DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

USING TEXT MINING ALGORITHMS OF HEALTH DATA FOR ANALYSIS PURPOSES

by
Müslüm Serdar AKİS

November, 2019
İZMİR

USING TEXT MINING ALGORITHMS OF HEALTH DATA FOR ANALYSIS PURPOSES

A Thesis submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Degree of Master of Science
in Computer Engineering Program

by

Müslüm Serdar AKİS

November, 2019
İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"USING TEXT MINING ALGORITHMS OF HEALTH DATA FOR ANALYSIS PURPOSES"** completed by **MÜSLÜM SERDAR AKİS** under supervision of **ASST. PROF. DR. SEMİH UTKU** 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.

Asst. Prof. Dr. Semih Utku

Supervisor

(Jury Member)

Prof. Dr. Alp Kut

(Jury Member)

Prof. Dr. Emin Anarım

Prof.Dr. Kadriye ERTEKİN

Director
Graduate School of Natural and Applied Sciences

ACKNOWLEDGMENTS

I would like to express my special thanks to "Asst. Prof. Semih UTKU". As my teacher and mentor, he gave me the opportunity to work on this project. Without his dedicated works and recommends, this work would not have been possible.

I would like to dedicate this work to my mom and my father for their supports, prayers and sacrifices since I was born.

At last but not least, I would like to thank Hilal BUCAN for her constant help for all these years.

Müslüm Serdar AKİS

USING TEXT MINING ALGORITHMS OF HEALTH DATA FOR ANALYSIS PURPOSES

ABSTRACT

Patients' narratives were analyzed through qualitative and quantitative textual analysis in order to create a system that can help the specialist in diagnosis. The patients' narratives were examined and checked if there were any relation between diagnosis and the knowledge as an outcome of mining the narratives. The different methods of text mining have been used to extract useful patterns and knowledge from these diagnostic groups. The most successful algorithms have been determined and selected to be used in the recommended system. Simultaneous use of more than one algorithm was examined and the results were compared. By selecting the most effective combination and system, success of the estimation process was tried to be increased. The result with the highest probability of success is suggested to the doctor. The results of the studies show that successful suggestions can be made with the proposed system and it has been observed that the system can have significant contributions in increasing the quality of the treatment of patients.

Keywords: Health care, natural language processing, text mining, treatment follow-up

SAĞLIK VERİLERİNDE METİN MADENCİLİĞİ ALGORİTMALARININ ANALİZ AMAÇLI KULLANILMASI

ÖZ

Hastaların hikayeleri nitel ve nicel metinsel analizi yoluyla, tanı koymada uzmana yardımcı olmak için bir öneri sistemi oluşturulması amaçlandı. Hastaların hikayeleri incelendi ve hikayeler ile teşhis bilgisinin herhangi bir ilişkisi olup olmadığı kontrol edildi. Bu tanı guruplarından bilgi çıkarmak için çeşitli metin madenciliği yöntemleri kullanıldı. En başarılı algoritmalar belirlenerek başarılı bir öneri sistemi geliştirilmek için kullanıldı. Birden fazla algoritmanın kullanılması durumunda elde edilen sonuçlar karşılaştırıldı. En etkili seçimler yapılarak başarı oranları artırılmasına çalışıldı. En yüksek olasılığa sahip sonuç ise, öneri olarak doktora verildi. Çalışmanın sonucu göstermektedir ki, bu verilerin ışığında başarılı bir öneri sistemi oluşturulabilir ve tedavi kalitesini arttırma konusunda sisteme önemli katkılarda bulunulabilir.

Anahtar kelimeler: Sağlık hizmeti, doğal dil işleme, metin madenciliği, hasta takip form

CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER ONE
INTRODUCTION

People have always been searching for new ways to express themselves and to understand the nature basics. Everything that humanity achieved is landed on that instinct. Our current technology and knowledge are based on these foundations. At the beginning, people were communicating via cave paintings and tools (Gönenç, 2007). However, after the invention of writing, communication stepped into a new dimension and the first era began.

