DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

INVESTIGATION OF TEXT MINING METHODS ON TURKISH TEXT

by
Ezgi PASİN

September, 2018
İZMİR

INVESTIGATION OF TEXT MINING METHODS ON TURKISH TEXT

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Master of Science in
Statistics, Statistics Program

by

Ezgi PASİN

September, 2018

İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"INVESTIGATION OF TEXT MINING METHODS ON TURKISH TEXT"** completed by **EZGİ PASİN** under supervision of **ASSOC. PROF. DR. SEDAT ÇAPAR** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Assoc. Prof. Dr. Sedat ÇAPAR

Supervisor

Assoc.Prof.Dr. Emel Kuruoğlu

(Jury Member)

Assoc.Prof.Dr. Hidayet Savaş SAZLI

(Jury Member)

Prof.Dr. Kadriye ERTEKİN

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGEMENT

I would like to express my gratitude to my advisor Assoc. Prof. Dr. Sedat ÇAPAR for his patience, motivation and support throughout my thesis.

Besides my advisor, I would like to thank Assoc. Prof. Dr. Emel KURUOĞLU KANDEMİR for her support and motivation.

Finally, I want to thank to my father Tuncay PASİN, my mother Hatice PASİN and my sister Başak PASİN for their endless love, support and patience during my thesis.

Ezgi PASİN

INVESTIGATION OF TEXT MINING METHODS ON TURKISH TEXT

ABSTRACT

Today, with the widespread use of the internet the size and value of the data have increased. Making meaningful information from large amounts of data is one step ahead for others and for companies. Various mining techniques must be applied to obtain meaningful information from the data. Data mining processes on structured data. Text mining is the sub-study area of data mining that works on texts. Text mining is the process of analyzing text to extract valuable information from text for special purposes.

Before the text mining techniques are applied, the data must be prepared and pre-processed. Extracting meaningful information from texts, classifying text, and reaching the desired information in a short time increase the importance of text mining. Text classification is the process of deciding the class of the given text using the training documents for the predefined categories.

The aim in thesis is to classify Turkish data as text. The categories were examined in three categories as “Gender Identification”, “Author Identification” and “Species Determination”. Naive Bayes method and the bit-score weighting k-NN method were used for classification. The accuracy rates of the two methods are compared. The R programming language is used for classification. In this thesis, a dataset consisting of Turkish columns was created to work on text classification.

Keywords: Text mining, k-NN algorithms, Naive Bayes method, data mining, unstructured data, text categorization

TÜRKÇE METİNLER ÜZERİNDE METİN MADENCİLİĞİ YÖNTEMLERİNİN İNCELENMESİ

ÖZ

Günümüzde internetin yaygın kullanılması ile birlikte verilerin boyutu ve değeri artmıştır. Büyük miktarlardaki verilerden anlamlı bilgi çıkarmak kişi ve firmalar için diğerlerinden bir adım öne geçmektir. Verilerden anlamlı bilgilerin elde edilmesi için çeşitli madencilik teknikleri uygulanmalıdır. Veri madenciliği yapısal veriler üzerinde işlem yapar. Veri madenciliğinin, metinler üzerinde çalışan alt çalışma alanı ise metin madenciliğidir. Metin madenciliği, özel amaçlar için metinlerden değerli bilgiler çıkarmak adına metnin analiz edilmesi işlemidir.

Metin madenciliği teknikleri uygulanmadan önce verilerin hazırlanması ve ön işlemden geçirilmesi gerekmektedir. Metinlerden anlamlı bilgi çıkarmak, metni sınıflandırmak, aranan bilgiye kısa sürede ulaşmak metin madenciliğinin önemini arttırmıştır. Metin sınıflandırma, önceden tanımlanmış kategorilere eğitim dokümanlarını kullanarak verilen metnin sınıfına karar vermesi işlemidir.

Tezde amaç, metin halindeki Türkçe verilerin sınıflandırılmasıdır. Kategoriler "Cinsiyet Tanımlama", "Yazar Tanımlama" ve "Tür Belirleme" olmak üzere üç kategoride incelenmiştir. Sınıflandırma yapılırken Naive Bayes metot ve bit skor ağırlıklandırılmış k-NN metot kullanılmıştır. İki metodun doğruluk oranları kıyaslanmıştır. Sınıflandırma için R programlama dili kullanılmıştır. Bu tezde metin sınıflandırılması üzerine çalışmak için Türkçe köşe yazılarından oluşan bir veri kümesi oluşturulmuştur.

Anahtar kelimeler: Metin madenciliği, k-NN algoritmaları, Naive Bayes metodu, veri madenciliği, yapılandırılmamış veri, metin sınıflandırma

CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER ONE
INTRODUCTION

In recent years with the inevitable increase at the use of internet technologies, people use the web for many reasons like entertainment, personal communication, online shopping and so on. Along with the widespread use of the internet, non-structural data in the virtual environment has increased the amount of data (Uytun, 2013). They do not mean anything on their own. However, this data becomes meaningful when it is processed for a certain purpose.

Information is a processed data for a purpose. It is not possible to make decisions based on "raw data" or "information", which is just a visualization of what happened in the past. It is also not possible to prevent loss caused by a past bad experience. It is important to discover confidential information about past events and to make forward-looking predictions (İnan, 2003). Therefore, it is very important to be able to use techniques to analyze large amounts of data. The data mining concept is encountered when it is desired to obtain meaningful information that cannot be estimated.

Data Mining, began to be used to solve computer data analysis problems in the 1960s. Data mining is the removal of large quantities of data, very unclear, previously unknown but potentially useful information (Han et al., 2006). It is possible to use data mining anywhere with large amounts of data. Nowadays, many field data mining applications are widely used. For example, marketing, biology, banking, insurance, stock exchange, medicine, science and engineering, industry and so on. successful applications are seen in many branches. Data mining has been used in different fields and sub-studies such as Text Mining, Web Mining, Temporal Data Mining and Spatial Data Mining have been created.

Text mining can be broadly defined as a knowledge-intensive process in which a user interacts with a document collection over time by using a suite of analysis tools. Text mining is the process of analyzing texts to extract important information from

texts for specific purposes. Text mining seeks to extract useful information from data sources through the identification and exploration of interesting patterns (Feldman & Sanger, 2007). The techniques such as information extraction, summarization, classification, clustering and information visualization are text mining techniques.

Text classification is the classification of natural language texts according to predetermined categories. The number of text classes may vary according to the size or the text of the text (Rana et al., 2014). Nowadays, it is possible to see many practically applied areas such as text classification, indexing of documents based on a controlled vocabulary, filtering of documents, automatic metadata creation, automatic hierarchical arrangement of web pages (Fabrizio, 2002). When text categorization is used correctly, the text will assign the most accurate class that may be possible.

Text categorization techniques which are used in text type data classification and various weighting methods are examined in Pilavcılar's study (Pilavcılar, 2007). In the study, k-NN and Naive Bayes classification techniques were used and the success rates between them were compared. As a result of the application, the success rate of Naive Bayes was found to be higher.

Bai designed an algorithm that uses the Markov model classifier in this study in 2011 (Bai, 2011). In this way, the relationships between the words are defined and the performance of the classification method is increased. He achieved a 92.7% classification success in his laboratory study.

The aim of this study is to find the proper class of Turkish columns. The bit-score weighting k-NN method and Naive Bayes method is used and applied on a real data set. 800 Turkish columns were used as dataset. Turkish requires different text processing techniques from English and similar languages. Various processes such as separating the roots of the words, removing punctuation marks were done for the categorization of columns.

The main purpose of the study is to transform the unstructured columns into structured data and analyze the transformed data with the k-NN method and Naive Bayes method and make the correct classification using R.

CHAPTER TWO
DATA MINING

The structures of databases have changed and more data has begun to be kept in these databases with the development of technology. As well as processing the data into the database, it has become important to examine them in the database and to extract meaningful information. Statistical or mathematical analyzes are performed on the data in the databases with different operations and different purposes. In fact, one of the analysis techniques is data mining.

Nowadays, institutions produce large amounts of data but they have difficulties in revealing meaningful and useful information within these data. It is not easy to solve large-scale data with traditional statistical methods. For this reason, special methods are needed to process and analyze the data. Data mining methods have emerged to meet this need (Özkan, 2008).

Data mining has emerged conceptually in the 1960s when computers began to be used to solve data analysis problems. With the 1960s and 1970s, software was started to be programmed and developed for database management. Although the data mining emerged in the 1970s, these software's mostly came to the forefront in the 1980s and 1990s (Grad & Bergin, 2009).

Data mining is the exploration and analysis of large quantities of data in order to discover valid, novel, potentially useful, and ultimately understandable patterns in data (Kalikov, 2006). The application area of these processes is quite wide. In these areas, there are disciplines such as data base systems, data visualization, artificial neural networks, statistics, artificial learning, as shown in Figure 2.1 (Savaş et al., 2012).

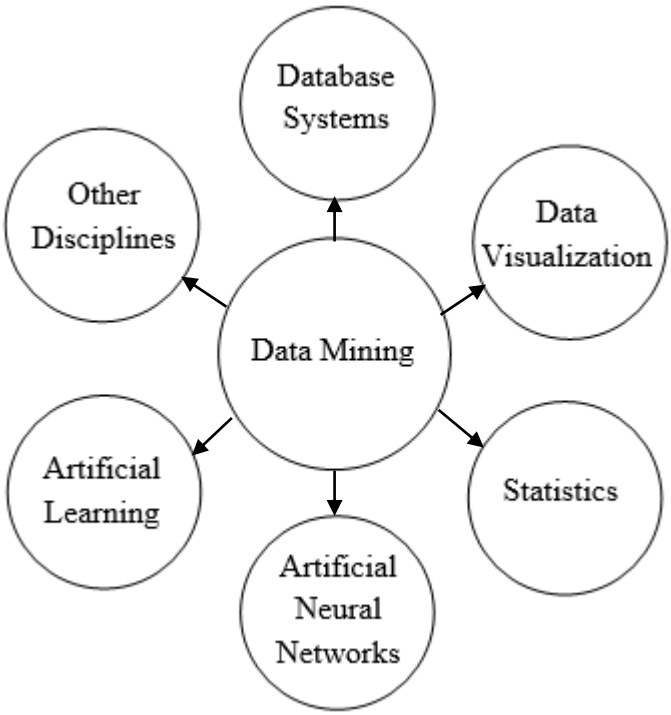

Figure 2.1 Data mining and disciplines

2.1 Application Areas of Data Mining

Data mining techniques have been applied successfully in many areas from business to science and sports. Data mining has been used in database marketing, retail data analysis, stock selection, credit approval, etc. Data mining techniques have been used in astronomy, molecular biology, medicine, geology, and many more fields. It has also been used in health care management, tax fraud detection, money laundering monitoring, and even sports (Sumathi & Sivanandam, 2006).

Güllüoğlu conducted a preliminary study to diagnose cancer in 2011. Data mining techniques have been used in Güllüoğlu's study. Cancer types have been identified in the study and an early detection approach has been explored for early detection of cancer, which may be used for people who have been or are likely to be infected using data mining (Güllüoğlu, 2011).

2.2 Database in Data Mining

A system that works quickly and accurately in small data clusters can behave completely differently when applied to very large databases. While a data mining system works well on consistent data, it may not produce successful results when applying different data. The reason for this is the difficulties in the data mining process.

2.2.1 Database Size

The database size is rapidly increasing. Many algorithms developed have been developed to be applicable to small databases. In order to be able to use the same algorithm in large samples, some issues need to be considered. The important thing is to choose the correct method. Data mining is a method that gives meaningful results on large data.

2.2.2 Null Values

Many datasets encounter a null value problem. This problem may have been due to incomplete data entry, incorrect measurements, hardware errors. Each null value is represented by "NULL", "*", and "?". Empty values can be removed by deleting or filling (Cios et al., 2007).

2.2.3 Missing Data

The dataset used in the data mining a value may be unknown or missing. In data mining methods, missing data is problematic in analysis because each data is important. For this reason, missing data must be filled by various methods.

2.2.4 Noisy Data

The noisy is an out-of-system error during data entry or data collection. If the dataset is noisy, the system should recognize the broken data and ignore them. Today's databases are trying to correction as automatically correct errors that occur during data entry. Incorrect data can cause serious problems in real databases (Çalışkan, 2006).

