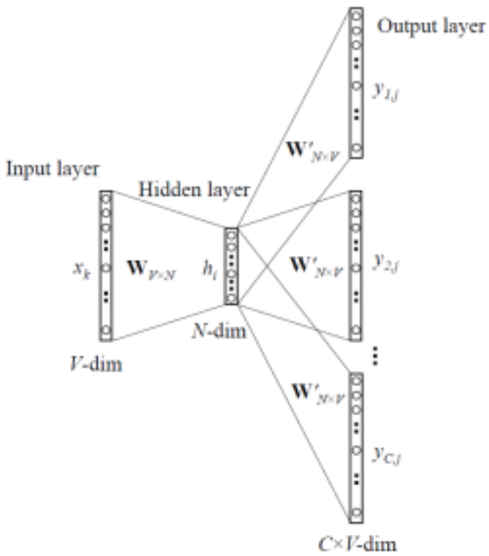



Figure 4.7. Demostration of Skip-gram on the neural network (Source: Büyükkınacı, 2018)

The Skip-gram model of the same sentence as given Figure 4.8 with window_size = 1 works like this: Briefly, the algorithm takes the center word as an input and tries to predict context words. Firstly, the word "bir" is the center word and try to predict the word "yandan" by using center word. Then, it shifts the center word to right. Now, the center word is "yandan". It tries to predict "bir" and "da" by using "yandan" and it continues like this until the end of sentence.

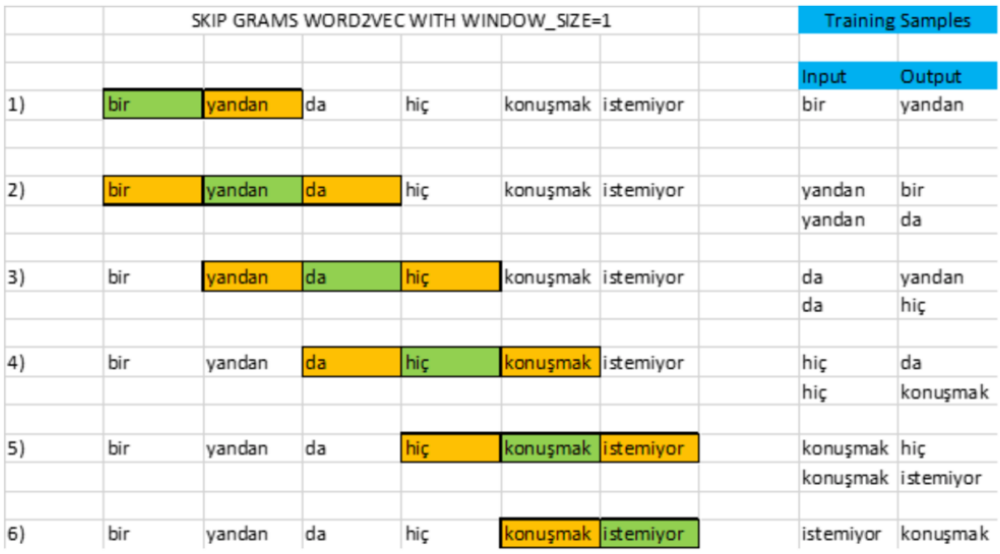



Figure 4.8. Skip-gram model. green represents the input, orange represents the output (Source: Büyükkınacı, 2018)

The main differences between CBOW and Skip-gram are

1. While CBOW models generally work better on small datasets, Skip-gram works better on large datasets.
2. CBOW requires less computation power, while Skip-Gram requires more computation power.
3. While CBOW is not good at understanding 2 or more meaningful words, Skip-Gram can learn 2 or more meaningful words better.

### *4.3. Keyword Spotting*

In keyword spotting method, lexicon words that are labeled with one of the 6 emotions are considered as keywords. Each word in given sentence is compared to all the keywords in lexicon. The comparison operation may be a basic string-matching task or it may be a similarity check where complex operations are required. Following the comparison operation, an emotion value for each individual emotion is obtained for the sentence. In other words, an emotion vector of 6 values represents the sentence where each value refers to a specific emotion. Finally, sentence is classified to the emotion category that holds the highest emotion value in the vector.

In our study, the *emotion vectors* are built in 4 different ways based on the comparison procedure and the strategy to obtain emotion values. These are *tf*, *MI_weighted-tf, max-similarity* and *average-similarity* methods.

In our experiments, *tf* method is tested by *Lexicon1* and *Lexicon2* and *Lexicon3. MI_weighted_similarity* method is applied using *Lexicon3. Max_similarity and average_similarity methods are tested using Lexicon1, Lexicon 2 and Lexicon 3.* In experiments, the sentence is classified to the emotion category that holds the highest value. In case where there exists (at most) 2 equal results, the sentence is accepted to be classified to both emotions and if one of them is the true class the result is accepted to be a hit (true positive).

#### *4.3.1. TF Method*

In *tf* method, the total number of matches to lexicon is considered. Simply, in *tf,* the high number of matches to the emotion words in lexicon to a specific emotion is considered as a strong indicator for the sentence to be assigned to the regarding emotion category. While *tf method* is applied the sentences that includes no matches are omitted.

In experiments, the sentence is classified to the emotion category that holds the highest value. In case where there exists (at most) 2 equal results, the sentence is accepted to be classified to both emotions and if one of them is the true class the result is accepted to be a hit (true positive).

In this method, according to the frequency of using these lexicon words in Tremo sentences, a matrix was created for each word and this was done for six emotions. Then the resulting matrices were combined and a large matrix was created. If any sentence in the Tremo dataset contains pivot or emotion words, the table was created by determining the amount of these words. For example, the phrase *'Fenerbahçe bu sene şampiyon olduğunda çok mutluluk duygusu hissedeceğim'* in Figure 4.9 changed to *'fenerbahçe şampiyo mutluluk duygus hissedecek'* after the preprocessings steps. Since the total number of *'mutluluk'* and *'duygus'* words in the sentence is 1 and any of the other emotion words are not included in the sentence, their number is written to the matrix as 0. The matrix of this sentence created for the feeling of Happiness [1,0,0,1,0,0,0,0,0,0,0]. All counting operations were applied for 21,558 Tremo sentences, six emotions and three lexicons. As a resulting matrix values were summed up and for each sentence, a six-digit matrix is obtained. However, in some sentences, no pivot or emotion words were found. Matrices of such sentences are expressed as [0,0,0,0,0,0]. These matrices were eliminated because they did not give us any information.

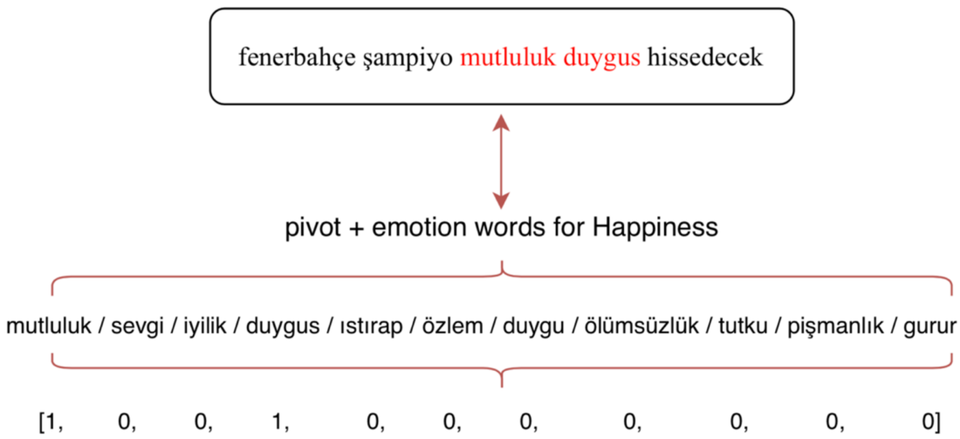



Figure 4.9. TF method matrix example

### 4.3.2. *MI_weighted tf*

In *MI-weighted-tf*, matches to lexicon words are obtained for each sentence as in tf-method. But in this method, MI values given in (Toçoğlu, and Alpkoçak, 2019) for matching words are summed up to obtain emotion values. Similar to *tf* method, sentences lack of lexicon matches are ignored during the experiments.

### 4.3.3. *Max similarity*

In *max_similarity* method, the cosine similarity of each word in sentence to each lexicon word is calculated employing CBOW vectors. For each emotion, the most similar lexicon word is determined for each word in sentence and the similarity value is recorded as maximum similarity of the word. Following, these maximum similarity values are averaged to obtain sentence emotion value.

#### *4.3.4. Average Similarity*

In *average_similarity* method, similar to *max_similarity* method, the cosine similarity to each lexicon word is calculated for each word in sentence. For each emotion, the average of these similarity values is assigned as emotion value to the regarding emotion.

### *4.4. Supervised Learning*

In supervised learning, emotion detection is accepted to be a classification task where given sentence is to be assigned to one of the six emotions. The vectors built in keyword spotting approaches are given as inputs to different classifiers. We employed WEKA in supervised learning experiments. Weka (2014) is the software developed at Waikato University for machine learning and consisting of the initials of the words "Waikato Environment for Knowledge Analysis". It contains many machine learning algorithms and methods that are widely used today. Thanks to its development in Java language and its libraries coming in .jar files, it can be easily integrated into projects written in Java language, making its use even more widespread (Şeker, 2013).

In supervised learning approach, commonly two data sets are employed in experiments. The first is known to be the training set. In training set, the samples together with their labels must be provided. The second set that is used to measure the performance of supervised approach is the testing set. In testing set, the samples are given to the learning machine without their labels. Following the testing stage, the supervised machine generates the labels to unlabeled sample based on the experience and information it obtained from the training set.

Supervised learning is good for classification problems such as determining which category a product belongs to or predicting the company's sales volume for a future date. All these type of problems are known as classification problems. In classification problems, several attributes are determined as indicators. For each sample, the values of attributes are measured and this set of values is given to the classifier. The supervised learning machine (classifier) simply tries to find the class that the sample belongs to based on attribute values. There exist many different algorithms employed while the attribute values are considered. These algorithms provides different types of classifiers such the Naïve Bayes that employs simple Bayes method, decision trees and neural networks.

In this thesis, we employed the sentence vectors as inputs to classifiers and split the data set in two as training and testing sets. In order to compare the performances of different classifiers in emotion detection, we employed Naïve Bayes, BayesNet, SMO, Random Forest and J48 methods in Weka. In Appendix A9 definitions of regarding classifiers are provided.

# CHAPTER 5 : EXPERIMENTAL RESULTS

In our experiments, punctuation marks numeric characters and extras spaces are removed, stemming is applied and a stop word filtering process is executed before the application of key word spotting and supervised learning methods. after preprocessing data sets, word vectors for sentence in TREMO data set is built employing Wikipedia. Following the stop word filtering, Tremo sentences consisting of one or two words are also omitted. Likewise, it is observed that for some of the words in TREMO sentences, there exists no available vector that can be obtained from Wikipedia. Such sentences that do not have at least 3 words that have valid vectors are ignored. After this filtering process, a total of 21,558 Tremo sentences remained.

In this section experimental results of keyword spotting approach and supervised methods will be given respectively.

## *5.1. Keyword Spotting Experiments*

Keyword spotting experiments are performed employing tf, MI-weighted-tf, max-similarity and average similarity methods respectively.

Tf method is applied employing Lexicon1, 2 and 3. The sentences that do not involve any lexicon words are not considered in tf experiment. The final data sets include 4474, 3583 and 15180 sentences respectively when Lexicon1, 2 and 3 are used. In table 5.1, 5.2 and 5.3, the numbers of sentences for the six emotions are given for Lexicon1 (66 words), 2 and 3 respectively. In these tables the values in rows refer to the number of sentences that are assigned by the method to the emotions in columns. For example, in table 5.1, 159 of happy sentences are classified as happy, 4 of happy sentences are classified as fear, 2 of them are classified as anger by tf method.

Moreover, in Table 5.1, 5.2 and 5.3; the cells given shaded refers to the highest number of sentences that belong to the emotion category given in the regarding row. As a result, for example in table 5.1, it is seen that most of the happy sentences are labeled as happy by the tf method. On the other hand, in same table, fear sentences are mostly classified to disgust category.

Table 5.4 involves the performance of tf method in terms of (detailed) accuracy values for all sets for all emotions in experiments. Accuracy is measured as the ratio of true positives (correctly classified sentences) to all sentences in the experiment.