Since the first era, the knowledge and culture obtained by humanity was able to be legated throughout the generations. Inventions and experiences have been released from limited human life; and the knowledge has become with other people and other culture. The studies have been accessible by different people and become operable from different perspectives. As a result of the transferring of all data obtained over the years, concepts of theories, scientific laws and history were discovered. Despite all of these, if there is another invention that is more significant than writing that begins a new era, it must be the internet.

The Internet has managed to reduce access time for information obtained from research. Getting that information before internet could take years or could take even a lifetime. However, the internet decreases this to seconds. Every study produced result and idea became shareable with the whole world within seconds. As a result, scientists from different research fields can examine these results and improve methods by making criticisms, improvements and additions. However, increasing speed of communication and information has brought its own problems.

The amount of generated information is increasing rapidly every day because of developing technology and the outcome of the digitalized world. According to opinion in the 1950s, doubling time of medical knowledge was 50 years. When the year became 1980, the doubling time decreased to only 7 years. In 2010, doubling time was 3.5 years. Also, it is estimated that this period could decrease to 73 days in 2020 (Densen, 2011). However, extracting patterns and knowledge from this information is

the biggest problem of this era. The resulting of mass data is huge and time-consuming to be examined by humans. Exactly at this point, text mining is one of the most important technologies.

The history of the beginning of text mining dates back to 1980's. However, text mining has become very popular during the recent years with the help of developing technology and increased processing power. The main purpose of text mining is to get knowledge over unstructured data such as text. It is known that the majority of information is stored as text (common estimation says over 80%) (Ghosh, Roy, & Bandyopadhyay, 2012), therefore, it is quite natural that text mining has become so popular. Text analytics is an answer to the "unstructured data" problem, which is best expressed by the truism that eighty percent of the information is originated and locked in "unstructured" form (Luhn, 1958). Thus, we are able to extract knowledge from the information explosion that we have today.

For over 50 years, computer scientists have developed algorithms to analyze natural language text, using either sets of hand-written rules or machine learning techniques. With the help of the developing technology, it has become possible to digitize this information entered in text format and stored in digital environment. If the obtained data are transformed into a structure that can assist the medical personnel, it will make the patient's satisfaction and the quality of treatment. It is also possible that the obtained experiences will be used in future treatments and the human errors that may occur are avoided.

The main issue that needed to be resolved is the vast majority of knowledge in the medical field, where limited staff and patient follow-up are vital, is inaccessible state. The information and results obtained have been stuck in a hospital. However, this information has a very high potential knowledge about patients and treatments.

In these days, a great majority of the information in hospitals and clinics about patients is kept as text. This situation is not much different in the field of spinal surgery. The data recorded over the years is not reviewed in any ways and the knowledge of the content remains as confidential. Hence, we cannot get answers to the questions like

“Are there some similarities between patients?” or “Is the patient recovering or getting worse?”. In a vital subject such as spinal cord surgery, it is terrifying that such a great accumulation of knowledge remains as idle.

The majority of patient follow-up forms, treatment reports and emergency service reports are kept as text and remains idle. However, if this information stores in a structured format, similar situations can be examined, the treatment and interventions to be applied can be predicted with the help of this information. Especially in an area where the limited number of experienced personnel exists, the knowledge of treatment has a vital importance. However, once that knowledge is uncovered and also supported with a decision assist system, medical errors can be eliminated with the suggestions. The quality of treatment which affects patient satisfaction directly could be increased.

Medical errors are associated with inexperienced physicians and nurses, new procedures, extreme of age, and complex or urgent care (Weingart, Wilson, Gibberd, & Harrison, 2000). The vast majority of medical errors results from faulty systems and poorly designed processes versus poor practices or incompetent practitioners (Palmieri, DeLucia, Peterson, Ott, & Green, 2008). Frequently reported sentinel events are shared in Figure 1.1. These facts (The Joint Commission, n.d.) show that medical errors are still one of the major issues.