2.3 Data Mining Processes

When a data set is first created, it may contain various properties that are not required in the data set. It is subjected to some operations in order to obtain correct results in a data set. For this, the data mining process takes place in four basic stages: data collection, data conversion, data mining and interpretation of results.

The steps of the data mining process are shown in Figure 2.2.

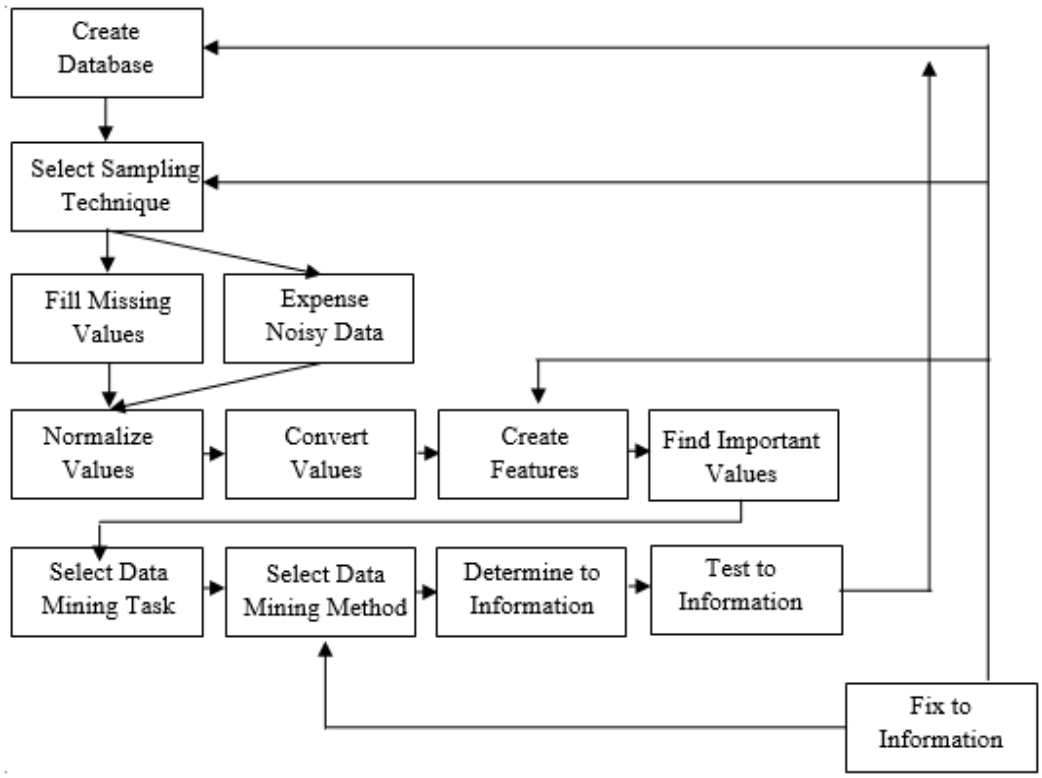

Figure 2.2 Data mining process

2.3.1 Data Selection

The first step in data mining is data selection. Every value in the dataset may not be meaningful. For example, in a market analysis, when customers want to examine products or services that customers take, they might include information such as demographics or lifestyles of customers in the data set but they are unnecessary and should not be included to the analysis. Therefore, the data selected from the data set is very important.

2.3.2 Data Cleansing and Conversion

The second step of data mining is data cleansing and conversion. After the desired database tables are selected for the data mining and the data to be mined is defined, corrections must be made on the data. Data transformation is the process of organizing data into the desired format and transforming the data from one type to another.

The purpose of data conversion is to transform the source data into different formats or values. For example, a logical field in the database can be converted to an integer type. The reason for this is that some of the data mining algorithms used are more successful in this way (Tang & Maclennan, 2005).

The purpose of the data cleansing process is to remove inappropriate or incorrectly entered data in the data. In this process, the missing values are filled with the appropriate values. If there is a lot of missing data, this record should be deleted.

2.3.3 Data Mining

The third step is data mining. The data mining phase is the mining of transformed data with one or more techniques to obtain the desired information.

2.3.4 Interpretation of Results

The final step is to interpret the results. In data mining, finding the results is the half of the work. Interpretation of the analysis made is the most important step of the study. The result interpretation phase is the analysis of the metallized information (Sumathi & Sivanandam, 2006).

2.4 Data Mining Methods

Data mining is a very common a method in recent years. Significant results on large datasets have become important both for the person and the company. As a result, data mining has begun to be used in many areas and methods have been developed. Methods of data mining is classification clustering and association rule.

Classification is a frequently used procedure in everyday life. With classification, objects are divided and separated and objects can be assigned as classes, either private or general categories (Bramer, 2007). Classification rules are established by using a part of the existing database as training data set. With the help of these rules, it is determined how to decide when a new situation arises.

The most frequently used method of data mining from the classification group is decision trees. At the same time, logistic regression, discriminant analysis, neural networks and fuzzy sets are used.

The classification process of the data consists of two steps. First, a model is set up for the data sets. It is used as a part of training data from the database to obtain the classification model. This data is randomly selected from the database. In the second step, the classification rules are specified on the test data. then the rules are applied to the test data.

The second method is clustering. Clustering is the process of grouping data based on similarities between themselves. The clustering analysis divides the information

in a data set into groups according to certain proximity criteria. Each of these groups is called a "cluster." In the clustering process, the similarity between the elements in the cluster should be high and the similarity between the clusters should be small (Sarıman, 2011).

Clustering analysis has been developed to elaborate the classification of individuals or objects. It divides the analyzed samples into groups according to their similarities and analyzes them. Another goal is to reduce the data set by grouping similar elements (Tan et al., 2006).

The third method is the association rule. The association is an interesting issue that has been studied extensively in the field of data mining. Association is the method used to determine relations between objects. It is one of the first techniques used in data mining and the mathematical model of the association rule was presented by Agrawal, Imielinski and Swami in 1993 (Agrawal et al., 1993).

The most typical example of the association rule is the market basket application. This process solves the purchasing habits of the customers by finding the relationship between the customer's purchases. The discovery of these types of associations reveals the knowledge of what products customers receive together and market managers can increase sales rates by identifying shelf layouts with this information (Döşlü, 2008).

2.5 Fields Related to Data Mining

It is possible to use data mining anywhere with large data. As a result, the use of data mining has increased in recent times. People or companies use many areas for a specific purpose. There are same areas of direct relevance with data mining. Machine learning and artificial intelligence are some of them.

Machine learning is the general name of computer algorithms that model a problem according to the problem belonging to it. The data set and the model created

by the algorithm used are set up to give the highest performance. For this reason, many machine learning methods have been developed. Some of those; k-nearest neighbors algorithm, Naive Bayes classifier, decision trees, logistic regression analysis, k-average algorithms, support vector machines and artificial neural networks (Atalay & Çelik, 2017). Machine learning uses two methods, supervised and unsupervised learning.

In the supervised learning method, it is expected that the desired result set is obtained from the input data. To train the machine in the supervised learning method, a training set and outputs for the values are given. Using this training data, the machine learns to use in the future. In the future, if you want something on the machine, look at the training set you have given and and compare the desired data with the similar data in the training set and making inferences.

Unlike supervised learning, unsupervised learning does not give an exit value. Unsupervised learning does not include any training data. Unsupervised learning aims to learn the structure of the data.

Artificial intelligence (AI) systems are designed to adapt and learn. Artificial intelligence models the human brain. It is used in different areas of daily life and is also used for purposes such as prediction, classification, clustering. Techniques such as expert systems, genetic algorithms, fuzzy logic, artificial neural networks, machine learning are generally referred to as artificial intelligence technologies (Atalay & Çelik, 2017).

CHAPTER THREE
TEXT MINING

With the rapid increase in internet usage and the widespread use of personal computers, a growing number of document stacks are emerging. A significant number of these datasets contain unprocessed and unaltered data in non-structural form. Texts, photos, videos, audio files are some of these data. It is very difficult to analyze such data. For this reason, techniques and methods in the field of Text Mining, which is the sub-field of Data Mining, are used to analyze the data sets in text form (Kılınç et al., 2016).

Text mining studies are often accompanied by natural language processing, which is another work area in text-based literature. Natural language processing studies are mostly based on linguistic knowledge under artificial intelligence. Text mining studies are mostly aimed at reaching the results statistically via text (Seker, 2015). The purpose of text mining is to manipulate unstructured information, extract meaningful numerical contents from the text, and thus access the information contained in the text for various data mining algorithms (Melek, 2012).

Text mining is a method that facilitates our work in many areas. For example, a researcher trying to find a medication for a disease needs to quickly examine the work done before and reach out to the different concepts in the documents (Güven, 2007).

3.1 Data Structures

Generally, data describing facts, ideas, and concepts that can be collected, transmitted, counted, and processed are very important for every research. Data word comes from the Latin root of the word. The information is processed into meaningful information that is presented to the user. The data are used in two ways as structured data and unstructured data.

3.1.1 Structured Data

Structured data is data that has been organized into a formatted repository, typically a database, so that its elements can be made addressable for more effective processing and analysis. Structured data is information, usually text files, displayed in titled columns and rows which can easily be ordered and processed by data mining tools. This could be visualized as a perfectly organized filing cabinet where everything is identified, labeled and easy to access. Table 3.1 is an example of structured data.

Table 3.1 An example of structured data

Year	**Population**	**Annual Increase (%)**
1990	56,473,653	2.29
2000	67,804,543	2
2007	70,586,256	0.58
2008	71,517,100	1.31
2009	72,561,312	1.48
2010	73,722,988	1.6
2011	74,724,269	1.35
2012	75,627,384	1.2
2013	76,667,864	1.37
2014	77,695,904	1.34
2015	78,741,053	1.34

3.1.2 Unstructured Data

90% of the world's data are in unstructured format. Unstructured data refers to data that follows a form that is less ordered than items like spreadsheet pages, database tables or other linear or ordered data sets. For example, email is a fine illustration of unstructured textual data (Weiss et al., 2010).

One of the most common types of unstructured data is text. Unstructured text is generated and collected in a wide range of forms, including Word documents, PowerPoint presentations, survey responses, and posts from blogs and social media sites.

Other types of unstructured data include images, audio and video files. Like texts, we often encounter pictures, sounds and videos in our daily lives. The widespread use of the internet has caused me to encounter this data more often.

3.2 Application Areas of Text Mining

Text mining, used in different areas such as national security and company security practices, legal-lawyer-legal cases, corporate finance, patent analysis, public relations, web pages comparison of enterprises.

In a questionnaire study, questions can be prepared in an open-ended manner to express the views and opinions of respondents without a specific constraint. Thus, previously unfamiliar customer opinions and ideas can be obtained from the structured questions designed by the persons. An internet page or a large document can be scanned, and we can automatically extract the list of terms found. The most important features or terms that have been defined can be identified (Melek, 2012).

In market research, published documents, press releases and web pages are inculcated with text mining techniques to measure market impact. Also, text mining is used in the evaluation of open-ended questions and interviews. In customer relationship management, qualified information is extracted from the text information obtained from the access points of all customers such as email, transaction, call center and questionnaire. This qualified information is used to estimate the sales of the customer (Dolgun et al., 2009).

Another common practice of text mining is to help automatically classify texts. For example, it is possible to automatically filter the message from this mail by

specifying certain words or phrases that may pass through in an unwanted unnecessary mail. Or, as with most automated systems, electronic messages can be categorized and directed to agencies that want messages to go or to departments that need to be diverted. An example of this is an automatic message sent from a company that has applied for more than one job posting to only those who apply for an illegal job.

3.3 Comparison of Data Mining and Text Mining

Data mining essentially analyze figures that may be described as homogeneous and universal. Text mining tools have to face major technical challenges such as heterogeneous document formats (text documents, emails, social media posts, etc.). Unlike data mining, text mining works with unstructured or semi-structured collections of text documents. These differences are shown in Table 3.2.

Table 3.2 Comparison of data mining and text mining

	Data Mining	**Text Mining**
Data Object	Numerical & Categorical	Textual data
Data Structure	Structured data are used.	Unstructured data are used.
Maturity	The widespread use began in 1990.	The widespread use began in 2000.