Table 5.1. Keyword spotting -TF method Lexicon1 results

| | **TF METHOD → *Lexicon1*** | | | | | |
|---|---|---|---|---|---|---|
| **Observed Emotion →** | **HAPPY** | **FEAR** | **ANGER** | **SADNESS** | **DISGUST** | **SURPRISE** |
| **HAPPY** | 159 | 4 | 2 | 20 | 4 | 49 |
| **FEAR** | 5 | 289 | 1 | 2 | 647 | 380 |
| **ANGER** | 19 | 48 | 50 | 584 | 21 | 103 |
| **SADNESS** | 30 | 12 | 31 | 1011 | 2 | 52 |
| **DISGUST** | 9 | 4 | 3 | 2 | 49 | 7 |
| **SURPRISE** | 13 | 2 | 35 | 6 | 3 | 1058 |
| **AMBIGIOUS** | 5 | 3 | 0 | 8 | 3 | 10 |

Table 5.2. Keyword spotting -TF method Lexicon2 results

| | **TF METHOD → *Lexicon2*** | | | | | |
|---|---|---|---|---|---|---|
| **Observed Emotion →** | **HAPPY** | **FEAR** | **ANGER** | **SADNESS** | **DISGUST** | **SURPRISE** |
| **HAPPY** | 170 | 4 | 2 | 2 | 38 | 61 |
| **FEAR** | 13 | 294 | 1 | 6 | 12 | 660 |
| **ANGER** | 18 | 2 | 52 | 6 | 35 | 36 |
| **SADNESS** | 32 | 14 | 11 | 1011 | 14 | 34 |
| **DISGUST** | 6 | 3 | 3 | 3 | 64 | 8 |
| **SURPRISE** | 12 | 2 | 2 | 4 | 16 | 1061 |
| **AMBIGIOUS** | 5 | 5 | 1 | 8 | 6 | 11 |

Table 5.3. Keyword spotting -TF method Lexicon3 results

| | **TF METHOD → *Lexicon3*** | | | | | |
|---|---|---|---|---|---|---|
| **Observed Emotion →** | **HAPPY** | **FEAR** | **ANGER** | **SADNESS** | **DISGUST** | **SURPRISE** |
| **HAPPY** | 2694 | 286 | 257 | 175 | 163 | 1439 |
| **FEAR** | 661 | 1203 | 238 | 325 | 342 | 270 |
| **ANGER** | 389 | 287 | 1670 | 259 | 320 | 320 |
| **SADNESS** | 762 | 711 | 437 | 1197 | 215 | 777 |
| **DISGUST** | 291 | 264 | 288 | 70 | 1327 | 101 |
| **SURPRISE** | 1028 | 275 | 243 | 116 | 168 | 1477 |
| **AMBIGIOUS** | 216 | 102 | 135 | 75 | 93 | 143 |

Table 5.4. Tf- method accuracy table

| | **Lexicon1** | | **Lexicon2** | | **Lexicon3** | |
|---|---|---|---|---|---|---|
| **Emotion** | **Accuracy** | **Count** | **Accuracy** | **Count** | **Accuracy** | **Count** |
| **Happy** | 0,697 | 228 | 0,726 | 234 | 0,813 | 3312 |
| **Fear** | 0,230 | 1256 | 0,316 | 930 | 0,556 | 2164 |
| **Anger** | 0,073 | 686 | 0,391 | 133 | 0,675 | 2475 |
| **Sadness** | 0,913 | 1107 | 0,932 | 1085 | 0,432 | 2772 |
| **Disgust** | 0,731 | 67 | 0,771 | 83 | 0,747 | 1776 |
| **Surprise** | 0,957 | 1105 | 0,976 | 1087 | 0,683 | 2163 |
| **TOTAL** | | **4474** | | **3583** | | **15180** |

Considering the results in Table 5.4, it may be stated that

1. tf method succeeds in happy, surprise and disgust sentences (Accuracy range [0,683- 0,976]) for all lexicons.
2. Though Lexicon1 and 2 succeed, Lexicon3 does not in emotion sadness
3. Lexicon1 and Lexicon2 fail in emotions fear and anger. On the other hand, Lexicon3 fails in sadness sentences.

When we look at the accuracy results for Lexicon 1 and 2 in Table 5.4, we can see that some emotions have very high accuracy and some emotions have very low accuracy. This is because the emotion words in Lexicon 1 and 2, which we actually aim to generate automatically, are based on the automatic decision making, not the similarity with the corresponding emotion.

For example, as seen in Table 5.4, the emotion with the lowest accuracy value for Lexicon 1 is 'Anger' (0.073). When the emotion words that were automatically decided for Lexicon 1 were examined, it was observed that a few words that were not very close meaning to Anger emotion were actually selected (such as *'umutsuzluk', 'şaşkınlık'* and *'üzüntü'*). To give another example, the emotion with the second lowest accuracy value is 'Fear' (0.230). When the emotion words determined automatically for fear were examined, it was observed that words such as *'umutsuzluk', 'öfke'* or *'duygusallık'* do not actually belong to the emotion of fear.

Emotions mixed with each other were thought to be the main reason for such low accuracy rates. The reason for this is that emotion words that are determined automatically include other emotions. In other words, the reason why the automatically created dictionaries do not have a balance of accuracy is that they are not created in a controlled manner. If a controlled lexicon were created, only the words with the closest meaning with the relevant emotion could be selected and higher accuracy rates could be obtained.

The emotion words *'korkak'* or *'korkar'* in the feeling of 'Disgust' cause the emotions of the sentences to be made wrongly in machine learning. Therefore, emotions that are actually labeled with Fear emotion are assigned to Disgust emotion as a result of wrong decision making, as seen in Table 5.1.

When overall accuracy in Table 5.5 was examined, it was observed that Lexicon2, which consists of 66 words, has the highest accuracy value. Next comes Lexicon3, which consists of 1320 words (Happy: 250, Fear: 214, Anger: 261, Sadness: 212, Disgust: 228, Surprise: 155 words) and has a value of 0.630 overall accuracy. The lowest overall accuracy value is Lexicon1 with 0,585.

Table 5.5. Lexicons overall accuracy table

| Lexicon Name | Overall Accuracy | Number of Words |
|---|---|---|
| Lexicon1 | 0,585 | 66 words |
| Lexicon2 | 0,740 | 66 words |
| Lexicon3 | 0,630 | 1320 words |

In *MI-weighted-tf*, is tested with Lexicon3 as mentioned in earlier sections. At this stage, MI words in Tremo sentences were examined and an MI value was created for each emotion in each sentence. As a result, as shown in Figure 5.1, each sentence has a value for six emotions.

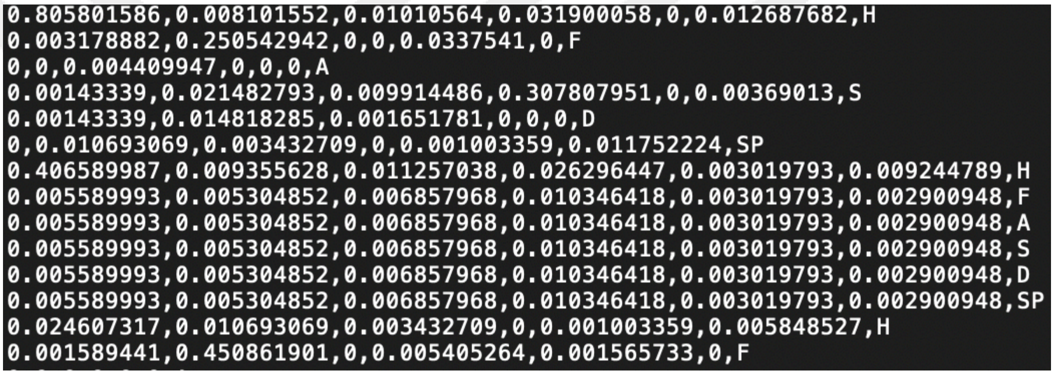
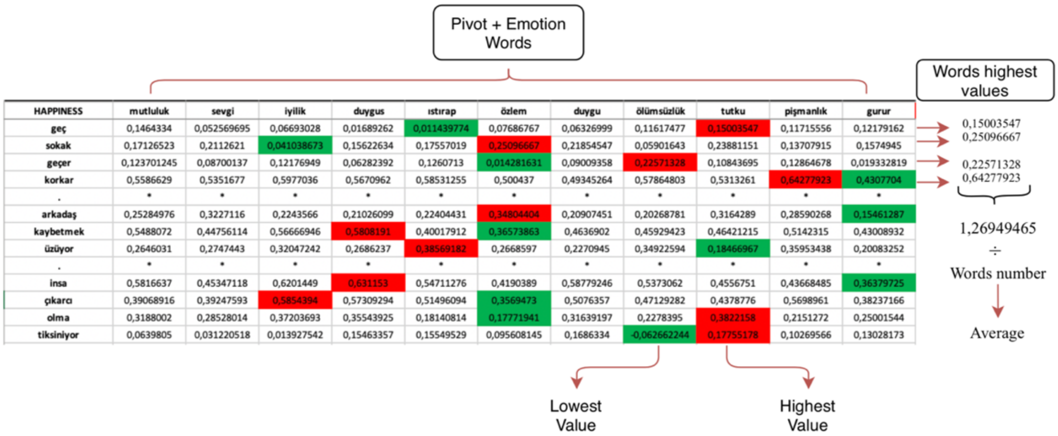

```
0.805801586,0.008101552,0.01010564,0.031900058,0,0.012687682,H
0.003178882,0.250542942,0,0,0.0337541,0,F
0,0,0.004409947,0,0,0,A
0.00143339,0.021482793,0.009914486,0.307807951,0,0.00369013,S
0.00143339,0.014818285,0.001651781,0,0,0,D
0,0.010693069,0.003432709,0,0.001003359,0.011752224,SP
0.406589987,0.009355628,0.011257038,0.026296447,0.003019793,0.009244789,H
0.005589993,0.005304852,0.006857968,0.010346418,0.003019793,0.002900948,F
0.005589993,0.005304852,0.006857968,0.010346418,0.003019793,0.002900948,A
0.005589993,0.005304852,0.006857968,0.010346418,0.003019793,0.002900948,S
0.005589993,0.005304852,0.006857968,0.010346418,0.003019793,0.002900948,D
0.005589993,0.005304852,0.006857968,0.010346418,0.003019793,0.002900948,SP
0.024607317,0.010693069,0.003432709,0,0.001003359,0.005848527,H
0.001589441,0.450861901,0,0.005405264,0.001565733,0,F
```

Figure 5.1. MI_Words table

In the table created for Lexicon3, emotions with the highest two values of each sentence were marked as in other studies. It is observed that there exists 12047 correctly classified sentences from a total of 15180 sentences in experiment. This results to an overall accuracy of 0,794.

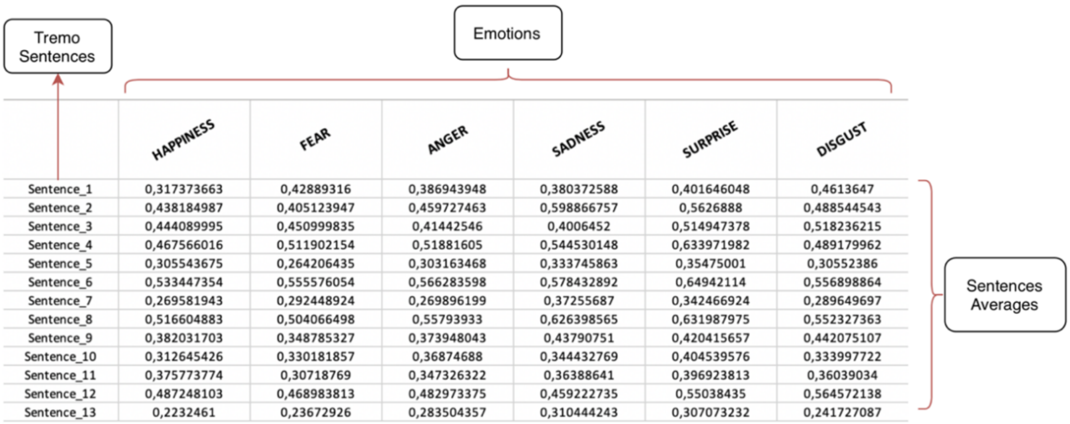
As mentioned in previous sections, in *max_similarity* method, the main idea is measuring the maximum cosine similarity of each word in sentence to each lexicon employing CBOW vectors. In experiments, every word of each Tremo sentence is examined one by one. The highest values of the words are summed up and divided by the number of words. Thus, the average is calculated as shown in Figure 5.2. For example, the highest of the similarity values of the words *'geç, sokak, geçer, korkak'* to lexicon words are added. The average is calculated by dividing the total number of words.




| HAPPINESS | mutluluk | sevgi | iyilik | duygus | ıstırap | özlem | duygu | ölümsüzlük | tutku | pişmanlık | gurur |
|---|---|---|---|---|---|---|---|---|---|---|---|
| geç | 0,1464334 | 0,052569695 | 0,06693028 | 0,01689262 | 0,011439774 | 0,07686767 | 0,06326999 | 0,11617477 | 0,15003547 | 0,11715556 | 0,12179162 |
| sokak | 0,17126523 | 0,2112621 | 0,041038673 | 0,15622634 | 0,17557019 | 0,25096667 | 0,21854547 | 0,05901643 | 0,23881151 | 0,13707915 | 0,1574945 |
| geçer | 0,123701245 | 0,08700137 | 0,12176949 | 0,06282392 | 0,1260713 | 0,014281631 | 0,09009358 | 0,22571328 | 0,10843695 | 0,12864678 | 0,019332819 |
| korkar | 0,5586629 | 0,5351677 | 0,5977036 | 0,5670962 | 0,58531255 | 0,500437 | 0,49345264 | 0,57864803 | 0,5313261 | 0,64277923 | 0,4307704 |
| . | • | • | • | • | • | • | • | • | • | • | • |
| arkadaş | 0,25284976 | 0,3227116 | 0,2243566 | 0,21026099 | 0,22404431 | 0,34804404 | 0,20907451 | 0,20268781 | 0,3164289 | 0,28590268 | 0,15461287 |
| kaybetmek | 0,5488072 | 0,44756114 | 0,56666946 | 0,5808191 | 0,40017912 | 0,36573863 | 0,4636902 | 0,45929423 | 0,46421215 | 0,5142315 | 0,43008932 |
| üzüyor | 0,2646031 | 0,2747443 | 0,32047242 | 0,2686237 | 0,38569182 | 0,2668597 | 0,2270945 | 0,34922594 | 0,18466967 | 0,35953438 | 0,20083252 |
| . | • | • | • | • | • | • | • | • | • | • | • |
| insa | 0,5816637 | 0,45347118 | 0,6201449 | 0,631153 | 0,54711276 | 0,4190389 | 0,58779246 | 0,5373062 | 0,4556751 | 0,43668485 | 0,36379725 |
| çıkarcı | 0,39068916 | 0,39247593 | 0,5854394 | 0,57309294 | 0,51496094 | 0,3569473 | 0,5076357 | 0,47129282 | 0,4378776 | 0,5698961 | 0,38237166 |
| olma | 0,3188002 | 0,28528014 | 0,37203693 | 0,35543925 | 0,18140814 | 0,17771941 | 0,31639197 | 0,2278395 | 0,3822158 | 0,2151272 | 0,25001544 |
| tiksiniyor | 0,0639805 | 0,031220518 | 0,013927542 | 0,15463357 | 0,15549529 | 0,095608145 | 0,1686334 | -0,062662244 | 0,17755178 | 0,10269566 | 0,13028173 |