Top 10 most frequently reported Sentinel Events in 2018	
Type of Sentinel Event	**Events Reported**
Fall	111
Unintended retention of a foreign body	111
Wrong-site surgery	94
Unassigned*	68
Other unanticipated event**	59
Suicide	50
Delay in treatment	43
Product or device events	29
Criminal event	28
Medication error	24
*This category was unassigned at the time of the report. **Includes asphyxiation, burns, choking on food, drowning, and being found unresponsive.	

Figure 1.1 Statistic of sentinel events (The Joint Commission, n.d.)

Another statistic from the Minnesota Department shows same issues (Health, 2009). Statistics shared in Figure 1.2. Same problems are spread in United States. For example, medical errors constitute a significant part of the causes of death in the United States (Henneman et al., 2005). This information confirms that that quality improvement of health care is necessary. Wrong treatment still causes of losing lives which can be saved with a little caution.

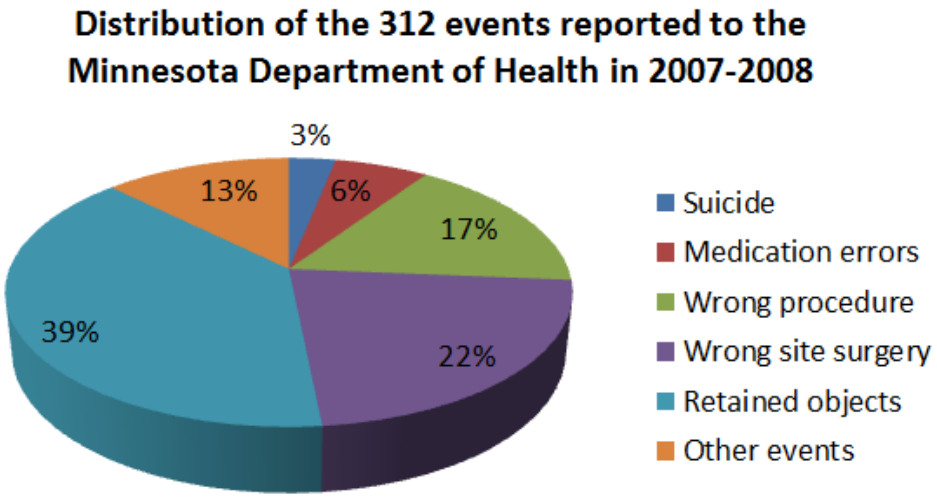

Figure 1.2 Statistics II

The main motivation of the this study is to save lives by helping to prevent these and similar situations and to help them to provide a better quality of treatment. Especially in an area where the limited number of experienced personnel exists, the knowledge of treatment has a vital importance. However, such a structure is not used in the existing system. Experiences acquired during the treatment process cannot be transferred to other doctors. This directly affects patient satisfaction and may cause misdiagnosis.

The objective of this study is to show that, through qualitative and quantitative textual analysis of a spinal surgery and spinal data, it is possible to build a recommendation system and generate a new body of knowledge by adding relevant information to describe patients' narratives. It seems that such a research is lacking in spinal surgery. Additionally, it is thought that mistakes can be prevented by creating awareness amongst the medical personnel by the results that can be obtained.

In our proposed system, the patients' narratives were examined and checked if there was any relation between diagnosis and the knowledge as an outcome of mining the narratives. In addition, a prototype suggestion system has been developed to assist the doctor in decision-making period with the help of the result. The suggested system and the applied methods are described in details and the results of the tests performed are

shared in the following sections. However, there were some difficulties we face during this study.