Data mining and text mining are somewhat similar. Because it works with very large datasets in both and aims to extract meaningful information from the information stack.

3.4 Steps of the Text Mining

Text mining is a process that transforms texts obtained in a certain period of time using meaningful analysis tools into meaningful information. In the first stage, texts

are collected from specific databases or search engines. The resulting text is pre-processed to remove it from meaningless words and finally the results obtained using data mining techniques are evaluated. The steps taken in the text mining process are explained under the heading five (Oğuz et al., 2009). These steps are shown in Figure 3.1.

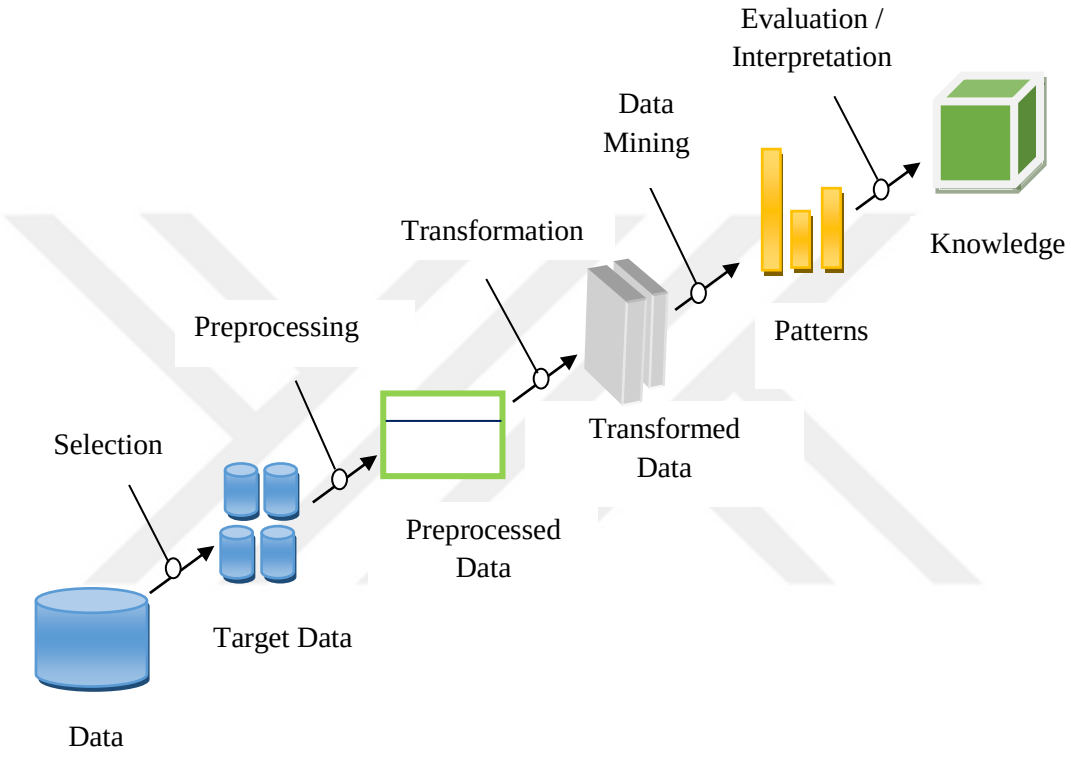

Figure 3.1 Text mining steps

3.4.1 Creating a Text Collection

The first step of the text mining steps is to create a text collection. It is the process of using information access systems that exist in the topics of interest. Nowadays, text collection is mostly done using search engines like google, yahoo, etc. In addition, data collected in different databases or personal computers can also be used in text mining.

3.4.2 Text Preprocessing

The second step is text preprocessing. The text collection of the previous step can contain complex and unnecessary information. For this reason, a number of operations are performed in order to provide convenience for large amounts of data. In this step, the text is divided into words, the roots of the words are found and the unnecessary words are deleted. In addition, the semantic values of these words are found and their conformity to the writing rules is made at this step. Ineffective words do not make sense for information extraction and classification, for example; consists of words such as “ve (and)”, “veya (or)”, “biraz (some)”, “hepsi (all)” in the Turkish language. Frequently used words in Turkish have been removed from the text.

3.4.3 Feature Selection

The third step is the step of feature selection. This phase is the most important phase of the text mining process. In this phase, the important words in the texts are identified and the unrelated features are extracted. At the end of this step, the unstructured data is transformed to structured data.

3.4.4 Data Mining

The fourth step is the step of data mining. Data mining is a technique used to obtain the desired and valuable knowledge for research from large scale data. Nowadays, the increase of the data in the databases has necessitated many studies to analyze the data. Data mining deals with the analysis of numerical data in databases using various statistical and analytical methods and interpretation of the results obtained.

Data mining is actually very similar to classical statistical applications. However, classical statistical applications are run on organized and summary data. Data mining works with millions or even billions of data. As the number of data was large, some special analysis algorithms had to be developed and started to do work for it. Data

mining is used in many areas such as banking, finance, marketing, communications, astronomy, biology, medicine.

Estimation of main expenditure groups' portion with data mining methods are examined in Ahi's study. An implementation of statistical mapping was performed using data mining methods to predict spending variables for the 12 main spending groups in the study (Ahi, 2015).

3.4.5 Evaluation and Interpretation

The fifth and the last step of the text mining steps is the step of evaluation and interpretation. In this step, the analysis made is evaluated and interpreted.

3.5 Techniques Used in Text Mining

To teach computers how to analyze, understand and generate text, technologies are produced by natural language processing. The technologies like information extraction, summarization, categorization, clustering and information visualization, are used in the text mining process (Gaikwad et al., 2014).

3.5.1 Information Extraction

Information extraction refers to the automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from unstructured sources (Sarawagi, 2008). Information extraction is usually aimed at obtaining information on specific criteria using natural language processing on a text. The goal is to create a software that automatically processes large amounts of data and minimize human interference. The environment in which information can be extracted is usually written texts, but the environment in which these texts are found may change. For example, databases, documents on the internet, or scanned text can create the source of this data.

3.5.2 Categorization

Categorization is the assignment of normal language documents to predefined set of topics according to their content. It is a collection of text documents, the process of finding the accurate topic or topics for each document (Dang & Ahmad, 2014). Statistical classification techniques like Naive Bayes probability method, k nearest neighbor algorithm, decision trees, artificial neural networks and genetic algorithms can be used to categorize text.

In the classification, for example, there are three target groups such as high income, middle income and low income, or a target variable such as income category which can be categorized. The data mining model examines large sets of each record containing information on the target variables as in the input or predictor variable set (Lorese, 2005).

3.5.3 Clustering

Clustering allows to classify the units studied in a research by gathering them within certain groups according to their similarities, to reveal the common features of the units and to make general definitions about these classes. The general purpose of the clustering analysis is to group the ungrouped data according to their similarities and to help the researcher obtain useful, useful summarized information (Dinler, 2014).

Classes are predefined in the categorization, but clustering analyzes the data without a specific class. In general, the classes are not readily available in training data because they are not known at the beginning. Class number can be increased when analyzing in clustering. The data are classified or grouped by maximizing similarity within the classes.

3.5.4 Visualization

The human brain processes visual information better than it processes text so using charts, graphs, and design elements, data visualization can help you explain trends and stats much more easily (Ergün, 2017). Visualization is a mental image or a visual representation of an object or scene or person or abstraction that is similar to visual perception.

Visualization has many definitions but the most referred one, which is found in literature is "the use of computer-supported, interactive, visual representations of data to amplify cognition", where cognition means the power of human perception or in simple words the acquisition or use of knowledge (Teyseyre & Campo, 2009).

3.5.5 Summarization

Text summarization is the process of automatically creating a compressed version of a given text that provides useful information for the user. In big organization or company, researcher do not have time to read all documents, so they summarize document and highlight summary with main points (Gupta & Lehal, 2009). The summary is a metric generated from one or more texts containing a significant portion of the information, reducing the length and keeping the general meaning as if it were in the original texts.

3.6 Fields Related to Text Mining

Today, text mining is used in many areas. Some of the areas related to text mining are explained.

3.6.1 Natural Language Processing

It is the biggest factor in the communication between people. Natural language processing aims to use language factor most effectively between human and

computer. The purpose of natural language processing is to provide natural language communication with the computer, so the computer must learn the natural language rules (Delibaş, 2008). For this, the computer needs a general dictionary and various algorithms to use it.

The natural language processing system generally has five basic elements. These are parser, dictionary, apprehend, knowledge base and producer. Figure 3.2 shows the interaction of these five elements (Delibaş, 2008).

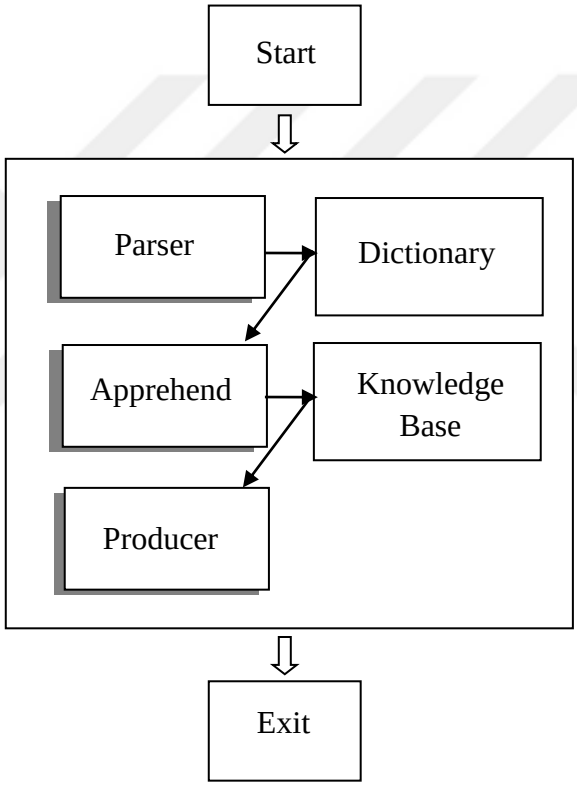

Figure 3.2 General diagram of natural language processing systems

The parser is the most basic element of natural language processing. The parser analyzes the given sentence syntactically. After parser, the roots of the words are determined. these words are subjected to a semantic analysis process to generate an output sequence according to the input sequence.

The dictionary is a document containing all the words that the program needs to be recognized. The parser works by doing syntactic analysis with the dictionary. The

dictionary consists of the roots and meanings of each lamb that are recognized by the natural language processing system. If taking into account the number of words in dictionaries used in natural language processing systems, it takes can a lot of time to create a dictionary.

The apprehend tries to identify the meaning of the clan together with the knowledge base. The main task of the apprehend is to find out the knowledge base of the generated parser tree.

The aim of natural language processing is to analyze and understand the canonical structure of natural languages (Bahadır, 2007). natural language processing is used in many applications such as text translation, text summarization, text categorization.

3.6.2 Information Retrieval Systems

In our daily lives, computers have become important places, leaving paper-based information place to digital environment. Transfer of information has become faster thanks to removable disks or social networks. In this way, knowledge has accumulated on the internet and it has become difficult to reach valuable information (Oğuz, 2009). Information retrieval systems have been developed to enable people to access the right information in a short period of time.

Information retrieval systems are used especially in academic and specialized areas. It is mostly used in medical field. Information retrieval systems work with the user's query. It compares the terms in the user's queries with the terms in the user's queries and it conveys the documents that the same terms pass through to the user (Onur, 2007). Search engines such as Google Yahoo are examples of systems developed for general purposes.

When users search in search engines or databases, there are a number of ways to get better results. Keyword selection should be done correctly and the writing of the words should be done correctly.

3.6.3 Information Extraction Systems

Information extraction inference systems are one of the effective approaches to text mining, which converts unstructured text into structured form (Gaizauskas, 2002). These systems deal with documents written in natural language. In texts, it tries to extract useful information such as person, place or time names. Information extraction systems aim to extract structured knowledge from unstructured text.

3.6.4 Question Answering Systems

Question answering systems are systems that respond directly to the question asked by the user. it does not need to list page addresses for person to find an answer. These systems are one of the types of information retrieval systems and it seems to be one step ahead of current search engine technologies (Erhardr et al., 2002).