Figure 5.2. Max similarity calculation table

The calculated values are accepted as the average of the sentences and the values for six emotions are tabulated as shown in Figure 5.3. Then, the first and second emotions with the highest value for each sentence are marked as in Figure 5.4. In the other stage, previously determined labels (validated emotions) are added to the table. If any of the two emotion values marked matches the label, it is considered correct classification (true positive).




| | HAPPINESS | FEAR | ANGER | SADNESS | SURPRISE | DISGUST |
|---|---|---|---|---|---|---|
| Sentence_1 | 0,317373663 | 0,42889316 | 0,386943948 | 0,380372588 | 0,401646048 | 0,4613647 |
| Sentence_2 | 0,438184987 | 0,405123947 | 0,459727463 | 0,598866757 | 0,5626888 | 0,488544543 |
| Sentence_3 | 0,444089995 | 0,450999835 | 0,41442546 | 0,4006452 | 0,514947378 | 0,518236215 |
| Sentence_4 | 0,467566016 | 0,511902154 | 0,51881605 | 0,544530148 | 0,633971982 | 0,489179962 |
| Sentence_5 | 0,305543675 | 0,264206435 | 0,303163468 | 0,333745863 | 0,35475001 | 0,30552386 |
| Sentence_6 | 0,533447354 | 0,555576054 | 0,566283598 | 0,578432892 | 0,64942114 | 0,556898864 |
| Sentence_7 | 0,269581943 | 0,292448924 | 0,269896199 | 0,37255687 | 0,342466924 | 0,289649697 |
| Sentence_8 | 0,516604883 | 0,504066498 | 0,55793933 | 0,626398565 | 0,631987975 | 0,552327363 |
| Sentence_9 | 0,382031703 | 0,348785327 | 0,373948043 | 0,43790751 | 0,420415657 | 0,442075107 |
| Sentence_10 | 0,312645426 | 0,330181857 | 0,36874688 | 0,344432769 | 0,404539576 | 0,333997722 |
| Sentence_11 | 0,375773774 | 0,30718769 | 0,347326322 | 0,36388641 | 0,396923813 | 0,36039034 |
| Sentence_12 | 0,487248103 | 0,468983813 | 0,482973375 | 0,459222735 | 0,55038435 | 0,564572138 |
| Sentence_13 | 0,2232461 | 0,23672926 | 0,283504357 | 0,310444243 | 0,307073232 | 0,241727087 |

Figure 5.3. Max similarity sentence averages table

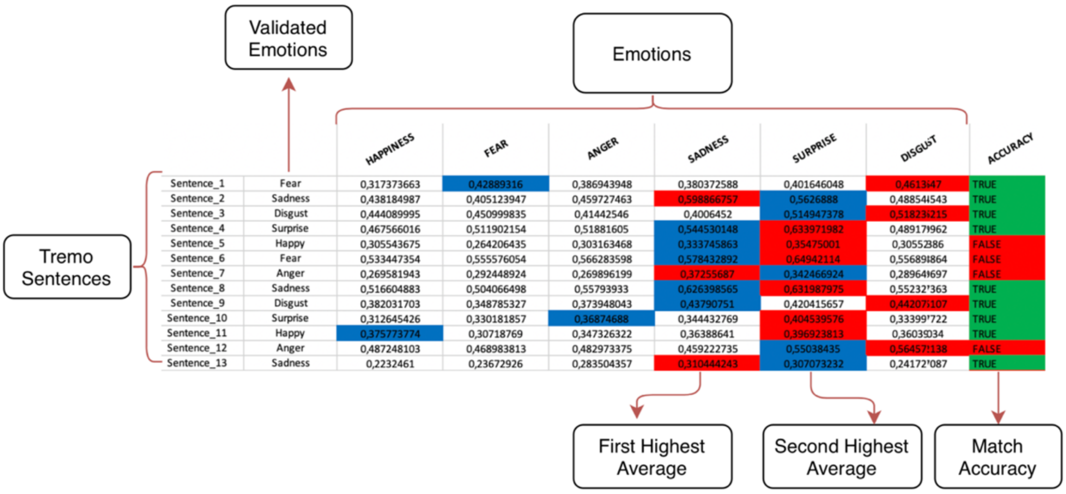



Figure 5.4. Max similarity accuracy table

When the overall results are examined; Lexicon3 provided the highest number of matches as shown in Table 5.6. The most mismatches were observed in Lexicon2.

Table 5.6. Max similarity matching accuracy table

| Lexicon Number | # of correctly classified sentences | # of incorrectly classified sentences | Accuracy |
|---|---|---|---|
| *Lexicon1* | 9820 | 10803 | 0,476 |
| *Lexicon2* | 9462 | 11161 | 0,459 |
| *Lexicon3* | 13568 | 7055 | 0,658 |

Table 5.7 gives the emotion-based accuracy values of *max_similarity* method with a similar structure to Table 5.4.

Table 5.7. Max_similarity method accuracy table

| | Lexicon1 | | Lexicon2 | | Lexicon3 | |
|---|---|---|---|---|---|---|
| **Emotion** | **Accuracy** | **Count** | **Accuracy** | **Count** | **Accuracy** | **Count** |
| Happy | 0,429 | 4135 | 0,399 | 4135 | 0,710 | 4135 |
| Fear | 0,409 | 3206 | 0,376 | 3206 | 0,615 | 3206 |
| Anger | 0,088 | 3877 | 0,078 | 3877 | 0,869 | 3877 |
| Sadness | 0,487 | 4048 | 0,375 | 4048 | 0,493 | 4048 |
| Disgust | 0,731 | 2634 | 0,928 | 2634 | 0,793 | 2634 |

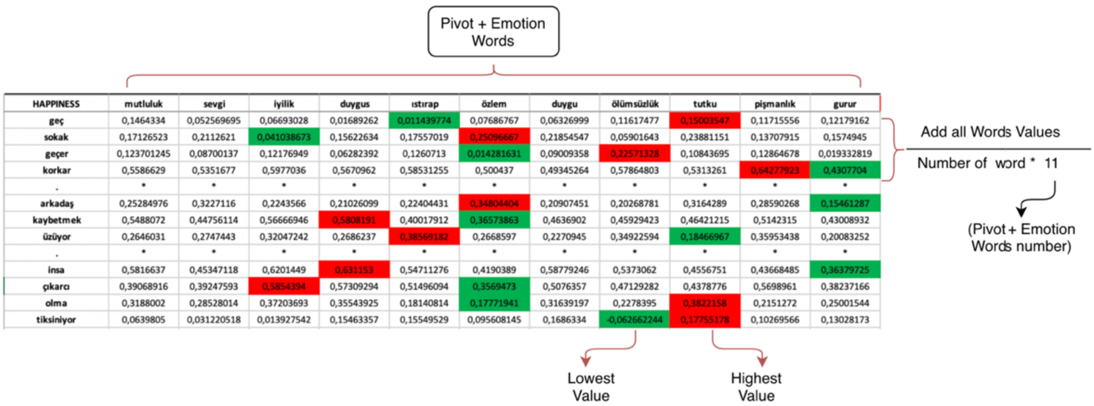
Table 5.7. (continued)

| | | | | | | |
|---|---|---|---|---|---|---|
| Surprise | 0,916 | 2723 | 0,859 | 2723 | 0,445 | 2723 |
| **TOTAL** | | **20623** | | **20623** | | **20623** |

Examining the results in Table 5.7, it may be stated that

1. *Max_similarity* succeeds in disgust (accuracy range [0.793-0.928]) and fails in sadness emotion (accuracy range [0.375-0.493]) for all lexicons.
2. Though Lexicon1 and 2 succeed, Lexicon3 does not in emotion surprise
3. Though Lexicon3 succeeds, Lexicon1 and 2 fails in emotions happy, fear and anger.

The last experiments in keyword spotting method involve the use of *average similarity* method. Instead of taking the highest values of the words, the similarity values of all the words of the sentence were taken as shown in Figure 5.5 in *average similarity* method. For example, the similarities of the words 'geç, sokak, geçer, korkak' to pivot and ten emotion words are all processed. Following, all the added values are then divided by the number of values. Using this method, the average of the sentence is calculated as shown in Figure 5.6.




| HAPPINESS | mutluluk | sevgi | iyilik | duygus | ıstırap | özlem | duygu | ölümsüzlük | tutku | pişmanlık | gurur |
|---|---|---|---|---|---|---|---|---|---|---|---|
| geç | 0,1464334 | 0,052569695 | 0,06693028 | 0,01689262 | 0,011439774 | 0,07686767 | 0,06326999 | 0,11617477 | 0,15003547 | 0,11715556 | 0,12179162 |
| sokak | 0,17126523 | 0,2112621 | 0,041038673 | 0,15622634 | 0,17557019 | 0,25096667 | 0,21854547 | 0,05901643 | 0,23881151 | 0,13707915 | 0,1574945 |
| geçer | 0,123701245 | 0,08700137 | 0,12176949 | 0,06282392 | 0,1260713 | 0,014281631 | 0,09009358 | 0,22571328 | 0,10843695 | 0,12864678 | 0,019332819 |
| korkar | 0,5586629 | 0,5351677 | 0,5977036 | 0,5670962 | 0,58531255 | 0,500437 | 0,49345264 | 0,57864803 | 0,5313261 | 0,64277923 | 0,4307704 |
| . | • | • | • | • | • | • | • | • | • | • | • |
| arkadaş | 0,25284976 | 0,3227116 | 0,2243566 | 0,21026099 | 0,22404431 | 0,34804404 | 0,20907451 | 0,20268781 | 0,3164289 | 0,28590268 | 0,15461287 |
| kaybetmek | 0,5488072 | 0,44756114 | 0,56666946 | 0,5808191 | 0,40017912 | 0,36573863 | 0,4636902 | 0,45929423 | 0,46421215 | 0,5142315 | 0,43008932 |
| üzüyor | 0,2646031 | 0,2747443 | 0,32047242 | 0,2686237 | 0,38569182 | 0,2668597 | 0,2270945 | 0,34922594 | 0,18466967 | 0,35953438 | 0,20083252 |
| . | • | • | • | • | • | • | • | • | • | • | • |
| insa | 0,5816637 | 0,45347118 | 0,6201449 | 0,631153 | 0,54711276 | 0,4190389 | 0,58779246 | 0,5373062 | 0,4556751 | 0,43668485 | 0,36379725 |
| çıkarcı | 0,39068916 | 0,39247593 | 0,5854394 | 0,57309294 | 0,51496094 | 0,3569473 | 0,5076357 | 0,47129282 | 0,4378776 | 0,5698961 | 0,38237166 |
| olma | 0,3188002 | 0,28528014 | 0,37203693 | 0,35543925 | 0,18140814 | 0,17771941 | 0,31639197 | 0,2278395 | 0,3822158 | 0,2151272 | 0,25001544 |
| tiksiniyor | 0,0639805 | 0,031220518 | 0,013927542 | 0,15463357 | 0,15549529 | 0,095608145 | 0,1686334 | -0,062662244 | 0,17755178 | 0,10269566 | 0,13028173 |

Figure 5.5. Average similarity calculation table

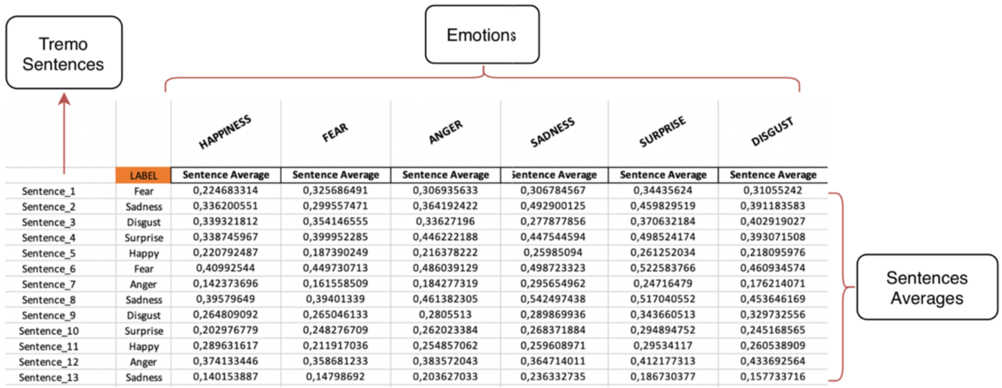



Figure 5.6. Average similarity sentence averages table

In *average similarity* method experiments, next step, in this created table, the first two emotions with the highest value for each sentence are marked as in Figure 5.7. When one of the two emotions with the highest average similarity value matched the sentence's label, it is accepted to be a true match.