Although text mining algorithms and methods have been developed for many years, the situation is different from other processes when the data in the health context is examined from the perspective of text mining. Healthcare data is complex, heterogeneous, and scattered. Therefore, it is difficult, time-consuming, and laborious to integrate them reliably. The majority of tools for analysis of text were trained on edited text genres such as newspaper articles or scientific papers (Carroll, Koeling, & Puri, 2012). Also, due to patient rights, information had to be taken anonymously. These conditions have been taken into consideration when studying.

The data was collected from patients by Neurological Surgeries and Spinal Surgeries who applied for (Utku, Baysal, & Zileli, 2010). This study contained a total of 31752 patients (14128 men and 17624 women) with 79 different diagnostic groups in the period of 01/01/2005 to 31/12/2015. During each examination, complaints of the patients were listened and their narratives were recorded electronically. The total number of narratives is 65606 and diagnoses were determined for each examination and recorded with the associated texts.

CHAPTER TWO
LITERATURE REVIEW

With the development of text and data mining, the number of studies has increased in the health area. In contrast to the number of studies using text and data mining in the health area, no studies were found in the spinal surgery. However, it has been seen that the studies performed in the health field obtained very successful results.

Michelson, Pariseau, & Paganelli (2014) have used text mining techniques to detect surgical site infections (SSI) in unstructured clinical notes to improve SSI detection. The study group included patients who underwent elective orthopedic surgery in the following areas in 2010: joint replacement, foot and ankle (F & A), sports medicine, and upper extremity. In the EMR (Epic Systems, Verona, WI) notes (in plain text), laboratory test results, drugs and problem lists were taken. This study confirms the effectiveness of the approach. The study shows 100% of the SSIs detected by traditional infection control. Also, it discovered 37 SSI which prevents traditional surveillance deviation.

Srinivas, Rani, & Govrdhan (2010) have examined the potential use of classification-based data mining techniques in a large number of healthcare services. In this study, intelligent and effective cardiac arrest prediction methods were presented using data mining techniques; Rule-based, Decision Tree, Naïve Bayes and Artificial Neural Network. These techniques focus on using different algorithms to predict combinations of various target properties. To predict the heart attack effectively, heart disease has provided an effective approach to extract important patterns from data warehouses. On the basis of the calculated significant weight, frequent molds having larger values than a predetermined threshold were selected for the strain-valued estimation. When the result values were checked, it was seen that this approach increased the current estimate by 5%.

Jen, Wang, Jiang, Chu, & Chen (2012) have developed an early warning system to classify chronic disease using K-NN and Linear Discriminant Analysis (LDA). In this study, K-NN was used to determine the risk factors of various chronic diseases, to

reduce the relationship between cardiovascular disease and hypertension and to establish an early warning system for the complications of these diseases.

Delespierre, Denormandie, Bar-Hen, & Josseran (2017) have extracted the meaningful words through Standard Query Language (SQL) with the LIKE function and wildcards to perform pattern matching followed by text mining and a word cloud using R packages. This textual experiment using a two-step textual procedure has shown that text mining and data mining techniques provide appropriate means to improve residents' health and quality of care by adding new, simple and useful data on electronic health record (EHR).

Jonnagaddala, et al. (2015) have developed a system to extract risk factors of coronary artery disease (CAD) to prevent or intervene disease in early stages. Risk factors such as social history, family history, medication history and comorbidities are usually embedded in unstructured clinical narratives. Risk factor data is extracted from unstructured clinical notes with clinical text mining methods. This study uses clinical text mining methods to extract Framingham risk factors from unstructured electronic health records and calculates 10-year coronary artery disease risk scores in a cohort of diabetic patients. The text mining system performance was reliable and consistent in extracting the risk factor data. However, this study also observed that a rule-based system may fail in a few situations.

Gupta et al. (2016) have developed a system called miRiaD (microRNAs in association with Disease). The system is a text-mining tool that automatically extracts associations between microRNAs and diseases from the literature. miRiaD, a high-performance system that can capture a wide variety of microRNA-disease related information, extends beyond the scope of existing microRNA-disease resources. miRiaD was applied to the entire Medline corpus, identifying 8301 PMIDs with miR-disease associations. In this study, authors evaluated the recall and the precision of miRiaD according to information of high interest to public microRNA-disease database curators (expression and target gene associations), obtaining a recall of 88.46–90.78.