Figure 2.3 shows the general diagram of the question answering systems (Waheeb & Babu, 2017). The system first registers and analyzes the user query. Then, according to the question, it determines the answer types (e.g. date, place, name etc.). At the same time using an information retrieval system, it accesses documents containing the words of the questionnaire.

The answers are searched in documents determined by information access systems and these answers are scored and ranked. High-scoring answers are presented as a convenient interface to the user. The difference between information access systems of question answering systems; While information retrieval systems presenting the document list to the user, question answering systems are presented as answers to phrases or words.

The general diagram shows of the survey answer system in Figure 3.3.

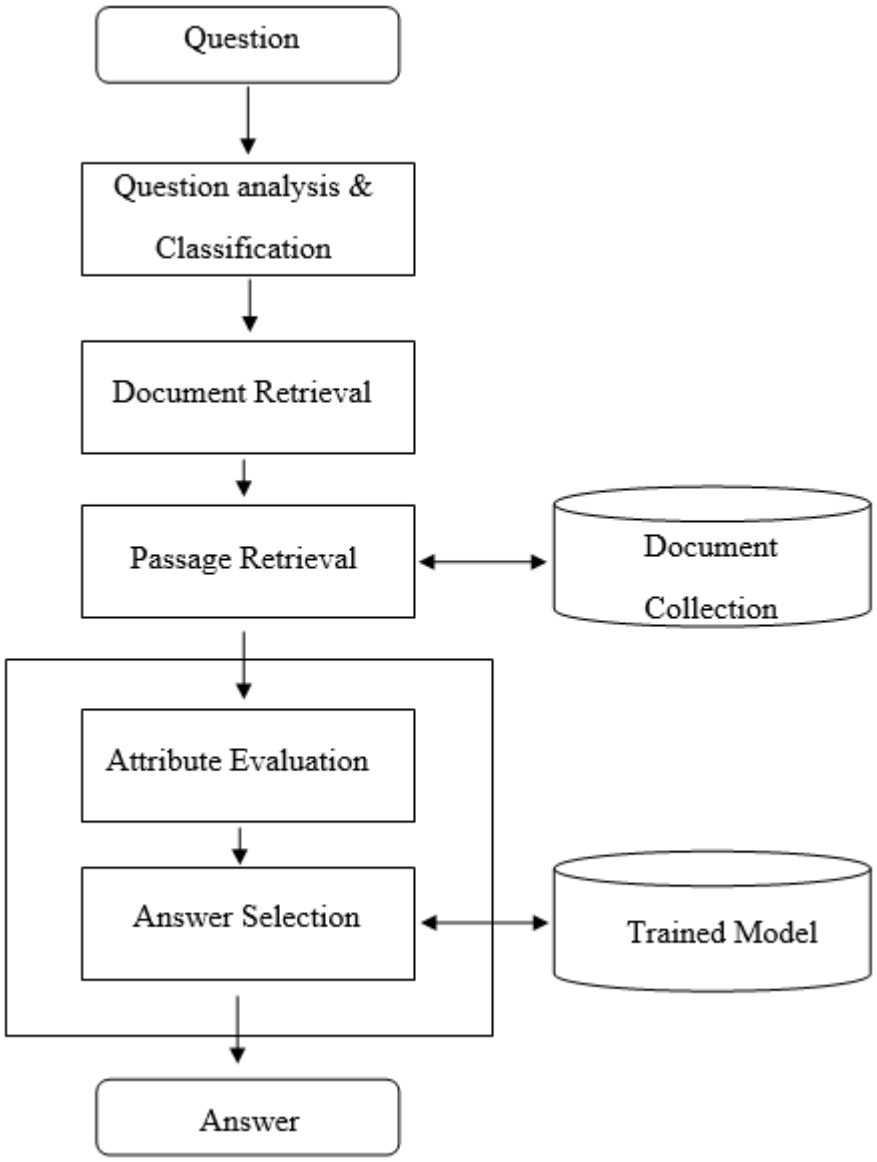

Figure 3.3 General diagram of question answering system

The most important areas related to text mining shows in Figure 3.4. Today, text mining is used in many fields and in the future, the areas of use will increase more. The most common used in these areas are data mining, statistics and natural language processing.

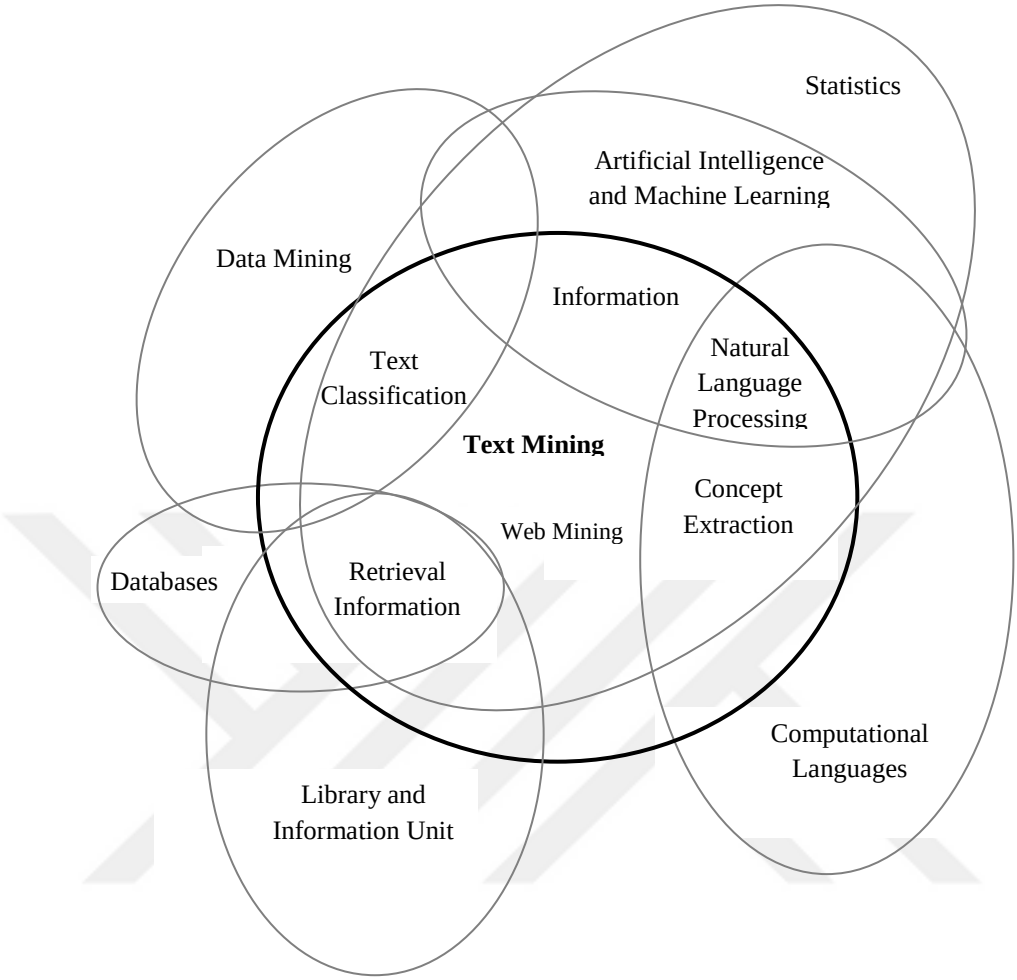

Figure 3.4 Text mining and related areas

CHAPTER FOUR
CATEGORIZATION TECHNIQUES

The purpose of text categorization is to classify the documents into a fixed number of predetermined categories. Each document may belong to one categorical, multiple categorization, and may not belong to any category. There are many text classification methods. In this thesis, k-NN method and Naive Bayes method are used to classify the text.

4.1 K-NN Algorithm

One of the categorization algorithms is k nearest neighbor algorithm. The nearest neighbor algorithm is a learning algorithm that is a result of classifying the query vector with the nearest k neighbor vector (Koyuncugil & Özgülbaş, 2009). The purpose of the k Nearest Neighbor (k-NN) algorithm is to use a database in which the data points are separated into several separate classes to predict the classification of a new sample point.

The classification rules are generated by the training samples themselves without any additional data. The k-NN classification algorithm predicts the test sample's category according to the k training samples which are the nearest neighbors to the test sample and decide it to that category which has the largest category probability (Lihua et al., 2006). Classification process of k-NN algorithm:

- Assume that there are training documents belonging to categories *C* is from 1 to *j*.
- *d* is the vector lengths of the test data from 1 to *m*.
- *q* is test data.
- Calculate the similarities between all training samples and *q*. Taking the *i*th sample di as an example, the similarity sim (*q*, *di*) is as Equation 4.1 (Suguna & Thanushkodi, 2010).

$$sim(q, di) = \frac{\sum_{j=1}^{m} q_j \cdot d_{ij}}{\sqrt{(\sum_{j=1}^{m} q_i)^2}\sqrt{(\sum_{j=1}^{m} d_{ij})^2}} \quad (4.1)$$

As a result, the k-NN classifier selects the class with the greatest probability as the result of Equality 4.1.

The k nearest neighbor algorithm is applicability is a simple algorithm and it is effective if the number of training documents is high. There are also disadvantages such as the advantages of the nearest neighborhood algorithm. For example, k parameter is needed, and it is not clear which distance type and what quality to use to get the best results.

As seen in Table 4.1, for some test specimens the correct estimate is made when the k value is increased, decreased or taken at certain intervals. In order to increase the classification success, it is necessary to know that a proper k value should be selected instead of a constant k value for each test case. A large selection of k does not increase the likelihood of success (Bulut & Amasyalı, 2014).

Table 4.1 Successful impact of k value

	k parameter for k-NN									
	1	2	3	4	5	6	7	8	9	10
Example 1	0	0	0	0	0	0	0	1	1	1
Example 2	1	1	1	1	1	1	1	1	1	1
Example 3	0	0	0	0	0	0	0	0	0	0
Example 4	0	0	1	1	1	1	1	1	1	1
Example 5	0	0	0	1	1	1	0	0	0	1
Example 6	1	1	1	1	1	1	1	1	0	0
Example 7	1	1	1	1	1	0	1	1	1	1
Example 8	0	0	0	1	0	0	0	0	0	0
The number of correct predictions	3	3	4	6	5	4	4	5	4	5

The nearest neighbor algorithm helps to find which of the k classes the data belongs to. For example, the ones shown with a square show a class and the ones shown with a circle show the other class in Figure 4.1. We need to find out which group belongs to which is indicated by the star.

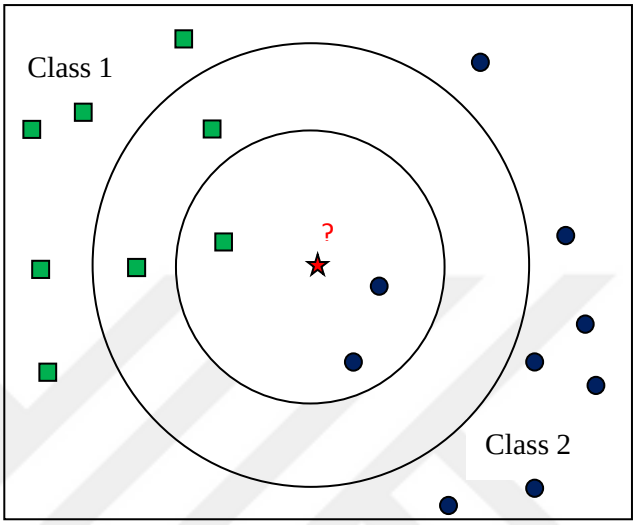

Figure 4.1 Classification example

Assume that d1, d2 and d3 are train vectors belonging to different groups as seen in Figure 4.2. q is the vector we want to find the class (Pilavcılar, 2007).