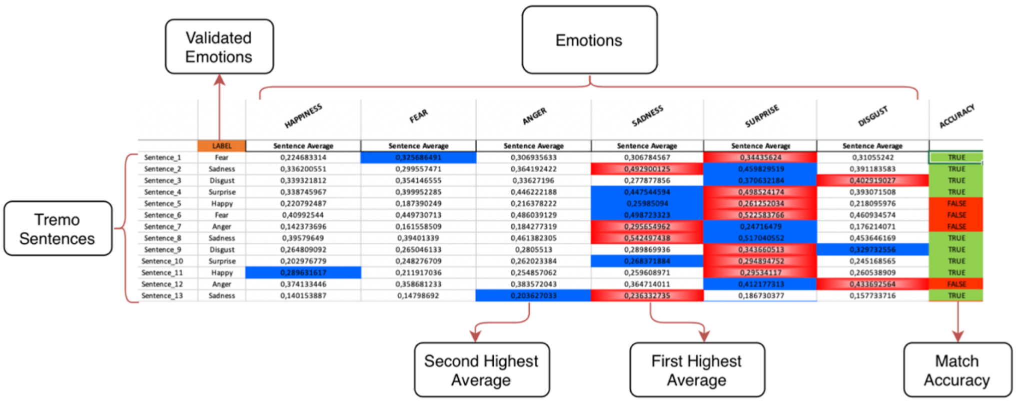



Figure 5.7. Average similarity accuracy table

Table 5.8 gives the emotion-based accuracy values of average similarity method with a similar structure to Table 5.4.

Table 5.8. Average similarity method accuracy table

| | **Lexicon1** | | **Lexicon2** | | **Lexicon3** | |
|---|---|---|---|---|---|---|
| **Emotion** | **Accuracy** | **Count** | **Accuracy** | **Count** | **Accuracy** | **Count** |
| Happy | 0,389 | 4135 | 0,535 | 4135 | 0,169 | 4135 |
| Fear | 0,287 | 3206 | 0,347 | 3206 | 0,723 | 3206 |
| Anger | 0,192 | 3877 | 0,190 | 3877 | 0,802 | 3877 |

Table 5.8. (continued)

| | | | | | | |
|---|---|---|---|---|---|---|
| Sadness | 0,536 | 4048 | 0,496 | 4048 | 0,355 | 4048 |
| Disgust | 0,679 | 2634 | 0,745 | 2634 | 0,159 | 2634 |
| Surprise | 0,862 | 2723 | 0,872 | 2723 | 0,087 | 2723 |
| **TOTAL** | | **20623** | | **20623** | | **20623** |

Examining the results in Table 5.8, it may be stated that

1. In *average similarity* method, there exists no emotion that all lexicons succeed.
2. Though Lexicon1 and 2 succeed, Lexicon3 does not in emotion surprise and disgust
3. Though Lexicon3 succeeds in fear and anger, Lexicon1 and 2 fails in both of these emotions.

In Table 5.9, total number of correctly/incorrectly classified sentences and accuracy values are presented. As the results in Table 5.9 are examined, the highest performance is observed to be obtained by Lexicon2.

Table 5.9. Average similarity matching accuracy

| **Lexicon Number** | **# of correctly classified sentences** | **# of incorrectly classified sentences** | **Accuracy** |
|---|---|---|---|
| *Lexicon1* | 9579 | 11044 | 0,465 |
| *Lexicon2* | 10406 | 10217 | 0,505 |
| *Lexicon3* | 8218 | 12405 | 0,399 |

Table 5.10 gives overall accuracy results of keyword spotting method. To exemplify, when Lexicon 1 is used in tf method, the overall accuracy is observed to be 0,585. It is examined that the highest accuracy values 0,740 and 0,794 belong to *tf* method with Lexicon 2 and *MI_weighted_tf* with Lexicon3. Considering that Lexicon3 is actually built up utilizing the TREMO itself, such a high accuracy value is not surprising. On the other hand, when Lexicon2 is used, though it is not possible to cover a large set of sentences in the experiment. (only 3583 sentences), tf method and Lexicon2 data set duo succeeds in that limited number of sentences.

Table 5.10. Accuracy results of key spotting method

| **Keyword Spotting method** | **Lexicon1** | **Lexicon2** | **Lexicon3** |
|---:|---:|---:|---:|
| tf | 0,585 | 0,740 | 0,630 |
| *MI-weighted-tf* | - | - | 0,794 |
| *max_similarity* | 0,476 | 0,459 | 0,658 |
| *average similarity* | 0,465 | 0,505 | 0,399 |

### *5.2. Supervised Learning Experiments*

In supervised learning experiments, BayesNet, Naïve Bayes, SMO, J48 and random forest methods are utilized with datasets that are used in keyword spotting methods. The performance of methods is measured by average precision, recall and F1 values. Precision and recall are simply defined as follows

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

where TP refers to true positives (number of correctly classified instances), FN is false negative, and FP is false positive. F1 is given as

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3}$$

The datasets and their brief descriptions are given in Table 5.11.

Table 5.11. The datasets that are utilized in supervised method experiments

| **Dataset #** | **Description** |
|---:|---|
| 1 | The dataset is built using lexicon1 and tf keyword spotting method. |
| 2 | The dataset is built using lexicon2 and tf keyword spotting method. |
| 3 | The dataset is built using lexicon3 and tf keyword spotting method. |
| 4 | The dataset is built using lexicon1 and average similarity keyword spotting method. |
| 5 | The dataset is built using lexicon2 and average similarity keyword spotting method. |
| 6 | The dataset is built using lexicon3 and average similarity keyword spotting method. |
| 7 | The dataset is built using lexicon1 and max similarity keyword spotting method. |

Table 5.11. (continued)

| | |
|---|---|
| 8 | The dataset is built using lexicon2 and max similarity keyword spotting method. |
| 9 | The dataset is built using lexicon3 and max similarity keyword spotting method. |
| 10 | The dataset is built using lexicon3 and MI_weighted keyword spotting method. |

In Table 5.12, the performance results of learning experiments are depicted for all supervised methods. The values in shaded cells refer to the highest F1 value for regarding dataset in Table 5.12. For example, for Dataset1 the highest F1 value is provided by random forest method. Examining the classification methods that give the highest F1 values in Table 5.12, it is observed that in 4 of 10 sets random forest method give the acceptable highest scores of classifications. Though SMO method gives highest scores for 6 datasets, it cannot be considered as a succeeding method due to the F1 values lower than 0.5.

Table 5.12. Supervised methods-results

| Data set | Data set size | Classification Methods | TP Rate | Precision | Recall | F1-Measure |
|---|---|---|---|---|---|---|
| Dataset 1 (Lexicon1+ tf method) | 4474 | BayesNet | 0,720 | 0,713 | 0,720 | 0,686 |
| | | Naive Bayes | 0,718 | 0,708 | 0,718 | 0,681 |
| | | SMO | 0,728 | 0,742 | 0,728 | 0,692 |
| | | J48 | 0,727 | 0,745 | 0,727 | 0,690 |
| | | Random Forest | 0,731 | 0,743 | 0,731 | 0,697 |
| Dataset 2 (Lexicon2+ tf method) | 3583 | BayesNet | 0,723 | 0,731 | 0,723 | 0,687 |
| | | Naive Bayes | 0,722 | 0,738 | 0,722 | 0,694 |
| | | SMO | 0,744 | 0,800 | 0,744 | 0,716 |
| | | J48 | 0,745 | 0,800 | 0,745 | 0,717 |
| | | Random Forest | 0,745 | 0,794 | 0,745 | 0,718 |

Table 5.12. (continued)

| | | | | | | |
|---|---|---|---|---|---|---|
| Dataset 3 (Lexicon3+ tf method) | 15180 | BayesNet | 0,546 | 0,535 | 0,546 | 0,536 |
| | | Naive Bayes | 0,496 | 0,523 | 0,496 | 0,473 |
| | | SMO | 0,593 | 0,575 | 0,593 | 0,587 |
| | | J48 | 0,618 | 0,599 | 0,618 | 0,604 |
| | | Random Forest | 0,623 | 0,609 | 0,623 | 0,610 |
| Dataset 4 (Lexicon1+ average similarity method) | 21558 | BayesNet | 0,239 | 0,258 | 0,239 | 0,165 |
| | | Naive Bayes | 0,238 | 0,365 | 0,238 | 0,164 |
| | | SMO | 0,425 | 0,438 | 0,425 | 0,382 |
| | | J48 | 0,355 | 0,346 | 0,355 | 0,350 |
| | | Random Forest | 0,239 | 0,258 | 0,239 | 0,165 |
| Dataset 5 (Lexicon2+ average similarity method) | 21558 | BayesNet | 0,246 | 0,283 | 0,246 | 0,171 |
| | | Naive Bayes | 0,244 | 0,346 | 0,244 | 0,169 |
| | | SMO | 0,451 | 0,431 | 0,451 | 0,409 |
| | | J48 | 0,377 | 0,366 | 0,377 | 0,371 |
| | | Random Forest | 0,246 | 0,283 | 0,246 | 0,171 |
| Dataset 6 (Lexicon3+ average similarity method) | 21558 | BayesNet | 0,342 | 0,366 | 0,342 | 0,300 |
| | | Naive Bayes | 0,346 | 0,402 | 0,346 | 0,306 |
| | | SMO | 0,511 | 0,484 | 0,511 | 0,490 |
| | | J48 | 0,422 | 0,412 | 0,422 | 0,416 |
| | | Random Forest | 0,342 | 0,366 | 0,342 | 0,300 |
| Dataset 7 (Lexicon1+ max similarity method) | 21558 | BayesNet | 0,249 | 0,348 | 0,249 | 0,179 |
| | | Naive Bayes | 0,249 | 0,481 | 0,249 | 0,182 |
| | | SMO | 0,420 | 0,412 | 0,420 | 0,399 |
| | | J48 | 0,343 | 0,336 | 0,343 | 0,339 |
| | | Random Forest | 0,249 | 0,348 | 0,249 | 0,179 |

Table 5.12. (continued)

| | | | | | | |
|---|---|---|---|---|---|---|
| Dataset 8 (Lexicon2+ max similarity method) | 21558 | BayesNet | 0,249 | 0,332 | 0,249 | 0,182 |
| | | Naive Bayes | 0,248 | 0,506 | 0,248 | 0,182 |
| | | SMO | 0,436 | 0,416 | 0,436 | 0,414 |
| | | J48 | 0,358 | 0,349 | 0,358 | 0,353 |
| | | Random Forest | 0,249 | 0,332 | 0,249 | 0,182 |
| Dataset 9 (Lexicon3+ max similarity method) | 21558 | BayesNet | 0,430 | 0,437 | 0,430 | 0,422 |
| | | Naive Bayes | 0,431 | 0,442 | 0,431 | 0,422 |
| | | SMO | 0,497 | 0,476 | 0,497 | 0,482 |
| | | J48 | 0,436 | 0,428 | 0,436 | 0,432 |
| | | Random Forest | 0,430 | 0,437 | 0,430 | 0,422 |
| Dataset 10 (Lexicon3+ MI-weighted method) | 15180 | BayesNet | 0,613 | 0,650 | 0,613 | 0,616 |
| | | Naive Bayes | 0,433 | 0,641 | 0,433 | 0,462 |
| | | SMO | 0,462 | 0,697 | 0,462 | 0,487 |
| | | J48 | 0,656 | 0,641 | 0,656 | 0,644 |
| | | Random Forest | 0,654 | 0,639 | 0,654 | 0,644 |

In Table 5.13, the average performance results of supervised learning approaches are given. Examining the values in Table 5.13 it is observed that

1. Dataset 1 and 2 provide the highest performance values. On the other hand, the number of sentences that are involved in experiments is lower compared to other sets.
2. Datasets 4-9 give lowest classification results. As a result, it may be stated that when max-similarity or average similarity methods are used all the lexicons fail to detect emotions using supervised approach.
3. Though Lexicon3 is constructed from TREMO set, the classification performance given as F1 Measure of Dataset3 is not as high as expected.

Table 5.13. Average performance results of supervised learning for different sets

| Dataset | Dataset Size | TP Rate | Precision | Recall | F1-Measure |
|---|---|---|---|---|---|
| Dataset 1 (Lexicon1+ tf method) | 4474 | 0,725 | 0,730 | 0,725 | 0,689 |
| Dataset 2 (Lexicon2+ tf method) | 3583 | 0,736 | 0,773 | 0,736 | 0,706 |
| Dataset 3 (Lexicon3+ tf method) | 15180 | 0,575 | 0,568 | 0,575 | 0,562 |
| Dataset 4 (Lexicon1+ average similarity method) | 21558 | 0,299 | 0,333 | 0,299 | 0,245 |
| Dataset 5 (Lexicon2+ average similarity method) | 21558 | 0,313 | 0,342 | 0,313 | 0,258 |
| Dataset 6 (Lexicon3+ average similarity method) | 21558 | 0,393 | 0,406 | 0,393 | 0,362 |
| Dataset 7 (Lexicon1+ max similarity method) | 21558 | 0,302 | 0,385 | 0,302 | 0,256 |
| Dataset 8 (Lexicon2+ max similarity method) | 21558 | 0,308 | 0,387 | 0,308 | 0,263 |
| Dataset 9 (Lexicon3+ max similarity method) | 21558 | 0,445 | 0,444 | 0,445 | 0,436 |
| Dataset 10 (Lexicon3+ MI-weighted method) | 15180 | 0,564 | 0,654 | 0,564 | 0,571 |

# CHAPTER 6 : CONCLUSION

In this thesis, the success of emotion lexicons and word vector representations in the emotion detection task was investigated. Two datasets were used in the studies. These datasets are 'Tremo' and 'Wikipedia'. Related pre-processing operations have been applied to the datasets. The contributions and regarding experiments together with findings in thesis are as follows:

- Two approaches to build emotion lexicon are proposed. These approaches are implemented as Lexicon1 and Lexicon2 where vector-based similarity is used to empower the selection of emotion words. In order to compare the performance of our proposed lexicons, an existing lexicon (was named as Lexicon3 in thesis) (Toçoğlu, and Alpkoçak, 2019) that was build from Tremo dataset is also employed in the experiments. It is observed that it was not possible to label a large set of sentences using the proposed lexicons due to limited number of emotion words in lexicons. On the other hand, it is examined that especially second lexicon, given as Lexicon2, succeeded in the sentences that may be included in the experiment.
- The performance of both unsupervised, given as keyword spotting, and supervised methods are explored in emotion detection in Turkish texts. The experiments revealed that the neither supervised not unsupervised methods outperforms the other.
- It is shown that the emotion detection scores vary for different emotions. For example, occurrence frequency based tf method succeeds in detection of happy, surprise and disgust sentences but it fails in other emotions.
- The lexicon-based emotion detection is performed by two different ways of comparison to lexicon words. The first group of methods involves basic string comparison to lexicon words and the second group requires the vector-based similarity of sentence words to emotion words.