In Alex, et al. (2019), CT and MRI have used text mining methods to classify brain scans. The aim of the study is to classify the reports using the symptoms and observations of stroke occurrences. The system is called as Edinburgh Information Extraction for Radiology reports (EdIE-R). This paper is based on 1,168 radiology reports. Besides of developing the rule-based EdIE-R system to identify phenotype information related to stroke in radiology reports, they also created manually annotations for this data in parallel. The equivalent system scores in the blind test results for the first annotator were 95.49 for entities, 98.27 for relationships, 94.41 for negation and 96.39 for labels and second annotator results are 96.86, 96.01, 96 for respectively.

CHAPTER THREE
MINING AND DECISION SUPPORT SYSTEMS

People have always been inclined to make mistakes. Problems of daily life, heavy working conditions, busy schedules and the stressful living of the big cities cause people to make wrong decisions. These wrong decisions can sometimes cost people lives (Bernstein, Hebert, & Etchells, 2003). These problems could be preventable with the help of data mining techniques and decision support systems.

Huge amount of data is recording on each day. Storing that amount of data and editing data in line of the requirements has become very easy after computer has entered our lives. However, data only by itself has no use. Data gathered needs to be processed through the line of the requirements and needs to extract useful information. Otherwise, information that is not used for any purpose is nothing but a wasted memory. However, numerous rows of data are exhaustive and troublesome to be examined by humans. Therefore, there was a need for a method to examine large amounts of data. Precisely at this point, data mining has become popular.

Although data mining and decision support systems are not the same thing, they can be used together at many points and become the systems that can make a difference. Numerous rows of data could be analyzed by using data mining techniques and combined with a useful system and then human errors could be prevented. These two domains will be briefed detailly at following titles.

3.1 Data Mining

Data mining is the process of analyzing data to find useful information that has not yet been discovered, revealing previously unknown patterns and relationships between attributes hidden in the data (Rupnik, Kukar, Bajec, & Krisper, 2006). The purpose of data mining is to extract information from the data set and make it understandable for further use (Marwaha, 2014).

Manual knowledge extraction from data has been performed over the centuries. First examples of knowledge extraction from data could be Bayes' Theorem at 1700s and regression analysis 1800s. Data mining first appeared with primitive methods by economists and statisticians but after 1960, it has been transformed to a data fishing. However, data mining has appeared as a term firstly used in 1983 by Michael Lovell (1983).

Huge amount of information holds as idle without uncovered even today. The similarities or differences in these data stay covered. However, due to increasing amount of data and data complexity, efforts have moved to an automatic knowledge extraction system than manual knowledge extraction. Using data mining techniques, meaningful patterns can be extracted from large data sets that cannot be examined by humans. In addition, the system is self-feeding, so, the results could be getting better by the time.

The main reason why data mining has become so popular and effective, data mining could identify similarity trends and patterns among groups of data. In this way, processes can be automated. Quick decisions could be made at critical points with the help of data mining. Classification or decision-making mechanism that will last for years or months can be achieved in seconds thanks to data mining techniques and a great workforce could be saved. Data mining could inspected in five topics as visualized in Figure 3.1 (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