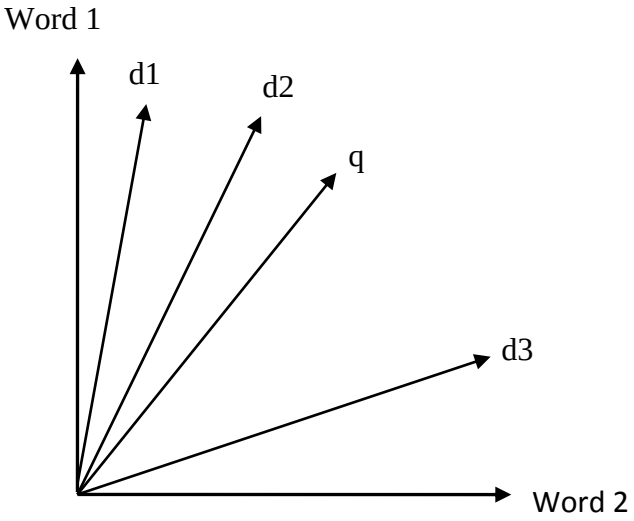

Figure 4.2 Text classification

Three kinds of methods can be used for the term weighting in the k-NN algorithm as bit-score weighted k-NN method, frequency weighted k-NN method and TF-IDF weighted k-NN method.

4.1.1 Bit-Score Weighted k-NN Method

Bit-score weighted k-NN method, the weights of the vector are determined by whether or not they are in the document. If a word goes through the document twice, it will not change this statement. It will be expressed as one if the word is in the text, as zero otherwise. For each word in the dictionary, it is checked whether or not the document has passed through and the weights are found.

The bit-score weighted k-NN method uses the Equation 4.1 formula. the data is assigned to the largest of the probabilities computed as the result of the equality.

For example: The answers of the six students to the exam are shown in Table 4.2. It will be expressed as one if students answer correctly of questions, as zero otherwise. According to the answers given by the students, they have success or fail.

Table 4.2 An example of k-NN method

Questions							
Students	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Status**
S_1	0	1	0	1	0	0	Unsuccessful
S_2	0	0	0	1	0	0	Unsuccessful
S_3	0	0	0	1	1	0	Successful
S_4	0	0	0	0	0	1	Unsuccessful
S_5	0	0	1	0	0	0	Successful
S_6	1	0	0	0	0	0	Unsuccessful

Suppose a student answers the questions $S_t = (0, 0, 1, 0, 1, 0)$. If we calculate the success according to the answer given by this student, we must use Equality 4.1.

$$sim(s_1, s_t) = 0,\ sim(s_2, s_t) = 0,\ sim(s_3, s_t) = \frac{1}{\sqrt{2}*\sqrt{2}} = 0.5,\ sim(s_4, s_t) = 0,$$

$$sim(s_5, s_t) = \frac{1}{1*\sqrt{2}} = 0.707\ ,\ sim(s_6, s_t) = 0$$

The greatest similarity should be chosen. The similarity values for the third and fifth documents are different from zero. The greatest value is the fifth document. Because the success of the fifth document is successful, we can say that it is also successful in the case of this student.

4.1.2 Frequency Weighted k-NN Method

Frequency Weighted k-NN Method, calculates the weights of the vectors according to the number of documents found. The number of words in the dictionary is important. For example, if a word is found twice in the document, the weight is two.

4.1.3 Term Frequency – Inverse Document Frequency Weighted k-NN Method

Term frequency inverse document frequency (TF-IDF) documents are one of the most important weighting methods that we can use to define in the vector space model. TF-IDF method determines the relative frequency of words in a specific document through an inverse proportion of the word over the entire document corpus (Trstenjak et al., 2013). This value is calculated from all training documents. So, if a word is not common in the document, it creates a defining feature.

TF-IDF is generally used to find the similarity ratio between the test vector and the training vector. Various versions of the TF-IDF function are available. The TF-IDF function can be calculated from the Equation 4.2 and Equation 4.3 (Soucy & Mineau, 2005).

$$W_{d,t} = tf_{d,t} \cdot idf_t \tag{4.2}$$

$$idf_t = log \frac{D}{dft} \tag{4.3}$$

where,

tf is the frequency of the words found in the dictionary. *D* is total number of documents and *dft* is document frequency (number of documents passed through the word).

4.2 Naive Bayes Probability Method

The Naive Bayes classification method is one of the most successful algorithms used to classify text documents. Its applicability is an easy algorithm and is also successful in terms of performance. Naive Bayes is a classification technique based on Bayes' theorem.

Assume that *D* represents the data set and that each class of *X* in *D* is known. *X* is a vector whose length is the number of words in the dictionary and $X = (x_1, x_2, \ldots, x_n)$ is expressed. Suppose you have *m* classes represented by $C_1, C_2, \ldots, C_m$. Naive Bayes is used to find the C_i class that an *X* vector belongs to. It tries to find the highest $P(C_i \mid X)$ probability in all classes. This probability is found in the Equation 4.4.

$$P(C_i|X) = \frac{P(X \mid C_i) \cdot P(C_i)}{P(X)} \tag{4.4}$$

Since the value of *P(X)* is the same for all classes, only the expression $P(X \mid C_i).P(C_i)$ must be computed and maximized. $P(C_i)$ is the ratio of the number of elements in class *Ci* to the number of all elements. $P(X \mid C_i)$ is calculated by

Equation 4.5, assuming that *X* is an attribute vector containing n values (Kaynar et al., 2017).

$$P(S \mid C_i) = \prod_{k=1}^{n} P(X_k \mid C_i) \quad (4.5)$$

As a result, the Naive Bayes classifier selects the class C_i with the largest $P(X \mid C_i) . P(C_i)$ expression as the class of the *X* vector.

For example: Information is given of 14 days in Table 4.3. Data were given weather forecast, temperature, humidity rate, wind and playing tennis for 14 days (Burak, 2017).

Table 4.3 An example of Naive Bayes

Days	**Weather Forecast**	**Temperature**	**Humidity Rate**	**Wind**	**Do you play tennis?**
Day 1	Sunny	Hot	High	Weak	No
Day 2	Sunny	Hot	High	Strong	No
Day 3	Cloudy	Hot	High	Weak	Yes
Day 4	Rainy	Warm	High	Weak	Yes
Day5	Rainy	Cold	Normal	Weak	Yes
Day 6	Rainy	Cold	Normal	Strong	No
Day 7	Cloudy	Cold	Normal	Strong	Yes
Day 8	Sunny	Warm	High	Weak	No
Day 9	Sunny	Cold	Normal	Weak	Yes
Day 10	Rainy	Warm	Normal	Weak	Yes
Day11	Sunny	Warm	Normal	Strong	Yes
Day 12	Cloudy	Warm	High	Strong	Yes
Day 13	Cloudy	Hot	Normal	Weak	Yes
Day 14	Rainy	Warm	High	Strong	No

The tennis playing situation of the person is calculated according to the weather on the fifteenth day.

The conditions required on day 15 are:

Weather Forecast: “Sunny”

Temperature: “Cold”

Humidity Rate: “High”

Gale Force: “Strong”

Do you play tennis?: “Yes”

Let's write the total numbers first.

Total number of data: 14, total number of yes: 9, total number of no: 5

We must calculate the probability of “Yes” and “No” from Equation 4.5.

Probability of yes = $P(Y) = \frac{9}{14} \; x \; \frac{2}{9} \; x \; \frac{3}{9} \; x \; \frac{3}{9} \; x \; \frac{3}{9} \; = \; 0.0053$

Probability of no = $P(N) = \frac{5}{14} \; x \; \frac{3}{5} \; x \; \frac{1}{5} \; x \; \frac{4}{5} \; x \; \frac{3}{5} \; = \; 0.0205$

Naive Bayes probability method is calculated for "Yes" and "No" class. This method accepts the highest possible probability. Under these conditions the person will not play tennis because it is more likely not to play tennis.

4.3 Decision Trees

Decision trees are a classification algorithm that has been widely used in the literature in recent years. The most important reason for the widespread use of this method is understandable and simple. Decision trees use a multistage or sequential approach while classification (Kavzoğlu & Çölkesen, 2010).

Classification of the data using decision tree technique is a two-step as learning and classification. In the learning step, a training data which is known beforehand is

analyzed by a classification algorithm in order to create a model. The learned model is shown as a classification rule or decision tree. In the classification step, test data is used to determine the correctness of classification rules or decision tree (Çalış et al., 2014).

The general diagram of the decision tree is shown in Figure 4.3. Decision trees are structures that resemble flow charts. Branches and leaves are elements of a tree structure. The last structure is called "leaf", the top structure is called "root" and the structures among them are called "branch".

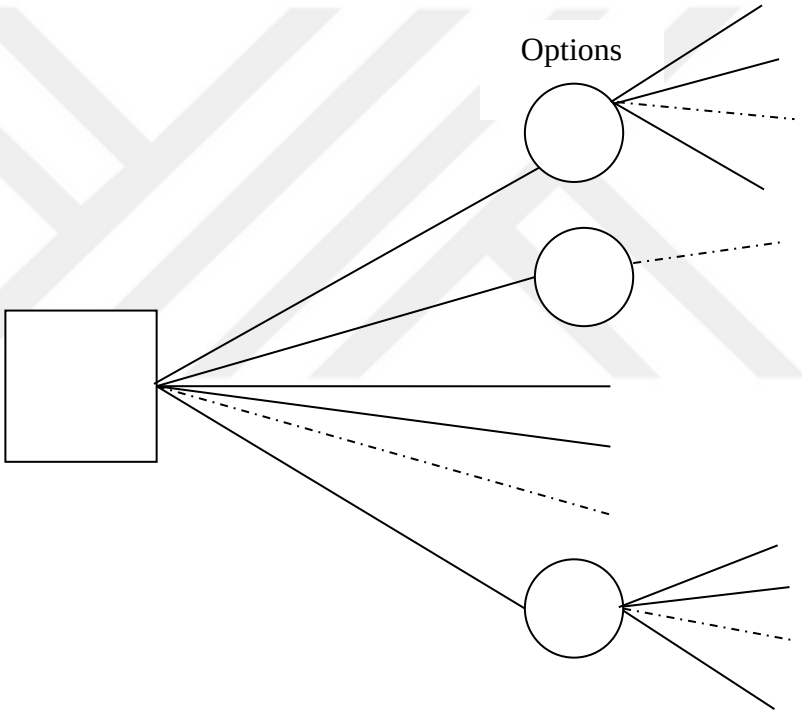

Figure 4.3 General diagram of decision tree

Decision tree algorithm can be improved by dividing the data set into small pieces. A decision tree can consist of both categorical and numerical data. A decision node may contain one or more branches and branching increases as options increase. The most important thing to note is the correct selection of the branches to which the options depend when the decision tree is created.

For example: Information is given of 14 days in Table 4.4. Data were given weather forecast, temperature, humidity rate, wind and playing tennis for 14 days.

Table 4.4 An example of decision tree

Days	**Weather Forecast**	**Temperature**	**Humidity Rate**	**Wind**	**Do you play tennis?**
Day 1	Rainy	Hot	High	No	No
Day 2	Rainy	Hot	High	Yes	No
Day 3	Cloudy	Hot	High	No	Yes
Day 4	Sunny	Warm	High	No	Yes
Day5	Sunny	Cold	Normal	No	Yes
Day 6	Sunny	Cold	Normal	Yes	No
Day 7	Cloudy	Cold	Normal	Yes	Yes
Day 8	Rainy	Warm	High	No	No
Day 9	Rainy	Cold	Normal	No	Yes
Day 10	Sunny	Warm	Normal	No	Yes
Day11	Rainy	Warm	Normal	No	Yes
Day 12	Cloudy	Warm	High	Yes	Yes
Day 13	Cloudy	Hot	Normal	No	Yes
Day 14	Sunny	Warm	High	Yes	No

The decision tree for the tennis player status is shown in Figure 4.4. There are three weather conditions: “Sunny”, “Cloudy” and “Rainy”. When the weather is cloudy, the person will play tennis. As a result, the temperature and humidity conditions have not been considered.

The situation that makes a difference to play tennis in the sunny and rainy weather conditions is looked at. When the weather is sunny, the person will play tennis

according to the wind conditions. If there is wind, tennis will play but if not, it will play.

Third weather condition is rainy. When the weather is rainy, the humidity status is important. If the weather is rainy, when the humidity is high, the person will not be able to play tennis, but when the humidity is normal the person will be able to play tennis.