As a future work, we plan to enhance our lexicons where word vector similarity is employed to determine lexicon words. In addition, we will run experiments with succeeding methods on different datasets.

# REFERENCES

Agarwal, A., Xie, B., Vovsha, I., Rambow, O. and Passonneau, R. (2011) *Sentiment analysis of twitter data.* Workshop on Language in Social Media, New York, pp. 30-38.

Akbaş, E. (2012) *Aspect based opinion mining on Turkish tweets.* Unpublished Master Thesis. Bilkent University.

Aue, A. and Gamon, M. (2005). *Customizing Sentiment Classifiers to New Domains: a Case Study.* International Conference on Recent Advances in Natural Language Processing, Borovets, Bulgaria.

Baccianella, S., Esuli, A. and Sebastiani, F. (2010). *SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining.* The 7th International Conference on Language Resources and Evaluation, Valletta, Malta, pp. 2200-2204.

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

Blinov, P. D., Klekovkine, M. V., Kotelnikov, E. V. and Pestov, O. A. (2013) *Research of lexical approach and machine learning methods for sentiment analysis*. Comput Linguist & Intellect Technol. Vol. 2, pp. 48-58.

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

Büyükkınacı, M. (2018, January 07). *Word2Vec Nedir? (Türkçe).* [Blog]. Available at: https://medium.com/@muhammedbuyukkinaci/word2vec-nedir-türkçe-f0cfab20d3ae

Çebi, C. (2018) *Rastgele Orman Algoritması.* [Online]. Available at: https://medium.com/@cemthecebi/rastgele-orman-algoritması-d1260a920f24 (Access Date: 3 Junuary 2020).

Dehkharghani, R., Saygın, Y., Yanıkoğlu, B. and Oflazer, K. (2015) S*entiturknet: a Turkish polarity lexicon for sentiment analysis*. Lang Resour & Eval, Issue 50, pp. 667-685.

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

Ding, X., Liu, B. and Yu, P. S. (2008) *A Holistic Lexicon-Based Approach to Opinion Mining.* Conference on Web Search and Web Data Mining (WSDM), Palo Alto, CA, USA.

Ekman, P. (1992) An argument for basic emotions. *Cognition & Emotion,* Vol. 6(3/4), pp. 169-200.

Envision Your Evolution. (2019, June 06). *Paul Ekman's Theory of Emotion* [Blog]. Available at: https://www.envisionyourevolution.com/evolution-emotion/paul-ekman-theory-of-emotion/2149/

Fellbaum, C. (1998) *WordNet: An Electronic Lexical Database.* London, UK: MIT Press.

García, J. (2019) *Introduction to emotion recognition in text.* [Online]. Available at: https://www.meaningcloud.com/blog/introduction-to-emotion-recognition-in-text (Access Date: 5 Junuary 2020).

Geneva University (2013) *International Survey On Emotion Antecedents And Reactions (ISEAR).* [Online]. Available at: https://www.unige.ch/cisa/research/materials-and-online-research/research-material/ (Access Date: 12 February 2020).

John, G. and Langley, P. (1995) *Estimating Continuous Distributions in Bayesian Classifiers.* Proceedings Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo, pp. 338-45.

Kaynar, O., Görmez, Y., Yıldız, M. and Albayrak, A. (2016) *Sentiment Analysis with Machine Learning Techniques.* International Artifical Intelligence and Data Processing Symposium (IDAP'16), Sivas, Turkey.

Mikolov, T., Sutskever, I., Chen, K., Corrado G. and Dean J. (2013) *Distributed Representations of Words and Phrases and their Compositionality.* Vol. 2, pp. 3111-3119

Mohammad, S. (2010) *NRC Word-Emotion Association Lexicon.* [Online]. Available at: http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.html (Access Date: 7 March 2020).

Mohammad, S. (2012) *#Emotional Tweets.* First Joint Conference on Lexical and Computational Semantics, Montréal, Canada, pp. 246-255.

Mohammad, S. M. and Kiritchenko, S. (2015) *Using Hashtags to Capture Fine Emotion Categories from Tweets*. Computational Intelligence, Vol. 31, pp. 301-326.

Mohammad, S. M. and Turney, P. D. (2012) *Crowdsourcing a Word–Emotion Association Lexicon.* Computational Intelligence, Vol. 29(3), pp. 436-465.

Nasilbe (2018) *Farklı Duygu Türleri Nelerdir?.* [Online]. Available at: Medium.com: https://medium.com/@nasilbe61/farklı-duygu-türleri-nelerdir-aa618b33341a (Access Date: 3 January 2020).

Nielsen, F. Å. (2011) *A new ANEW: Evaluation of a word list for sentiment analysis in microblogs.* Workshop on Making Sense of Microposts, Heraklion, Crete, Greece, pp. 93-98.

Oxford Dictionary (2020) *Oxford Learner's Dictionary.* [Online]. Available at: https://www.oxfordlearnersdictionaries.com (Access Date: 4 January 2020).

Pang, B., Lee, L. and Vaithyanathan, S. (2002) *Thumbs up? Sentiment Classification using Machine Learning Techniques.* Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), pp. 30-38.

Pervan, N. (2019) *Duygu Analizi Neden Önemli?.* [Online] Available at: SisaSoft: https://www.sisasoft.com.tr/duygu-analizi-neden-onemli/ (Access Date: 3 January 2020).

Plutchik, R. (1980) *A general psychoevolutionary theory of emotion.* In: Emotion: Theory, Research and Experience. New York: Elsevier Inc., pp. 1: 3-33.

Quinlan, R. (1993) *C4.5 Programs for Machine Learning.* San Mateo. CA. : Morgan Kaufmann Publishers.

Saif, H., He, Y. and Alani, H. (2012) *Alleviating data sparsity for Twitter sentiment analysis.* 2nd Workshop on Making Sense of Microposts (#MSM2012): Big things come in small packages at the 21st International Conference on theWorld Wide Web (WWW'12), Lyon, France, pp. 2-9.

Standard, E. (2017, December 13). *Tapping Into Emotion: 10 Tips for Driving Connections in Higher Ed Marketing.* [Blog]. Available at: https://medium.com/@easternstandard/tapping-into-emotion-10-tips-for-driving-connections-in-higher-ed-marketing-80a80dde7ca1

Stemmer, (2002). *Snowball Stemmer.* [Online]. Available at: https://snowballstem.org (Access Date: 2 Junuary 2020).

Strapparava, C. and Valitutti, A. (2004) *WordNet-Affect: an Affective Extension of WordNet.* The 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal, pp. 1083-1086

Şeker, S. E. (2013). *İş Zekası ve Veri Madenciliği (Weka ile).* ISBN 9786051276717, Cinius Publications, Istanbul, Turkey.

Taboada, M. Brooke, J., Tofiloski, M., Voll, K. and Stede, M. (2011) *Lexicon-Based Methods for Sentiment Analysis*. Computational Linguistics, Vol. 37(2), pp. 267-307.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D. and Kappas, A. (2010) *Sentiment Strength Detection in Short Informal Text.* Journal of the American Society for Information Science and Technology, Vol. 61(12), pp. 2544-2558.

Toçoğlu, M. A., Çelikten, A., Aygün, İ. and Alpkoçak, A. (2019) *Comparison of Different Machine Learning Approaches for Emotion Analysis in Turkish.* Dokuz Eylul University Faculty of Engineering Journal of Science and Engineering, Vol. 21(63), pp. 719-725.

Toçoğlu, M. A. and Alpkoçak, A. (2019) *Lexicon-based emotion analysis in Turkish.* Turkish Journal of Electrical Engineering & Computer Sciences, Vol. 27(2), pp. 1213-1227.

Toçoğlu, M. and Alpkoçak, A. (2005) *Tremo Dataset.* [Online] Available at: http://demir.cs.deu.edu.tr/tremo-dataset/ (Access Date: 2 Junuary 2020).

Uçan, A., Naderalvojoud, B., Sezer, E. A. and Sever, H. (2016) *SentiWordNet for New Language: Automatic Translation Approach.* 12th International Conference on Signal-Image Technology & Internet- Based Systems, Naples, Italy, pp. 308-315.

WEKA, (2014) *Waikato Environment for Knowledge Analysis.* [Online]. Available at: https://www.cs.waikato.ac.nz/ml/weka/ (Access Date: 2 January 2020).

Wiebe, J., Wilson, T. and Cardie, C. (2005). A*nnotating Expressions of Opinions and Emotions in Language.* Language Resources and Evaluation, Vol. 39, pp. 165-210.

[Wikipedia]. (2018, June 15) *Turkish Wikipedia articles.* [Wiki articles]. Available at: https://tr.wikipedia.org/wiki/Anasayfa

[Wikipedia]. (2020, January 6) *Bayesian Network.* [Wiki articles]. Available at: https://en.wikipedia.org/wiki/Bayesian_network

[Wikipedia]. (2020, January 6) *Sequential Minimal Optimization.* [Wiki articles]. Available at: https://en.wikipedia.org/wiki/Sequential_minimal_optimization

Yüce, D. M. (2019) *Duygu Nedir- Bilgi Açısından Önemi.* [Online] Available at: https://www.dmy.info/duygu-nedir-bilgi-acisindan-onemi/

# APPENDICES

## *A1. Stop Words List*

| STOPWORDS LIST TABLE | | | | | | |
|---|---|---|---|---|---|---|
| acaba | altı | altmış | ama | ancak | arada | artık |
| asla | aslında | ayrıca | az | bana | bazen | bazı |
| bazıları | belki | ben | benden | beni | benim | beri |
| beş | bile | bilhassa | bin | bir | biraz | birçoğu |
| birçok | biri | birisi | birkaç | birşey | biz | bizden |
| bize | bizi | bizim | böyle | böylece | bu | buna |
| bunda | bundan | bunlar | bunları | bunların | bunu | bunun |
| burada | bütün | çoğu | çoğunu | çok | çünkü | da |
| daha | dahi | dan | de | defa | değil | diğer |
| diğeri | diğerleri | diye | doksan | dokuz | dolayı | dolayısıyla |
| dört | edecek | eden | ederek | edilecek | ediliyor | edilmesi |
| ediyor | eğer | elbette | elli | en | etmesi | etti |
| ettiği | ettiğini | fakat | falan | filan | gene | gereği |
| gerek | gibi | göre | hala | halde | halen | hangi |
| hangisi | hani | hatta | hem | henüz | hep | hepsi |
| her | herhangi | herkes | herkese | herkesi | herkesin | hiç |
| hiçbir | hiçbiri | için | içinde | iki | ile | ilgili |
| ise | işte | itibaren | itibariyle | kaç | kadar | karşın |
| kendi | kendilerine | kendine | kendini | kendisi | kendisine | kendisini |
| kez | ki | kim | kime | kimi | kimin | kimisi |
| kimse | kırk | madem | mi | mı | milyar | milyon |
| mu | mü | nasıl | ne | neden | nedenle | nerde |
| nerede | nereye | neyse | niçin | nin | nın | niye |
| nun | nün | öbür | olan | olarak | oldu | olduğu |
| olduğunu | olduklarını | olmadı | olmadığı | olmak | olması | olmayan |
| olmaz | olsa | olsun | olup | olur | olursa | oluyor |
| on | ön | ona | önce | ondan | onlar | onlara |
| onlardan | onları | onların | onu | onun | orada | öte |
| ötürü | otuz | öyle | oysa | pek | rağmen | sana |
| sanki | şayet | şekilde | sekiz | seksen | sen | senden |
| seni | senin | şey | şeyden | şeye | şeyi | şeyler |
| şimdi | siz | sizden | size | sizi | sizin | sonra |
| şöyle | şu | şuna | şunları | şunu | ta | tabii |
| tam | tamam | tamamen | tarafından | trilyon | tüm | tümü |
| üç | un | ün | üzere | var | vardı | ve |
| veya | ya | yani | yapacak | yapılan | yapılması | yapıyor |
| yapmak | yaptı | yaptığı | yaptığını | yaptıkları | ye | yedi |
| yerine | yetmiş | yi | yı | yine | yirmi | yoksa |
| yu | yüz | zaten | zira | a | b | c |
| ç | d | e | f | g | ğ | h |
| ı | i | j | k | l | m | n |
| o | ö | p | r | s | ş | t |
| u | ü | v | y | z | w | q |
| x | | | | | | |