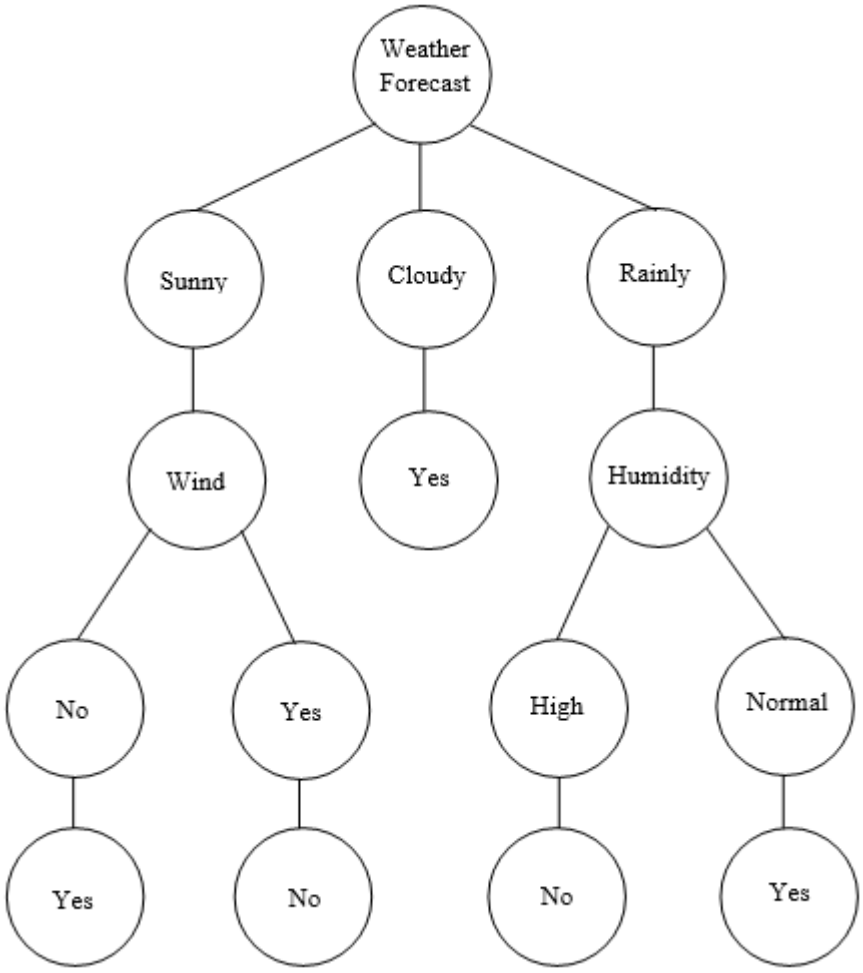

Figure 4.4 Decision tree for weather forecast

4.4 Artificial Neural Networks

Artificial neural networks can collect information about samples, generalize, and then compare them with samples that they have never seen before. Artificial neural networks are applied in many scientific fields due to these learning and

generalization features. At the same time, it can successfully solve complex problems (Ergezer et al., 2003).

Artificial neural networks are composed of artificial cells that are hierarchically connected to each other and can work to parallel. The basic task of an artificial neural network is to specify an output set that can correspond to an input set. In order to do this, the network is trained with examples of the event and be able to generalize. It specifies the output sets when it encounters similar events with this generalization (Altaş & Gülpınar, 2012).

General diagram of artificial neural networks is given in Figure 4.5.

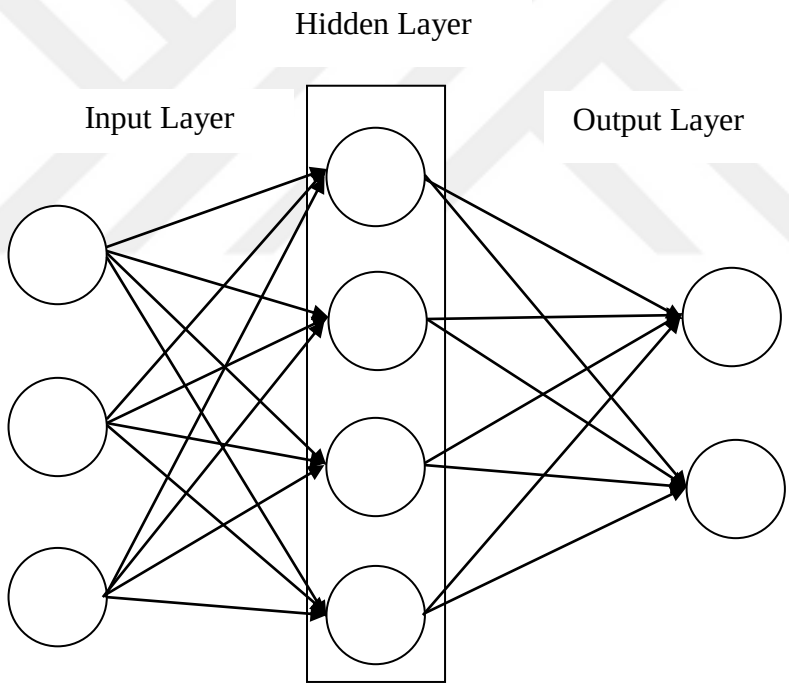

Figure 4.5 General diagram of artificial neural networks

4.5 Genetic Algorithms

In the 1970s, the basic principles of the Genetic Algorithm (GA) introduced by John Holland have been successfully applied in many types of problems. The Genetic Algorithm is an intuitive algorithm used to find approximate results in the

optimization or search problem. This algorithm was developed based on techniques in evolutionary biology such as inheritance, mutation, selection and crossover (Haznedar et al., 2016).

CHAPTER FIVE

APPLICATION

In this section, the classification of Turkish texts has been made using the k-NN method and the Naive Bayes probability method. Information about the database and comments on the results of the analysis are included.

5.1 Data Set

In this study, a collection of texts consisting of columns (a total of 800 columns belonging to 50 days) from Turkish newspaper companies was created. These 800 columns are our unstructured data to be categorized. The classification was examined in three categories as “Gender Identification”, “Author Identification” and “Species Determination”.

The first classification is gender identification. In this classification, columns are examined in two classes, "Female" and "Male". The authors are shown in Table 5.1 along with their gender. There are 10 columns of every gender. Columnists are abbreviated as the initials of their first and last names. Male columnists are designated as “E. S.”, “N. D.”, “M. V.”, “N. L.”, “S. K.”, “Y. A.”, “E. Ç.”, “Y. Ö.”, “A. Ç.” and “R. D.”. Female columnists are designated as “C. O”, “J. Ö.”, “E.S.Ö.”, “İ. Ç.”, “E.Ö.”, “T. D.”, “A. S.”, “Z. G.”, “E. K.” and “Ş. T.”.

The second classification is author identification. The author identification is divided into 20 classes because there are 20 columnists. Columns of these authors are shown in Table 5.1.

The third classification is species determination. The species determination has been divided into five categories, “Ekonomi (Economy)”, “Sağlık (Health)”, “Magazin (Magazine)”, “Siyasi (Political)” and “Spor (Sports)”. The frequency table of the news headings belonging to the defined groups is shown in Table 5.1.

Table 5.1 Numbers of columnist' columns

Columnists and Columns			
Male		**Female**	
Categories and columnists	**Total number of columns**	**Categories and columnists**	**Total number of columns**
Economy (E)		**Economy (E)**	
E. S.	40	C. O.	40
N. D.	40	J. Ö.	40
Health (H)		**Health (H)**	
M. V.	40	E. S. Ö.	40
N. L.	40	İ. Ç.	40
Magazine (M)		**Magazine (M)**	
S. K.	40	E. Ö.	40
Y. A.	40	T. D.	40
Political (P)		**Political (P)**	
E. Ç.	40	A. S.	40
Y. Ö.	40	Z. G.	40
Sport (S)		**Sport (S)**	
A. Ç.	40	E. K.	40
R. D.	40	Ş. T.	40

In order to be able to classify, we must first determine the training data. Training data is the set of data known to which group the data belongs to. In this study, training data for each category are defined as 20 columns. Training data set includes 400 observations.

Text preprocessing step applied to the data set in which the text collection was created. First, all words in the text are converted to lower case. Punctuation marks and frequently used words in Turkish language were removed from the text to ensure ease of operation. For example, “ile (with)”, “ve (and)”, “veya (or)”. Then the words are divided into their roots. The semantic values of the words in the text are found and the words that may belong to the category are added to the dictionary.

5.2 Dictionary of Model

This section lists all words that may be class of texts in our database. The number of words in the dictionary is approximately 320. It is used a joker-named method to separate the roots and their suffixes in Pilavcılar's study. A joker is a word that starts with the same word and has various suffixes but collects words that are close in meaning with a single representation (Pilavcılar, 2007). Each word in the dictionary is examined and words that may be jokers are found and added to the dictionary.

5.3 Methods Used in Application

In application, classification is made using bit-score weighting k-NN method and Naive Bayes method. The accuracy rates of these two methods are compared.

5.3.1 Application of K-NN Method

In application, bit-score weighted k-NN method was used first. The k-NN algorithm is a learning algorithm that results in the classification of the test vector with the nearest k neighbor vector. In this study, k values have been tested individually. The most successful result was found when k was 3. Therefore, the value of k is 3 in the study. The bit scored weighted k-NN method deals with the state of the word being in the text.

For example: In the third part of the application, the columns are classified according to their species. Let's explain this classification with a short example. A sentence belonging to the columns was taken and these sentences are shown in Table 5.2. The training data is the column known to belong to the category.

Table 5.2 Example of training data

Training Data	**Vector Length**											
Dictionary	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
E1 = yıllık enflasyon oranı bu sene de yükselişte	0	0	0	0	0	1	0	0	0	0	0	0
E2 = kırmızı et fiyatlarında müthiş düşüş	0	0	0	0	0	0	1	0	0	0	0	0
H1 = herkes bu şifalı bitkiyi konuşuyor	0	1	0	0	0	0	0	0	0	0	0	1
H2 = sağlık için yararlı bitkiler	0	1	0	0	0	0	0	0	0	0	1	0
M1 = ünlü şarkıcının yasak aşkı	1	0	0	0	0	0	0	0	0	0	0	0
M2 = diziden reyting rekoru	0	0	0	1	0	0	0	0	0	0	0	0
P1 = başkanı herkes ayakta alkışladı	0	0	1	0	0	0	0	0	0	0	0	0
P2 = partiden şok istifa	0	0	0	0	0	0	0	0	0	1	0	0
S1 = hakemin gözü önünde olmasına rağmen görmedi	0	0	0	0	0	0	0	0	1	0	0	0
S2 = gol atan futbolcunun büyük sevinci	0	0	0	0	0	0	0	1	0	0	0	0

The dictionary consists of 320 words but give an example with 320 words is difficult. So, suppose that our example dictionary includes 12 words.

Example Dictionary: "aşk=1, bitki=2, başkan=3, dizi=4, dolar=5, enflasyon=6, fiyat=7, futbolcu=8, hakem=9, parti=10, sağlık=11, şifa=12”

Example of test data= "taraftarlar hakeme büyük tepki gösterdi."

When creating a dictionary, the words that can be categorized are added. Therefore, when the words in the text are separated into their roots, only the roots of the words that could be categories have been found.

The vector determined according to whether or not the words in the dictionary are in the text:

T = (0,0,0,0,0,0,0,0,1,0,0,0)

sim(E1, T)= 0, sim(E2, T)= 0,

sim(H1, T)= 0, sim(H2, T)= 0,

sim(M1, T)= 0, sim(M2, T)= 0,

sim(P1, T)=0, sim(P2, T)=0

sim(S1, T)=1, sim(S2, T)=0.

The similarity between all training vectors of the vector belonging to the test data is calculated. If this similarity is equal to 1, it indicates that the two vectors belong to the same class. The similarity with the S1 vector was found to be 1. This test data belongs to sports column.

5.3.2 Application of Naive Bayes Algorithm

Naive Bayes is a predictive and descriptive classification algorithm that analyzes the relationship between variables. Naive Bayes is a known and frequently used algorithm for text categorization. A training data set is determined for the target function, new samples are identified that are defined by the attribute values, and the person estimates the class target value or class (Bidgoli & Boraghi, 2010).

For example: In the third part of the application, the columns are classified according to their species and the Naive Bayes method was used while classification. The total number of words in the corners according to the categories in this study is shown in Table 5.3.