*A2. Stemmed Stop Words List*

| STOPWORDS STEMFORM LIST TABLE | | | | | | |
|---|---|---|---|---|---|---|
| acap | al | altmış | am | ancak | ara | ar |
| as | asl | ayrıç | az | ba | baze | baz |
| bazı | belki | ben | be | be | be | ber |
| beş | bil | bilhas | bin | bir | biraz | birçok |
| birçok | bir | biris | birkaç | birşey | biz | biz |
| biz | biz | biz | bö | böyleç | bu | p |
| p | p | bun | bun | p | p | bu |
| bura | büt | çok | çok | çok | çünkü | da |
| dah | dahi | dan | de | defa | değil | diğer |
| diğer | diğer | di | dok | dok | dola | dolayı |
| dört | edecek | e | ederek | edilecek | ediliyor | edilmes |
| ediyor | eğer | elbet | elli | en | etmes | et |
| ettik | ettik | fakat | fala | fila | ge | gerek |
| gerek | gip | gör | hal | halde | hale | hangi |
| hangis | hani | hat | hem | he | hep | hepsi |
| her | herhangi | herkes | herkes | herkes | herke | hiç |
| hiçbir | hiçbir | iç | iç | ik | il | ilgil |
| is | iş | itibare | itibar | kaç | kadar | karş |
| ke | kendi | kendi | kendi | kendis | kendi | kendi |
| kez | ki | kim | k | k | k | kimis |
| k | kırk | made | mi | mı | milyar | milyo |
| mu | mü | nasıl | ne | ne | nede | ner |
| nere | nere | ne | niç | nin | nın | ni |
| nun | nün | öbür | ola | olarak | ol | olduk |
| olduk | olduk | olmadı | olmadık | olmak | olmas | olmaya |
| olmaz | ol | ol | olup | olur | olur | oluyor |
| on | ön | on | ö | on | on | on |
| on | on | on | on | o | ora | ö |
| ötür | ot | ö | o | pek | rağme | sa |
| sanki | şayet | şekil | sek | sek | sen | se |
| se | se | şey | şey | şe | şe | şey |
| ş | siz | siz | siz | siz | siz | sonra |
| şö | şu | ş | şun | ş | ta | tabii |
| tam | tama | tamame | taraf | trilyo | tüm | t |
| üç | un | ün | üzer | var | var | ve |
| veya | ya | yani | yapacak | yapıla | yapılmas | yapıyor |
| yapmak | yap | yaptık | yaptık | yaptık | ye | yedi |
| yer | yet | yi | yı | y | yirmi | yok |
| yu | yüz | zate | zira | a | b | c |
| ç | d | e | f | g | ğ | h |
| ı | i | j | k | l | m | n |
| o | ö | p | r | s | ş | t |
| u | ü | v | y | z | w | q |
| x | | | | | | |

## *A3. Deleted Symbols from Wikipedia Dataset*

| DELETED SYMBOLS | | | | | | |
|---|---|---|---|---|---|---|
| ¡ | ¢ | ¤ | ¥ | ¦ | § | © |
| ª | « | ¬ | ® | ¯ | ° | ± |
| ² | ³ | ´ | µ | ¶ | · | ¸ |
| ¹ | º | » | ¼ | ¾ | ¿ | ð |
| ÷ | Ŧ | ǂ | ǁ | ŋ | ȣ | ɭ |
| ʱ | ʲ | ʳ | ʴ | ʵ | ʶ | ʷ |
| ʸ | ʼ | ˀ | ˷ | ˟ | ˠ | ˘ |
| ˔ | ˕ | ˖ | ˙ | ˭ | ˽ | ˻ |
| ˿ | ֏ | ― | — | – | ‒ | • |
| ⋮ | ‾ | ⸮ | ⁰ | ‰ | … | ⁱ |
| ⁴ | ℅ | ₿ | Ω | Ⅎ | ™ | ℗ |
| № | ↔ | ↓ | → | ↑ | ← | ⅗ |
| ⅓ | ⅔ | ⅕ | ⅖ | ⅘ | ⅙ | ⅚ |
| ⅛ | ⅜ | ⅝ | ⅞ | √ | ∞ | ∑ |
| ∏ | ∆ | ∂ | ↨ | ↙️ | ∫ | ∟ |
| ≈ | ≠ | ≡ | ≤ | ≥ | ⌂ | ⌠ |
| ⌡ | □ | ▪ | ▫ | ┘ | └ | ┐ |
| ┌ | ○ | ◌ | ● | ◦ | ₺ | ł |
| ⸗ | › | ♦ | – | → | ∨ | ∧ |
| ξ | ψ | λ | π | μ | η | ≈ |
| · | ⟦ | ⟧ | c | ℓ | α | β |
| ω | k | ∑ | • | δ | γ | ′ |
| σ | τ | φ | ς | ³ | ² | ≠ |
| − | ± | ρ | ! | × | ≡ | ⇔ |
| κ | ● | | | | | |

*A4. TREMO & WIKIPEDIA selecting pivot and 10 emotion words survey results*

TREMO:

| ***Students*** | **Happiness** | | **Fear** | | **Anger** | |
|---|---|---|---|---|---|---|
| ***** | ***'mutlu'*** | *'mutluluk'* | ***'korku'*** | *'kork'* | *'öfke'* | ***'öfkelenir'*** |
| ***Student_1*** | X | | X | | X | |
| ***Student_2*** | | X | X | | | X |
| ***Student_3*** | X | | | X | | X |
| ***Student_4*** | X | | X | | | X |
| ***Student_5*** | | X | X | | | X |

| ***Students*** | **Sadness** | | **Disgust** | | **Surprise** | |
|---|---|---|---|---|---|---|
| ***** | ***'üz'*** | *'üzülür'* | *'iğrenme'* | ***'iğrenç'*** | ***'şaşır'*** | *'şaşırıyor'* |
| ***Student_1*** | X | | X | | | X |
| ***Student_2*** | X | | | X | X | |
| ***Student_3*** | X | | | X | | X |
| ***Student_4*** | X | | X | | X | |
| ***Student_5*** | | X | | X | X | |

WIKIPEDIA:

| ***Students*** | **Happiness** | | **Fear** | | **Anger** | |
|---|---|---|---|---|---|---|
| ***** | *'mutlu'* | ***'mutluluk'*** | ***'korku'*** | *'kork'* | ***'öfke'*** | *'öfkelenir'* |
| ***Student_1*** | X | | X | | X | |
| ***Student_2*** | | X | | X | | X |
| ***Student_3*** | | X | X | | X | |
| ***Student_4*** | X | | X | | | X |
| ***Student_5*** | | X | X | | X | |

| ***Students*** | **Sadness** | | **Disgust** | | **Surprise** | |
|---|---|---|---|---|---|---|
| ***** | *'üzülür'* | ***'üzül'*** | *'iğrenme'* | ***'iğrenç'*** | ***'şaşır'*** | *'şaşırır'* |
| ***Student_1*** | | X | X | | X | |
| ***Student_2*** | X | | X | | | X |

| | | | | | | |
|---|---|---|---|---|---|---|
| ***Student_3*** | | X | | X | | X |
| ***Student_4*** | | X | | X | X | |
| ***Student_5*** | | X | | X | X | |

| **TREMO** | |
|---|---|
| *Emotions:* | *Pivot Words:* |
| Happiness | mutlu |
| Fear | korku |
| Anger | öfkelenir |
| Sadness | üz |
| Disgust | iğrenç |
| Surprise | şaşır |

| **WIKIPEDIA** | |
|---|---|
| *Emotions:* | *Pivot Words:* |
| Happiness | mutluluk |
| Fear | korku |
| Anger | öfke |
| Sadness | üzül |
| Disgust | iğrenç |
| Surprise | şaşır |

### *A5. Wikipedia selecting pivot and 10 emotion words survey results*

Top 10 selections of students based on pivot words and emotions

Student_1

| Happiness | Fear | Anger | Sadness | Disgust | Surprise |
|---|---|---|---|---|---|
| *mutluluk* | *korku* | *öfke* | *üzül* | *iğrenç* | *şaşır* |
| sevgi | dehşet | kızgınlık | üzülür | ahlaksız | sanır |
| iyilik | umutsuzluk | şaşkınlık | affeder | gülünç | üzgü |
| duygus | korkus | öfkes | pişma | acayip | şaşırır |
| duygu | çaresizlik | acı | inandırır | delilik | sinirli |
| tutku | vahşet | pişmanlık | inanmaz | korkar | uyandırır |
| gurur | zombi | hırs | üzüle | utanma | hissettik |
| hayaller | kapılır | suçluluk | anlay | hissettik | aptal |
| şefkat | ızdırap | keder | sinirle | aşağıla | anlamış |
| heyecan | acımasızlık | dehşe | anlamaz | çıkarcı | hisset |
| sevgis | duygus | kin | hatırlıyor | çirk | hissediyor |

Student_2

| Happiness | Fear | Anger | Sadness | Disgust | Surprise |
|---|---|---|---|---|---|
| *mutluluk* | *korku* | *öfke* | *üzül* | *iğrenç* | *şaşır* |
| sevgi | dehşet | kızgınlık | üzülür | ahlaksız | sanır |
| iyilik | korkus | umutsuzluk | inandırır | aptal | şaşırır |
| özlem | çaresizlik | öfkes | öfkelenir | acayip | korkmuş |
| duygu | vahşet | acı | inanmaz | alaycı | sinirli |
| tutku | korkunç | çaresizlik | üzüle | bencil | uyandırır |
| hayaller | zombi | pişmanlık | anlamaz | hissettik | hissettik |
| şefkat | acı | hırs | angelica | yapmacık | anlamış |
| heyecan | şaşkınlık | suçluluk | şaşırır | kibirli | düşünüp |
| hissettik | ızdırap | kapılır | hatırlıyor | çirk | hisset |
| sevgis | yalnızlık | sempat | sanmak | ikiyüzlülük | söyley |

Student_3

| Happiness | Fear | Anger | Sadness | Disgust | Surprise |
|---|---|---|---|---|---|
| *mutluluk* | *korku* | *öfke* | *üzül* | *iğrenç* | *şaşır* |
| sevgi | dehşet | kızgınlık | üzülür | gülünç | sanır |
| iyilik | umutsuzluk | öfkes | affeder | acayip | şaşırır |
| duygus | çaresizlik | üzüntü | pişma | korkar | korkmuş |
| ölümsüzlük | vahşet | çaresizlik | sevinir | delilik | sinirlenir |
| tutku | zombi | korkus | öfkelenir | paranoyak | aptal |
| hayaller | acı | suçluluk | sinirlenir | utanma | düşünüp |
| şefkat | çılgı | keder | üzüle | bencil | hisset |
| inandık | yalnızlık | acımasız | şüphelenir | aşağıla | korkar |
| heyecan | acımazlık | dehşe | sanmak | çıkarcı | hissediyor |
| sevgis | duygus | kin | söyley | çirk | söyley |

Student_4

| Happiness | Fear | Anger | Sadness | Disgust | Surprise |
|---|---|---|---|---|---|
| *mutluluk* | *korku* | *öfke* | *üzül* | *iğrenç* | *şaşır* |
| sevgi | dehşet | kızgınlık | üzülür | şehvet | sanır |
| iyilik | umutsuzluk | umutsuzluk | affeder | gülünç | şaşırır |
| özlem | korkus | öfkes | pişma | korkak | uyandırır |
| duygu | öfke | kıskançlık | inanmaz | alaycı | hissettik |
| gurur | vahşet | çaresizlik | üzüle | utanma | sinirlenir |
| bilgelik | korkunç | korkus | sinirle | hissettik | aptal |
| şefkat | acı | suçluluk | setsuna | yapmacık | korkar |
| keder | çılgı | acımasız | şüphelenir | aşağıla | sinirlene |
| heyecan | yalnızlık | dehşe | hatırlıyor | kibirli | kibirli |
| sevgis | acımasızlık | kin | sanmak | ikiyüzlülük | söyley |

Student_5

| Happiness | Fear | Anger | Sadness | Disgust | Surprise |
|---|---|---|---|---|---|
| *mutluluk* | *korku* | *öfke* | *üzül* | *iğrenç* | *şaşır* |
| sevgi | dehşet | kızgınlık | üzülür | ahlaksız | sanır |
| iyilik | korkus | umutsuzluk | pişma | gülünç | şaşırır |
| özlem | öfke | öfkes | inandırır | korkak | üzgü |
| duygu | korkunç | üzüntü | sevinir | alaycı | hissettik |
| gurur | duygusallık | pişmanlık | üzüle | paranoyak | anlamış |
| hayaller | acı | hırs | anlay | utanma | hisset |
| bilgelik | şehvet | acımasız | anlamaz | korkutuç | korkar |
| şefkat | yalnızlık | dehşe | şüphelenir | yapmacık | kibirli |
| heyecan | iğrenç | kin | hatırlıyor | çıkarcı | hissediyor |
| sevgis | acımasızlık | kapılıp | söyley | çirk | söyley |

***A6. Selection numbers of the first 20 words according to the similarity of the pivot word***

| **HAPPINESS** | **Num** |
|---|---|
| ***mutluluk*** | ***5*** |
| **sevgi** | **5** |
| **iyilik** | **5** |
| duygus | 2 |
| ıstırap | 0 |
| **özlem** | **3** |
| **duygu** | **4** |
| ölümsüzlük | 1 |
| **tutku** | **3** |
| pişmanlık | 0 |
| **gurur** | **3** |
| yalnızlık | 0 |
| **hayaller** | **4** |
| acı | 0 |
| bilgelik | 2 |
| **şefkat** | **5** |
| keder | 1 |
| inandık | 1 |
| **heyecan** | **5** |
| hissettik | 1 |
| **sevgis** | **5** |

| **FEAR** | **Num** |
|---|---|
| ***korku*** | ***5*** |
| **dehşet** | **5** |
| **umutsuzluk** | **3** |
| **korkus** | **4** |
| öfke | 2 |
| **çaresizlik** | **3** |
| **vahşet** | **4** |
| **korkunç** | **3** |
| **zombi** | **3** |
| duygusallık | 1 |
| **acı** | **4** |
| şehvet | 1 |
| şaşkınlık | 1 |
| kapılır | 1 |
| ızdırap | 2 |
| sinsi | 0 |
| çılgı | 2 |
| **yalnızlık** | **4** |
| iğrenç | 1 |
| **acımasızlık** | **4** |
| duygus | 2 |

| **ANGER** | **Num** |
|---|---|
| ***öfke*** | ***5*** |
| **kızgınlık** | **5** |
| **umutsuzluk** | **3** |
| şaşkınlık | 1 |
| **öfkes** | **5** |
| acı | 2 |
| üzüntü | 2 |
| kıskançlık | 1 |
| **çaresizlik** | **3** |
| **pişmanlık** | **3** |
| **hırs** | **3** |
| korkus | 2 |
| **suçluluk** | **4** |
| keder | 2 |
| **acımasız** | **3** |
| kapılarak | 0 |
| **dehşe** | **4** |
| kapılır | 1 |
| **kin** | **3** |
| sempat | 1 |
| kapılıp | 1 |

| SADNESS | Num |
|---|---|
| ***üzül*** | ***5*** |
| **üzülür** | **5** |
| **affeder** | **3** |
| **pişma** | **4** |
| **inandırır** | **3** |
| sevinir | 2 |
| sinirlenir | 1 |
| öfkelenir | 2 |
| **inanmaz** | **3** |
| **üzüle** | **5** |
| anlay | 2 |
| sinirle | 2 |
| **anlamaz** | **3** |
| angelica | 1 |
| setsuna | 1 |
| şaşırır | 1 |
| **şüphelenir** | **3** |
| okumuşturkimse | 0 |
| **hatırlıyor** | **4** |
| **sanmak** | **3** |
| söyley | 2 |

| DISGUST | Num |
|---|---|
| ***iğrenç*** | ***5*** |
| **ahlaksız** | **3** |
| şehvet | 1 |
| **gülünç** | **4** |
| aptal | 1 |
| korkak | 2 |
| **acayip** | **3** |
| delilik | 2 |
| **alaycı** | **3** |
| korkar | 2 |
| paranoyak | 2 |
| **utanma** | **4** |
| korkutuç | 1 |
| bencil | 2 |
| **hissettik** | **3** |
| **yapmacık** | **3** |
| **aşağıla** | **3** |
| **çıkarcı** | **3** |
| kibirli | 2 |
| **çirk** | **4** |
| ikiyüzlülük | 2 |

| SURPRISE | Num |
|---|---|
| ***şaşır*** | ***5*** |
| **sanır** | **5** |
| üzgü | 2 |
| **şaşırır** | **5** |
| korkmuş | 2 |
| sinirli | 2 |
| **uyandırır** | **3** |
| sinirlenir | 2 |
| **hissettik** | **4** |
| **aptal** | **3** |
| **anlamış** | **3** |
| düşünüp | 2 |
| **hisset** | **4** |
| kapılır | 0 |
| **korkar** | **3** |
| sinirlene | 1 |
| pişma | 0 |
| korkmak | 0 |
| kibirli | 2 |
| **hissediyor** | **3** |
| **söyley** | **4** |

*A7. Unanimously decided pivot and helper 10 words*

| HAPPINESS |
|---|
| mutluluk |
| sevgi |
| iyilik |
| özlem |
| duygu |
| tutku |
| gurur |
| hayaller |
| şefkat |
| heyecan |
| sevgis |

| FEAR |
|---|
| korku |
| dehşet |
| umutsuzluk |
| korkus |
| çaresizlik |
| vahşet |
| korkunç |
| zombi |
| acı |
| yalnızlık |
| acımasızlık |

| ANGER |
|---|
| öfke |
| kızgınlık |
| umutsuzluk |
| öfkes |
| çaresizlik |
| pişmanlık |
| hırs |
| suçluluk |
| acımasız |
| dehşe |
| kin |

| SADNESS |
|---|
| üzül |
| üzülür |
| affeder |
| pişma |
| inandırır |
| inanmaz |
| üzüle |
| anlamaz |
| şüphelenir |
| hatırlıyor |
| sanmak |

| DISGUST |
|---|
| iğrenç |
| ahlaksız |
| gülünç |
| acayip |
| alaycı |
| utanma |
| hissettik |
| yapmacık |
| aşağıla |
| çıkarcı |
| çirk |

| SURPRISE |
|---|
| şaşır |
| sanır |
| şaşırır |
| uyandırır |
| hissettik |
| aptal |
| anlamış |
| hisset |
| korkar |
| hissediyor |
| söyley |

### *A8. MI_Words List*

| **Happy** | **Fear** | **Anger** | **Sadness** | **Disgust** | **Surprise** |
|---|---|---|---|---|---|
| mutlu | kork | öfkelen | üzül | tiksin | şaşır |
| mutlu oluru | karanlık | sinirlen | vefat | koku | şaşırma |
| mutlu olu | korku | haksızlık | öl üzül | tiksindir | görün şaşır |
| mutlu oldu | korkut | yalan | üzüntü | kus | bekleme |
| mutlu olmuş | korku film | insan öfkelen | düşük not | insan tiksin | alınca şaşır |
| al mutlu | kal kork | sinir | dedem vefat | koku tiksin | al şaşır |
| oluru | kaybetmek kork | haksızlık uğrat | düşük | gör tiksin | şaşırtı |
| oldu mutlu | öd | uğrat öfkelen | et üzül | ter | sürpriz |
| mutlu edi | karanlık kork | insan | üzer | tükür | bekleme olay |
| oldu | film | el şaka | kay | böcek tiksin | şaş |
| sevindi | gece | haksız | kay üzül | tükürme | duyu şaşır |
| kazan mutlu | kovalama | yalan söylen | üzüt | görün tiksin | şaşırt |
| yüksek not | kal | yalan söylenme | üz | pis | bekleme not |
| alınca mutlu | yalnız | izinsiz | görün üzül | tükürülme | olma şaşır |
| kazan | yalnız kal | al | dedem | burun karıştır | öğren şaşır |
| mutlu et | yükseklik | hakkı yenme | öl | ağız | karşılaş şaşır |
| mutluluk | deprem | olma öfkelen | sınav düşük | böcek | öğrendik şaşır |
| vakit geçir | ev yalnız | haksızlık yapıl | oldu üzül | kusmuk gör | görün |
| al | baş | yapıl öfkelen | mutlu | mide bulan | gör şaşır |
| mutlu eder | köpek kork | sınav | ölme | kus gör | çıkın şaşır |
| sınav yüksek | köpek | söz | ölümüne üzül | kötü koku | bekleme davranış |
| gör mutlu | köpek kovalama | üzül | al üzül | tükür insan | doğum günü |
| olun mutlu | ev | söylenme | sev | kusma | hayret |
| yüksek | baş kal | et öfkelen | not | tuvalet | görün şaşırma |

| | | | | | |
|---|---|---|---|---|---|
| aile birlikte | gör kork | verme öfkelen | anneanne vefat | bulandır | ilginç |
| olu | böcek kork | kızar | yakın kay | koka | görme arkadaş |
| olma mutlu | tırıs | uğrat | yakın | mide bulandır | şaşkınlık |
| arkadaş vakit | gece ev | dinleme | not al | gör | öğrendik şaşırma |
| başarılı oldu | izlet korku | yap öfkelen | üzüm | yer | alma şaşır |
| görün mutlu | asansör | hakkı | ölü üzül | yıkama | açık kal |
| kazanı mutlu | yılan | yapma öfkelen | dedem öl | tükürük | beklenmedi k |
| yap mutlu | arkadaş | dalga | gel üzül | koka insan | karşı görün |
| geçir mutlu | cin | not | anneanne | kıl | şaşırtma |
| vakit | kaybetmek | gör öfkelen | arkadaş ayrıl | yeme | duyduk şaşır |
| doğ | örümcek | veril söz | ölüm | şap | gelme şaşır |
| ol mutlu | dişçi | oldu | gör üzül | kus tiksin | insan şaşır |
| olmuş | kal korku | eşya | ayrıl | el yıkama | al şaşırma |
| gün mutlu | kabus | suçlan | köpek öl | insan yer | gör şaşırma |
| sevin | al | söylenme öfkelen | acıt | mide | çıkma şaşır |
| geçir | gör korkut | yapılma öfkelen | yüksek beklet | sigara koku | umma |
| birlikte | köpek kova | davranış öfkelen | öğrendik üzül | insan | sürpriz şaşır |
| mutlu olacak | film izlet | hakaret | alınca üzül | kokma | karşı şaşır |
| başarılı | ev döner | emir | babam vefat | burun karıştırma | doğum |
| lise kazan | gelme kork | gereksiz | olma üzül | yeme kıl | günü |
| tatil | asansör kal | yalan söyleme | olun üzül | yiye | beklenmedi k olay |
| an mutlu | karanlık kal | söz tutulma | kork | burun | olun şaşır |
| çikolata yedi | yılan kork | haksızlık öfkelen | kuşu | pis koku | mutlu |
| huzurlu | insan | söylet öfkelen | al | ağız şap | karşılaş |
| gitti mutlu | elektrik kesil | yapma | ayrıl üzül | sümük | olay |
| hediye | izlet | alınma öfkelen | edin üzül | sigara | arkadaş |

| | | | | | |
|---|---|---|---|---|---|
| doğ mutlu | duyduk ses | geçil | kuşu öl | hamam | bekleme oldu |
| karşılık al | yüksek kork | eşya izinsiz | kanser oldu | hamam böcek | gelin şaşır |
| kucak al | deprem oldu | yapılma | kötü not | geğirme | bekleme kişi |
| mutlu hisse | kapalı alan | etme öfkelen | yap üzül | tiksindirici | öğrendik |
| iyi | kaza | saygısızlık | sevgili ayrıl | yiye insan | olay şaşır |
| insan | izlet kork | söz veri | hastalan üzül | al | gel şaşır |
| oyna mutlu | elektrik | dalga geçil | bende üzül | ölü gör | beklet sınav |
| geçir an | kaybetmek korku | iftira | anne | kusma tiksin | kaldırım |
| arkadaş geçir | kopar | haksız yer | başarısız oldu | okul tuvalet | tahmin etme |
| et mutlu | öd kopar | söylet | kanser | karıştır | alınca |
| sevil hisset | yakalan kork | öfkeli | gün üzül | temiz olma | oldu şaşır |
| aile | sev kaybetmek | sev | hüzünlen | bulan | sürpriz doğum |
| şaşır | ses | söz kesilme | arkadaş baba | yanım kus | verme şaşır |
| düşük beklet | gelecek korku | konuşma öfkelen | düşün üzül | böcek gör | beklet |
| kazanı | gerçekten kork | benden izinsiz | şehit haber | oldu | olma şaşırtı |
| an | aniden önüm | haklı | vefat üzül | sınav | kişi bekleme |
| iyi not | yap kork | öfke | acı | yer ağız | görme |
| genel mutlu | allah | mutlu | başarısız | iğrenç | üzül |
| ulaştı | not | bağır öfkelen | kal üzül | fare | şok |
| sev birlikte | büyük korku | söyleme | kötü | çıkın tiksin | karşı |
| not | oldu kork | not al | yakın vefat | ağız koku | yüksek |
| öğrendik mutlu | kal korkut | arkadaş yalan | üzgün | yanım kusma | süre görme |
| arkadaş birlikte | sınav kork | yalan söylet | ölümüne | tükürme tiksin | bekleme bekleme |
| öğren mutlu | aniden | yapılan haksızlık | ölme üzül | pis tuvalet | günü sürpriz |
| hediye al | üzül | çile | ara bozulma | pis koka | al |
| not al | hırsız | görün öfkelen | ayrılı üzül | yanım | not |

| | | | | | |
|---|---|---|---|---|---|
| üniversit e | yalnızlık | gör | babaanne vefat | tükür tiksin | ağız açık |
| güzel | ol korku | konu | düşün içim | çöp | şaşkın |
| birlikte oldu | rüya | nefret | düşün üz | fare tiksin | arkadaş sürpriz |
| kucak | film izle | söylen | gerçekten üzül | yer balgam | bekleme tepki |
| sınav | film iz | kızdır | üzüntü duyar | dışkı | değişim |
| doğum | gece karanlık | hal | kazan üzül | sakız | düşük beklet |
| doğum günü | al kork | kay öfkelen | babam | tükür gör | sürpriz yapma |
| istet | olma korkut | izinsiz eşya | kavga et | su iç | sınav bekleme |
| hava uçarı | yakalan | sıra bek | hayvan öl | terli | yap şaşır |
| müzik dinlemek | kaybol kork | arka | anneanne öl | saç çıkma | oldu öğrendik |
| fatma | oldu | isteme | cenaze | önüm kus | sınav |
| büyük mutluluk | mutlu | yapıl | durum üzül | geğir | bekleme sınav |
| sevindir | hırsız girdi | arkadaş | sınav | bardak su | şok oldu |
| uçarı | trafik kaza | şaka | kavga | kirli | sürpriz yap |
| yapın mutlu | olacak kork | söz tutma | şehit | çıktı tiksin | günü parti |
| çıktı mutlu | kaybol | tutulma | duyduk üzül | hoşlanma | mars |
| lise | korkma | adaletsizlik | ede üzül | pislik | umma kişi |
| şampiyo n oldu | karanlık korku | nefret edi | duyu üzül | yeme tiksin | parti |
| sev | öl kork | yüksek | başarısız olun | tuvalet koku | arkadaş görün |
| günü | öl korku | sürekli | sev vefat | pis insan | arkadaş karşılaş |
| iyi vakit | gece kork | konuşma | sınav kötü | pis ol | olay karşı |
| saygın | gör kabus | haksızlık yapılma | üzül üzül | kıl çıkın | olay oldu |
| yer mutlu | endişelen | nefret eder | derinden | erkek tiksin | sınav yüksek |
| uç | tedirgin | haksızlık karşı | düşüt üzül | yeme saç | yüksek not |
| başarı | yalnız baş | alma öfkelen | mutlu oldu | iç | not alınca |
| sevil | not al | anne küfür | hastalan | içilme | duyu |
| çikolata | gelecek | ayar | yıkıl | balgam | bek yüksek |
| birlikte ol | ameliyat | ayrımcılık | babaanne | karşı burun | alınca şaşırma |