Table 5.3 The total number of words by category

Category	**Total**
Economy	1,006,338
Health	1,001,357
Magazine	1,003,664
Political	1,038,336
Sports	1,384,226
Total	5,433,921

A variable must be assigned to calculate the total value according to the table. Therefore, the total value is called *T* and the value is calculated as in Equation 5.1 (Değerli, 2012).

$$T = T_{economy} + T_{health} + T_{magazine} + T_{political} + T_{sports} \quad (5.1)$$

According to Equation 5.1, the *T* value is 5,433,921.

The category weight value of each category is obtained by dividing its total value by the total *T* value. This variable is also denoted by *W* and the value is calculated as in Equation 5.2 (Değerli, 2012).

$$W_C = \frac{T_C}{T} \tag{5.2}$$

Weights were found for each category using Equation 5.2. The weights of the categories are given in Table 5.4.

Table 5.4 Weights of categories

Category	**Weight (W)**
Economy	0.1851
Health	0.1842
Magazine	0.1847
Political	0.1910
Sports	0.2547

Let's consider the word "dolar" in the dictionary. The word does not contain any punctuation and punctuation mark. Therefore, the word will be used without any action. The number of passes of the "dollar" word by the category is given in Table 5.5.

Table 5.5 The sum of the words "dolar" have in categories

	Economy	Health	Magazine	Political	Sports
“dolar”	338	16	24	174	54

The total number of words is denoted by *F* and the number of words by category is indicated by *k*. The F value is calculated as Equation 5.3.

$$F = k_{economy} + k_{health} + k_{magazine} + k_{political} + k_{sports} \tag{5.3}$$

The F value was found to be 606 from Equation 5.3.

According to the algorithm, the weight of the word belonging to each category is found and the class is found after multiplying the total weight of the category. But there are “0” value problems in Naive Bayes applications. Since the values “0” will affect the multiplication, the desired result cannot be achieved. 0 value destroys the effects of probabilities and the end result is meaningless. To prevent this situation, a value such as *f* is added to each of the ratio. The *f* is a number between 0 and 1. Generally, *f* is chosen as 1.

The weight of the word for each category should be calculated using Equation 5.4. This calculation was made only for the "Economy" category.

$$P_C = \left(\frac{k+f}{F+f}\right) * W_C \tag{5.4}$$

When values are placed according to Equation 5.4, the P value of the economy category is calculated and the probability of *P(dolar | economy)* is found to be 0.1034.

$$P_{econamy} = \left(\frac{338+1}{606+1}\right) * W_{economy}$$

$$P_{economy} = 0.5584 * 0.1851 = 0.1034$$

The same procedure is applied for each category and the calculated category ratios are given in Table 5.6.

Table 5.6 Cross rates of "dollar" term in categories

	Economy	Health	Magazine	Political	Sports
“dolar”	0.1034	0.0051	0.0076	0.0550	0.0230

According to the calculated ratios, the probability of the economy category is 0.1034. Because the highest rate according to other categories belongs to the economy category, "dollar" word belongs economy category.

5.4 Application of Gender Identification

The first application is gender identification. There are two classes, male and female. A total of 800 columns of 400 male and female writers were collected.

5.4.1 K-NN Results of Gender Identification

According to the k-NN result, the classification of male authors gave better results. 312 of the 400 writers of the female authors and 336 of the male authors were correctly classified. Gender frequency values are shown in Table 5.7.

Table 5.7 Gender classification with bit-score weighted k-NN method

		Classified		
	Gender	**Female**	**Male**	**Total**
Real	**Female**	312	88	400
	Male	64	336	400

The accuracy ratios are shown in Table 5.8. The correct classification rate of female authors is 78% and that of males is 84%. As a result of the classification, male authors seem to use almost the same language. Female writers used a more ornate language. Thus, the classification of male authors has been more successful than that of female authors.

Table 5.8 Accuracy rates for gender classification using k-NN method

	Accuracy (%)
Female	0.78
Male	0.84

5.4.2 Naive Bayes Results of Gender Identification

The classification result using the Naive Bayes method is given in Table 5.9. According to Naive Bayes result, 342 of women writers are classified. Male authors classified 358 correctly. Male writers are better classified than female writers.

Table 5.9 Gender classification with Naive Bayes method

		Classified		
	Gender	**Female**	**Male**	**Total**
Real	**Female**	342	58	400
	Male	64	358	400

The correct classification rate of female authors is 85% and that of males is 90%. The Naive Bayes method is more successful than the k-NN method. The accuracy rate of both genders has increased by 7%. It is a good result to go up to 85% of accuracy rates. These accuracy ratios are given in Table 5.10.

Table 5.10 Accuracy rates for gender classification using Naive Bayes method

	Accuracy (%)
Female	0.85
Male	0.90

5.5 Application of Author Identification

The second application is author identification. In this application, the columns have been classified by the authors. The number of classes should be as much as the author. Therefore, the number of classes in this application is 20.

5.5.1 K-NN Results of Author Identification

Bit-score weighting k-NN classification method has been used to classify the columns according to the authors. Despite there are 40 columns belonging to each

author, the number of columns was found to be 40 over and 40 below. These columns found may not belong to the author. The correct classification number of columns is given in Table 5.11.

Table 5.11 Author classification with bit-score weighted k-NN method

Columnists	**Number of actual columns**	**Number of correct classifications**	**Number of misclassifications**
A. Ç.	40	34	4
A. S.	40	32	5
C. O.	40	35	7
E. Ç.	40	36	8
E. K.	40	30	5
E. Ö.	40	37	5
E. S.	40	39	7
E. S. Ö.	40	35	6
İ. Ç.	40	37	6
J. Ö.	40	34	3
M. V.	40	35	4
N. D.	40	38	6
N. L.	40	34	7
R. D.	40	32	7
S. K.	40	38	7
Ş. T.	40	34	4
T. D.	40	32	7
Y. A.	40	35	8
Y. Ö.	40	33	5
Z. G.	40	32	7

The correct classification ratios of the columns are shown in Table 5.12. The accuracy rates of classification are over 80%. The best classification is the author "J. Ö." with an accuracy of 91%. The author with the highest misclassified rate is “E. S.”.

Table 5.12 Accuracy rates for author classification using k-NN method

Columnists	**Accuracy (%)**
A. Ç.	0.89
A. S.	0.86
C. O.	0.83
E. Ç.	0.82
E. K.	0.86
E. Ö.	0.88
E. S.	0.80
E. S. Ö.	0.85
İ. Ç.	0.86
J. Ö.	0.91
M. V.	0.89
N. D.	0.86
N. L.	0.83
R. D.	0.82
S. K.	0.84
Ş. T.	0.89
T. D.	0.82
Y. A.	0.81
Y. Ö.	0.87
Z. G.	0.82

5.5.2 Naive Bayes Results of Author Identification

The second classification was made by Naive Bayes method. Classified columns are given in Table 5.13. The correct classification numbers are higher than the k-NN method. The number of correct classifications has increased.

Table 5.13 Author classification with Naive Bayes method

Columnists	**Number of real columns belonging to the author**	**Number of correct classifications**	**Number of misclassifications**
A. Ç.	40	37	2
A. S.	40	37	1
C. O.	40	39	2
E. Ç.	40	39	3
E. K.	40	37	1
E. Ö.	40	39	2
E. S.	40	36	1
E. S. Ö.	40	37	3
İ. Ç.	40	40	3
J. Ö.	40	40	0
M. V.	40	37	2
N. D.	40	39	3
N. L.	40	38	3
R. D.	40	37	3
S. K.	40	39	2
Ş. T.	40	36	3
T. D.	40	36	4
Y. A.	40	38	3
Y. Ö.	40	37	2
Z. G.	40	36	3

The correct classification ratios of the columns are found. These accuracy ratios are given in Table 5.14. The best classification is the author "J. Ö." with an accuracy of 100%. All columns belonging to the author are correctly classified as the application result. The accuracy rates of classification are over 90%.

Table 5.14 Accuracy rates for author classification using Naive Bayes method

Columnists	**Accuracy (%)**
A. Ç.	0.95
A. S.	0.97
C. O.	0.95
E. Ç.	0.93
E. K.	0.97
E. Ö.	0.95
E. S.	0.97
E. S. Ö.	0.93
İ. Ç.	0.93
J. Ö.	1
M. V.	0.95
N. D.	0.93
N. L.	0.93
R. D.	0.92
S. K.	0.95
Ş. T.	0.92
T. D.	0.90
Y. A.	0.93
Y. Ö.	0.95
Z. G.	0.92

5.6 Application of Species Determination

The third and final application is species determination. In this application, the species in which the columns belong is determined. Categories are defined as “Economy”, “Health”, “Magazine”, “Political” and “Sports”. There are five classes, and there is a total of 800 columns, 160 of which belong to each class.

5.6.1 K-NN Results of Species Determination

The frequency table of the k-NN results is shown in Table 5.15. There has been an incorrectly classified among the most economic columns and political columns. The reason for this is that the words used in economics and politics are similar. The magazine class has spread too much to other classes. The inclusion of words from other classes in magazine columns has caused incorrectly classified. Likewise, there are words in the health columns that can cover other classes.

Table 5.15 Species classification with bit-score weighted k-NN method

		Classified					
	Species	**Economy**	**Health**	**Magazine**	**Political**	**Sports**	**Total**
Real	**Economy**	140	4	4	10	2	160
	Health	8	136	6	2	8	160
	Magazine	8	8	128	6	10	160
	Political	14	5	8	131	2	160
	Sports	2	2	6	2	148	160

When looking at the accuracy rates, all categories are over 80%. The most successful results belong to the sports category. The reason why the accuracy rate of the sports category is high is that the words of the sports data are similar. The words used in sports data are more specific than the other categories. The category with the lowest accuracy is the magazine category. These accuracy ratios are given in Table 5.16.

Table 5.16 Accuracy rates for species classification using k-NN method

	Accuracy (%)
Economy	0.87
Health	0.85
Magazine	0.80
Political	0.82
Sports	0.93

5.6.2 Naive Bayes Results of Species Determination

The frequency table of the Naive Bayes results is given in Table 5.17. According to the Naive Bayes results, the classification was more successful than the k-NN method. The correct classification numbers have increased. The number of columns that do not belong to the specified class is less.

Table 5.17 Species classification with Naive Bayes method

		Classified					
	Species	**Economy**	**Health**	**Magazine**	**Political**	**Sports**	**Total**
Real	**Economy**	146	1	4	8	1	160
	Health	4	140	8	2	6	160
	Magazine	5	7	138	4	6	160
	Political	9	3	3	143	2	160
	Sports	0	0	3	1	156	160

The accuracy rates and incorrectly classified ratios of Naive Bayes results are given in Table 5.18. According to Naive Bayes results, the best classification belongs to the sports category. The accuracy rate of all categories is over 85%. According to the k-NN classification method, the accuracy rates of Naive Bayes are increased.

Table 5.18 Accuracy rates for species classification using Naive Bayes method

	Accuracy (%)
Economy	0.91
Health	0.87
Magazine	0.86
Political	0.89
Sports	0.97

The accuracy of the classification of columns is high with these techniques. But there are also incorrectly categorized columns. The biggest cause of misclassification in practice is the use of similar words. For example, suppose that a footballer is a column about love life. Although this column is a magazine category, it is assigned to the sports category. Because there is “football”, “football player” and “team name” in the column. Economic and political columns have been classified incorrectly as they have similar words.

CHAPTER SIX
CONCLUSION

In this section, the classification of the real database is made with the k-NN and Naive Bayes algorithms. As a data base, 800 columns of 50 days, published in 2018, were used. The classification has been done in three different ways as "Gender Identification", "Author Identification" and "Species Determination". As a result of the classification, the accuracy rates of these two methods are compared and it has been tried to make suggestions for future studies.

The classification algorithm in the study was written using the R studio program. Classification in application is done in three different ways. Gender identification was made as the first classification. In this application, there are two classes as male and female. The second application is author identification. Since there are 20 author, the number of classes in this section is 20. The third and last application is the species determination. In this application, there are five classes as economy, health, magazine, political and sports. The number of classes determined in the applications can be increased by adding words to the dictionary and adding training data.

According to the test results made during the application, the most successful algorithm was found as Naive Bayes. The Naive Bayes method is one of the most successful algorithms used for classifying text as the accuracy rate. The Naive Bayes algorithm is easy to implement and has high performance. When the accuracy rates are compared, the Naive Bayes accuracy rate is more successful than the k-NN method. The reason for this is that the vector to be categorized when calculating the probability of Naive Bayes algorithm does not need to calculate the cosine similarity between all training vectors.