| | | | | | |
|---|---|---|---|---|---|
| şampiyo n | gelen ses | davranma öfkelen | yenil üzül | tabak | tahmin |
| gel mutlu | görün kork | gelinme | ölüm haber | karıştırma | öğren |
| fatma saygın | görün korkut | gelme öfkelen | et | olma tiksin | uzun |
| güzel vakit | gir kork | düşük | kaybetmek | ağız şapırdatma k | tesadüf |
| mutluluk hava | korkutma | konuş | yüksek not | tükürülme tiksin | sürpriz yapı |
| mutluluk veri | olma kork | haksız durum | içim | şapırdatma k | alma |
| okul | karanlık yalnız | iş | hastalık | tırnak | arkadaş bekleme |
| insan birlikte | karanlık yerde | oluru | korkut | sigara içilme | sev |
| başarı mutlu | kaybetmek korkut | konuş insan | dedem ölüm | sokak yer | tepki |
| sınav iyi | havla | küfür | babam kay | bardak | sürpriz ziyaret |
| aile arkadaş | yürü kork | tutma öfkelen | bekleme | sinek | hamile oldu |
| aile mutlu | korkutan | geçilme | oluru | öfkelen | yapılan sürpriz |
| aile vakit | kaza geçir | arkadaş arka | yakın kaybetmek | tuvalet gir | bekleme yerde |
| alma mutlu | girdi kork | aptal | kaybetme | temiz | bekleme insan |
| bende mutlu | geçir kork | dalga geçme | insan üzül | tiksinti | insan bekleme |
| doğ sevindi | film kork | gereksiz yer | yakın arkadaş | tuvalet pis | ziyaret |
| kardeş doğ | kapalı | saygı | hastalanma | saç çıkın | yakışıklı |
| hava | şimşek | edil öfkelen | maç kay | iğrenç koku | beklem |
| okul kazan | kay kork | sinirli | değer | çatal | şaşı |
| maç kazan | elektrik kesilme | ısrarla | hayal kırıklık | insan ağız | tembel |
| birlikte mutlu | yılan gör | korkut | kırıklık | kaşık | sürpriz yapıl |
| ilk | kaybetme korku | bencil | yüksek | yumurta | farklı |
| yeğen | aniden karşı | öğrendik öfkelen | değer verme | karıştır gör | düşük bek |
| tatil mutlu | ameliyathan e | konuşu | sev kay | hapşırma | emin oldu |

| | | | | | |
|---|---|---|---|---|---|
| gör | ışık | arka iş | mülteci | temizle | yer |
| üniversit e kazan | ıssız | haksızlık uğra | çalış | sifon | aile |
| sev mutlu | araba | dinlenme | sınav başarısız | saç | davranış |
| yerde para | böcek | yer suçlan | yer | ağız şapırdatma | yüksek al |
| duyu mutlu | düşük | yalan sö | insan | şapırdatma | sonuç |
| al sevindi | ses duyduk | veril | aile | yağlı | yüksek alınca |
| kitap oku | kova | hoşlanma | taşınma | yedi | alakasız |
| yeğen doğ | yürü | laf | üzücü | not | bek |
| sevinç | takip | duyulma | çocuk sahip | el | bekleme hareket |
| hisset mutlu | kaza yap | umursamaz | canım | ölü | gel |
| al sevin | rüya gör | öğretmen | sınav yüksek | içil | yap sürpriz |
| doğum gün | ani hareket | dalga geçilme | çalış sınav | kereviz | çıkma |
| edi | dab | arka konuşulma | tartış | sev | günü hediye |
| sınav al | kovalan | etme | ala | yanım sigara | beklenme |
| sürpriz | kilitli | bağır | karşı | hayvan ölü | dışında |
| özgür | tıkırtı | verdi söz | yalan | bulaşık | şehir dışında |
| oldu an | tırsma | aile | anne babam | ağız açık | davranış gör |
| aile geçir | trafik | bekleme | hasta | tiksinme | yıl |
| eder | iz | yüksek not | yenil | anne | sandık |
| gitti | önce | uğra | düşük al | adam | yıl görme |
| arkadaş buluş | açıklan | dinleme öfkelen | dedem kay | sokak | bek düşük |
| yeni | ani | haklı oldu | anne vefat | fare gör | sevgili oldu |
| şampiyo n olma | ev hırsız | insan nefret | arkadaş vefat | arkadaş burun | şaşma |
| başlat mutlu | insan kaybetmek | haklı haksız | büyük anne | önüm kusma | tutul |
| kal | baş gelecek | inatla | sağlık sorun | ıslak | karşı gör |
| kazanma | başka korkma | küp | soma | yol | yanlış çıkma |
| eğlen | zarar gelme | bekletil | yokluk | sigara içme | kapıda |
| kötü | ilk kez | sorumsuz | kimsesiz | yüksek | |
| yaz | girdi | suçlanma | pişmanlık | sofra | |

| | | | | | |
|---|---|---|---|---|---|
| gün | aile kaybetmek | yırt | kır | toplu | |
| yer | havlama | salak | ölmüş | organ | |
| seçil | karşı geçer | kural uyma | sev insan | içme | |
| kavuş | gider | edilme | doğum günü | istifra | |
| karşı | disipline | saygısız | karşılık ala | pırasa | |
| geçit mutlu | atla | egoist insan | değer verdi | mayonez | |
| beşiktaş | cinli | düşük not | bozul | umumi | |
| müzik | şiddetli | edil | ağla | sümkürme | |
| elde et | bekleme | davran insan | çalış hal | solucan | |
| alınca | kez | ciddi alınma | ölen | sarımsak | |
| düşük not | döner | küfür edil | ev | ezilmiş | |
| yaz tatil | yer | verme | intihar | korku | |
| yedi mutlu | kazan | hak | küs | hayvan | |
| üniversit e kazanı | et | köpek | düşün | temizlik | |
| kardeş olacak | arı | egoist | haksızlık | koku duyduk | |
| huzur | iğne | yer | hal düşük | çorba | |
| mutlu hisset | sınav önce | davran | zarar görme | yenme | |
| hisset | kişi | konuşulma | arkadaş küs | oluru | |
| başar | sınav girme | davranma | ayrı kal | nefret edi | |
| olacak | ameliyat oldu | hakkında | beslet | dikkat etme | |
| yüksek al | atlat | davranıl | sinirlen | aile | |
| kitap | düşük not | iftira atılma | arkadaş kavga | yap | |
| oyna | araba kaza | verilme | şampiyonlu k | dokun | |
| yarışma | kesil | alınma | zorun | su | |
| kay | öğrendik | görün | haber | kişisel | |
| sarıl | önüm | mantıksız | ayrılı | ses çıkart | |
| günü sürpriz | geç saat | dağıt | ayrı | salyangoz | |
| spor yap | sınav gir | sorul | film | temizleme | |
| et | doğum | hareket | gör | tuvalet girdi | |
| korkut | fren | otobüs | verdi | açık | |
| korku | gölge | benden habersiz | hayal | bağırsak | |

| | | | | | |
|---|---|---|---|---|---|
| maç kazanma | bayılma | sorumluluk | kırıl | bakımsız | |
| insan vakit | palyaço | geçir | hasta oldu | eti | |
| şaşırma | karabasan | kişi | üzülme | kokulu | |
| maç | görün | yapmacık | taşın | izmarit | |
| buluş | akşam | kural | doğum | hijyen | |
| olun | doğal afet | kullanılma | evcil | hapşu | |
| köpek | sessizce | oldu hal | muhabbet kuşu | mantar | |
| bekleme | köpek gör | yap | aile kavga | öpme | |
| başarılı ol | uçak | rahatsız etme | özlü | temizlenme | |
| okul kazanı | yüksek | getirme | teşhis | tebeşir | |
| şampiyo n olun | günü | bildik hal | yapma | yıkanma | |
| anadolu | iyi | okul | çocuk | kal | |
| sağlıklı | baş gelme | davranış | elenme | babam | |
| galat | istet | kandırıl | felç | mutfak | |
| yakın | uyu | kardeş kavga | ede | kurbağa | |
| kazan an | harf | benden | olu | çorap | |
| kazan öğrendik | aile baş | karşı | kırılma | taşıma araç | |
| film | afet | dağınıklık | evcil hayvan | çıkma | |
| beklet sınav | kız | durum düş | zarar | dışkı gör | |
| alışveriş yap | kapı | alay | fener şampiyonlu k | arkadaş üst | |
| hedef | aşı | bencil insan | günü | not al | |
| bulduk | çekirge | haksız oldu | kaplumbağa | domates | |
| yılbaşı | olu | küfür et | sorun | sigara iç | |
| oyun oyna | kalp | soru sorma | bozulma | yola | |
| iltifat | karşı karşı | günü | | görüntü | |
| yardım et | etki | şiddet | | sürüngen | |
| sınav bit | | başka | | an | |
| insan mutlu | | sö | | deri | |
| eğlenceli | | beklet | | peçete | |
| oyun | | saçma | | sümüklü | |
| ders boş | | doğum günü | | toplu taşıma | |

| | | | | | |
|---|---|---|---|---|---|
| arkadaş aile | | kardeş | | yakın | |
| mutlu olma | | ukala | | öğrendik | |
| hisse | | alınca | | otobüs | |
| alındı | | kız | | arkadaş | |
| koku | | arka konuş | | yapmacık | |
| beklet | | kal | | ölü hayvan | |
| kut | | boş konuş | | et | |
| sınav düşük | | uyma | | çöp atan | |
| karşılık | | habersiz | | çıkaran | |
| okul tatil | | görmez | | | |
| hobi | | bozar | | | |
| sevinçli | | döven | | | |
| düşük | | çıldırt | | | |
| başarılı olun | | uyandırılma k | | | |
| takım | | toplama | | | |
| boş oldu | | tas | | | |
| istet oldu | | sorumsuzluk | | | |
| sev yeme | | söyleni | | | |
| turnuva | | taklit | | | |
| basketbo l | | kardeş oda | | | |
| yalnız | | irk | | | |
| takdir | | kavga | | | |
| beşiktaş şampiyo n | | başkası | | | |
| gitar | | olu | | | |
| beklet yüksek | | atılma | | | |
| ağız | | çalınma | | | |
| cuma | | oda | | | |
| galat şampiyo n | | arka konuşma | | | |
| kitap okur | | karar | | | |
| sev insan | | doğum | | | |
| bilgisaya r | | yapılan | | | |
| | | bağırma | | | |
| | | verdi | | | |

| | | | | | |
|---|---|---|---|---|---|
| | | ede | | | |
| | | kaba | | | |
| | | sorma | | | |
| | | tutma | | | |
| | | film | | | |
| | | ait | | | |
| | | yapma<br>gerek | | | |
| | | ilk | | | |
| | | gece | | | |

***A9. Classification Methods***

***1. Naive Bayes***

Naive Bayes classifier is a classification method based on Bayes Theorem. It is known as one of the simplest and fastest algorithms. In this classifier, properties / inputs are assumed to be independent of each other and all properties are linked to the classification label. Based on this assumption, the product of the conditional probability score is calculated for each class, and the sample is assigned to the relevant class. Details about the Naive Bayes classifier used in this study are given in John and Langley (1995).

$$\boldsymbol{P(A|B)} = \frac{\mathbf{P}\boldsymbol{(B|A)P(A)}}{\boldsymbol{P(B)}}$$

***2. BayesNet (Bayesian Network)***

A Bayesian network is a probable graphical model (a kind of statistical model) that represents a number of variables and conditional dependencies through a directed acyclic graph (DAG). Bayesian networks are ideal for capturing an incident and predicting the probability of any of the few possible known causes to be contributing factors. For example, a Bayesian network may represent probability relationships between diseases and symptoms. When symptoms are given, the network can be used to calculate the probability of the presence of various diseases. Effective algorithms can perform inference and learning in Bayes networks (Wikipedia, 2020).

### *3. SMO (Sequential Minimal Optimization)*

Sequential minimal optimization is an algorithm for solving the quadratic programming problem that occurs during the training of support vector machines. This algorithm was invented in 1998 by John Platt at Microsoft Research. Sequential minimal optimization (SMO) is widely used to train support vector machines and is implemented by the popular LIBSVM tool (Wikipedia, 2020).

### *4. Random Forest*

Random Forest algorithm is a Supervised classification algorithm. As the name implies, the algorithm simply creates a forest randomly. There is a direct relationship between the number of trees in the algorithm and the results they can achieve. As the number of trees increases, we get a definite result. One of the most important advantages of the Random Forest algorithm is that it can be used in both classification and regression problems (Çebi, 2018).

### *5. J48 (C4.5)*

J48 classifier is the implementation of C4.5 decision tree in WEKA tool. Although C4.5 tree is similar to ID3 decision tree (Iterative Dichotomiser 3), the main difference between them is that ID3 uses information gain while C4.5 uses gain ratio (Quinlan, 1993). On the other hand, different from ID3 tree C4.5 tree can be pruned. In C.4.5 algorithm, in order to create a new branch of the tree, the gain ratio value of the node is considered. After that, a sub-list is created under the new decision node and sub-decision trees are constructed. The unnecessary/useless branches are removed by pruning in order to narrow down the decision space by decreasing the size of tree.