In the first application gender classification was done. Classification was done by bit-score weighting k-NN method and Naive Bayes method. According to the bit-score weighting k-NN method, 78% of female authors and 84% of male authors were correctly classified. According to Naive Bayes binary weighting method, 85% of

female authors and 90% of male authors are correctly classified. The Naive Bayes method is more successful than the k-NN method. In both methods, the accuracy of male authors is higher than the accuracy of female authors. The reason for this is that male authors use more similar words than female authors.

In the second application, the columns were classified according to the authors. In this application, the Naive Bayes method gave better results than the k-NN method. In both methods, the correct classification rate is over 80%. It is the author J. Ö. who has the highest accuracy because of classification by Naive Bayes method. all the columns belonging to this author are correctly classified.

In the last application, columns are classified according to types. In both methods, the best result belongs to sports columns. There are more determinative words in sports data than in other categories. The fact that many words are different from other categories makes the sport category more successful in giving correct results.

The accuracy of the economic and political categories is less than that of the sports category. The reason for this is that these two categories have similar words as content. The lowest accuracy rate is the magazine category. The reason why the accuracy rate of the magazine category is low is that the columns of each category can be a magazine subject.

In the study, the Naive Bayes algorithm is more successful than the k-NN algorithm. The success of the k-NN algorithm can be increased by increasing the number of training data. However, as the number of training data increases, the algorithms used may experience problems in terms of speed performance.

To increase the success of the algorithms, the number of training data and the number of words in the dictionary can be increased. Apart from this, the words that do not express or frequently use can be excluded from the text. At the same time, indexing algorithms can be used to increase performance speed.

As a result, successful results can be obtained in large data using classification techniques. Text mining is a technique that yields success when some corrections are made on Turkish texts other than other languages. Therefore, text mining has been used in many areas in recent years. In the future automatic classification will take the place of manual classification. Human power saving and speed achieved are also an advantage of these techniques.

REFERENCES

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *In ACM sigmod record 22*(2), 207-216).

Ahi, L. (2015). *Veri madenciliği yöntemleri ile ana harcama gruplarının paylarının tahmini*. Master's thesis, Hacettepe University, Ankara.

Altaş, D., & Gülpınar, V. (2012). Karar ağaçları ve yapay sinir ağlarının sınıflandırma performanslarının karşılaştırılması: Avrupa Birliği örneği. *Trakya University Journal of Social Sciences, 14*(1), 1-22.

Atalay, M., & Çelik, E. (2017). Large data analysis and artificial intelligence and machine learning applications in big data analysis. *Mehmet Akif Ersoy University, Journal of Social Sciences Institute, 9* (22), 155-172.

Bai X. (2011). Predicting consumer sentiments from online text. *Decision Support Systems, 50*(4), 732-742.

Bahadır, Ö. (2007). *Aritmetik problemlerin çözümlenmesi ve simgelenmesi.* Master's thesis, Yıldız Teknik University, İstanbul.

Bidgoli, A. M., & Boraghi, M. (2010). A language independent text segmentation technique based on Naive Bayes classifier. *International Conference on Signal and Image Processing*, 11-16.

Bramer, M. (2007). *Principles of data mining* (2th ed.). London: Springer.

Bulut, F., & Amasyalı, M. F. (2014). *En yakın k komşuluk algoritmasında örneklere bağlı dinamik k seçimi.* Yıldız Teknik University, İstanbul.

Burak, A. (2017). *Naive Bayes Sınıflandırıcısı*. Retrieved June 28, 2018, from http://yazilimagiris.com/2017/11/naivenaif-bayes-siniflandiricisi.

Cios, K. J., Pedrycz, W., Swiniarski, R. W., & Kurgan, L. A. (2007). *Data mining: a knowledge discovery approach.* Springer Science & Business Media.

Çalış, A., Kayapınar, S., & Çetinyokuş, T. (2014). Veri madenciliğinde karar ağacı algoritmaları ile bilgisayar ve internet güvenliği üzerinde bir uygulama. *Journal of Industrial Engineering, 25*(3), 2-19.

Dang, S., & Ahmad, P. H. (2014). A comparative study on text mining techniques. *Global Journal of Advanced Research, 1*(2), 128-134.

Dinler, M. (2014). *Kümeleme analizi yöntemlerinin hayvancılık verilerinde karşılaştırmalı olarak incelenmesi.* Master's thesis, Bingöl University, Bingöl.

Değerli, O. (2012). *Naive Bayes yöntemi ile blog içeriklerinin sınıflandırılması*. Master's thesis, Gazi University, Ankara.

Delibaş, A. (2008). *Doğal dil işleme ile Tükçe yazım hatalarının denetlenmesi.* Master's thesis, Yıldız Teknik University, İstanbul.

Dolgun, Ö., Özdemir, T., & Doruk, O. (2009). Veri madenciliği'nde yapısal olmayan verinin analizi: Metin ve web madenciliği. *Statisticians Magazine, 2*, 48-58.

Döşlü, A. (2008). *Veri madenciliğinde market sepet analizi ve birliktelik kurallarının belirlenmesi.* Doctoral dissertation, Yıldız Teknik University.

Ergezer, H., Dikmen, M., & Özdemir, E. (2003). Yapay sinir ağları ve tanıma sistemleri. *Pivolka, 2*(6), 14-17.

Ergün, M. (2017). Using the techniques of data mining and text mining in educational research. *Electronic Journal of Education Sciences, 6(*12), 180-189.

Erhardt, R. A., Schneider, R., & Blaschke C. (2002). Status of text mining techniques applied to biomedical text. *Drug Dicovery Today, 11*(7), 315-325.

Fabrizio S. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34*(1), 1–47.

Feldman, R., & Sanger, J. (2007). *The text mining handbook advanced approaches in analyzing unstructured data* (2th ed.). Cambridge University Press, U.S.A.

Gaikwad, S. V., Chaugule, A., & Pramod, P. (2014). Text mining methods and techniques. *International Journal of Computer Applications, 85*(17), 43-44.

Gaizauskas, R. (2002). An information extraction perspective on text mining: Task technologies and prototype applications. *Euromap Text Mining Seminar,* Sheffield, UK.

Gökgöz, K. B. (2015). Veri madenciliği ile tıbbi cihaz bakım karar modeli. Master's thesis, Başkent University, Ankara.

Grad, B., & Bergin, T. J. (2009). Introduction: History of database management systems. *IEEE Annals of the History of Computing, 31(*4), 3-5.

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence, 1*(1), 60-63.

Güllüoğlu, S. (2011). Tıp ve sağlık hizmetlerinde veri madenciliği çalışmaları: Kanser teşhisine yönelik bir ön çalışma. *Online Academic Journal of Information Technology, 2*(5).

Güven A. (2007). *Türkçe belgelerin anlam tabanlı yöntemlerle madenciliği*. Doctoral dissertation, Yıldız Teknik University, İstanbul.

Han, J., Kamber, M., &. Pei J. (2006). *Data mining concepts and techniques* (3ed.). Morgan Kaufmann Publishers, Champaign.

Haznedar, B., Arslan, M. F., & Kalınlı, A. (2016). Karaciğer mikrodizi kanser verisinin sınıflandırılması için genetik algoritma kullanarak ANFIS'in eğitilmesi. *Sakarya University Graduate School of Natural and Applied Sciences, 21*(1), 54-62.

İnan, O., (2003). *Veri madenciliği,* Master's thesis, Selçuk University, Konya.

Kalikov, A. (2006). *Veri madenciliği ve bir e-ticaret uygulaması.* Master's thesis, Gazi University, Ankara.

Kavzoğlu, T., & Çölkesen, İ. (2010). Karar ağaçları ile uydu görüntülerinin sınıflandırılması: Kocaeli örneği. *Electronic Journal of Map Technologies, 2*(1), 36-45.

Kaynar, O., Tuna, M. F., Görmez, Y., & Deveci, M. A. (2017). Makine Öğrenmesi Yöntemleriyle Müşteri Kaybı Analizi. *Cumhuriyet University Journal of Economics and Administrative Sciences, 18*(1), 6-9.

Kılınç, D., Borandağ E., Yücalar, F., Tunalı, V., Şimşek, M., & Özçift, A. (2016). K-NN algoritması ile metin madenciliği kullanılarak bilimsel makale tasnifi. *Marmara Science Journal, 3*, 89-94.

Koyuncugil, A. S., & Özgülbaş, N. (2009). Veri madenciliğinin tıp ve sağlık alanında kullanımı. *Journal of Information Technologies, 2*(2), 70-80.

Larose, D. (2005). *Discovering knowledge in data an introduction to data mining.* New Jersey: Published by John Wiley & Sons, Inc.

Lihua, Y., Qi, D., & Yanjun, G. (2006). Study on k-NN text categorization algorithm. *Micro Computer Information, 3*(21), 269-271.

Melek, C. (2012). *Metin madenciliği teknikleri ile şirketlerin vizyon ifadelerinin analizi.* Master's thesis. Dokuz Eylül University, İzmir.

Onur, H. (2007). Dizinleme amacı ile kullanılabilecek yöntemlerin kıyaslanması ve arama sistemi geliştirilmesi. Master Thesis, Gazi University, Ankara.

Oğuz, B. (2009). *Metin madenciliği teknikleri kullanılarak kulak burun boğaz hasta bilgi formlarının analizi.* Master Thesis, Akdeniz University, Antalya.

Oğuz, B., Bilge U., & Saka, O. (2009). Tıpta metin madenciliği. *Journal of Biostatistics and Medical Informatics,* Akdeniz University, Antalya.

Özkan, Y. (2008). *Veri Madenciliği Yöntemleri*. Papatya Publications, İstanbul.

Pilavcılar, İ.F. (2007). *Metin madenciliği ile metin sınıflandırma*. Master Thesis, Yıldız Teknik University, İstanbul.

Rana, M. I., Khalid, S., & Akbar, M.U. (2014). News classification based on their headlines: A review. *17th International in Multi-Topic Conference,* 211-216.

Sarawagi, S. (2007). Information extraction. *Foundations and Trends in Databases, 1*(3), 261-377.

Sariman, G. (2011). Veri madenciliğinde kümeleme teknikleri üzerine bir çalışma: k-means ve k-medoids kümeleme algoritmalarının karşılaştırılması. *Süleyman Demirel University, Journal of the Institute of Science, 5*(3).

Savaş, S., Topaloğlu, N., & Yılmaz, M. (2012). Veri madenciliği ve Türkiye'deki uygulama örnekleri. *İstanbul Commerce University Journal of Science and Technology, 11*(21), 1-23.

Seker, S.E. (2015). Metin madenciliği. *YBS Ansiklopedi, 3*(2), 30-33.

Soucy, P., & Mineau, W. (2005). Beyond TFIDF weighting for text categorization in the vector space nodel, *IEEE International Canference*, 1130-1135.

Sumathi, S., & Sivanandam, S. N. (2006). *Introduction to Data Mining and its Applications, 29*. Berlin: Springer-Verlag.

Suguna, N., & Thanushkodi, K. (2010). An improved k-nearest neighbor classification using genetic algorithm. *International Journal of Computer Science Issues, 7*(4), 19-20.

Tan, P. N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*. Addison Wesley Publishers.

Tang, Z., & Maclennan, J. (2005). *Data mining with sql server 2005*. Wiley. New Jersey.

Teyseyre, A. R., & Campo, M. R. (2009). An overview of 3D software visualization. *IEEE Transactions on Visualization and Computer Graphics, 15*(1), 87-100.

Trstenjak, B., Mikac, S., & Donko, D. (2013). KNN with TF-IDF based framework for text categorization. *24th International Symposium on Intelligent Manufacturing and Automation, 69*, 1356 – 1364.

Uytun, M. B. (2013). *Opinion mining with text operations and extracting data from user reviews for e-commerce applications*, Master's thesis, Çankaya University, Ankara.

Waheeb, A., & Babu, A. P. (2017). Arabic question answering system based on data mining. *International Journal of Scientific & Technology Research*, *6*(2), 237-239.

Weiss, S.M., Indurkhya, N., & Zhang, T. (2010). *Information retrieval and text mining: In fundamentals of predictive text mining* (4th ed.). London.