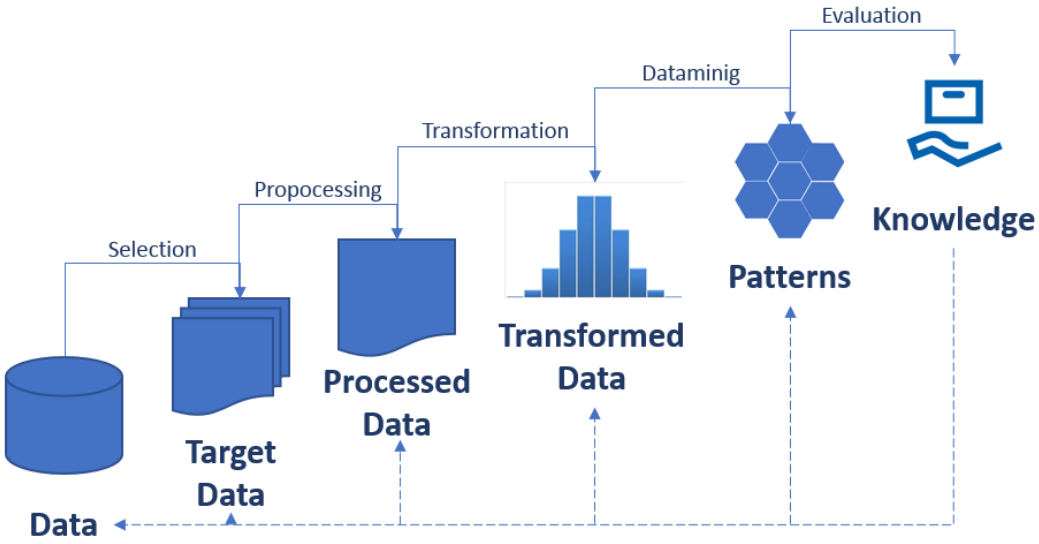

Figure 3.1 An overview steps of data mining

3.1.1 Selection

Selection is the first step to be taken before starting data mining. Not all data can be meaningful in large data warehouses. These data which do not make any sense can affect the operation of the algorithms and cause worse and meaningless outcomes. Significant data records and databases should be selected. Databases and data rows which will lead to intended results are selected at this stage.

3.1.2 Pre-Processing

Selected data could be noisy. After selection process, extracted data should be investigated and should be decided if data could be used in decision support or data mining. Data sets should be combined and systematized. Data should be cleaned and simplified according to the patterns that need to be exposed.

3.1.3 Transformation

Third step of the data mining is the transformation. In this stage, data cleared from previous step would transform into a supported structure for data mining algorithm. In

transformation step data are made for data mining. All data would be digitized after this step, according to algorithms which is going to use.

3.1.4 Mining

Mining step is the main step of learning and knowledge extraction. Mining could be split into six stages. (Fayyad et al., 1996).

Anomaly detection: Data would be investigated to see if there is any kind of anomaly. If there is an anomaly, it should be investigated again whether there is a chance to bypass it. Those process are made in this stage.

Association rule learning: In this stage, relations in features of data are investigated to determine the connection. If there is a connection, then, the related features can be used.

Clustering: Groups and structures are investigated to see if there are similarities in Clustering stage.

Classification: This is the stage of applying the extracted information to the data. The results are exposed at this stage.

Regression: It is the stage of extracting of the function that will estimate results with the least error.

Summarization: This is the last stage of data mining. Outputs would be visualized and report its success rate.

3.1.4 Interpretation or Evaluation

Result could be misleading in data mining. Although the results at first seem to be meaningful, they may not actually produce any meaningful results according to the future situations. This is usually due to lack of appropriate testing or a selection of not

appropriate features. This could be relatively prevented by separating the test set from the training set. However, only splitting sets may not be enough (Hawkins, 2004).

The final step of data mining is to verify that the pattern generated by the mining algorithms occurs in larger clusters. Even if not all the patterns are correct, it is tried to eliminate the patterns that are thought to be wrong and that damage the result. Performing tests with a data set that has not been used in training stage is very important at this point.

3.2 Text Mining

Manual text mining studies first started in 1980 (Schulman, Castellon, & Seligman, 1989). However, in the light of technological developments, text mining golden era of text mining was after 2000s. Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources (Hearst, 1999). Overall system process is shown in Figure 3.2.

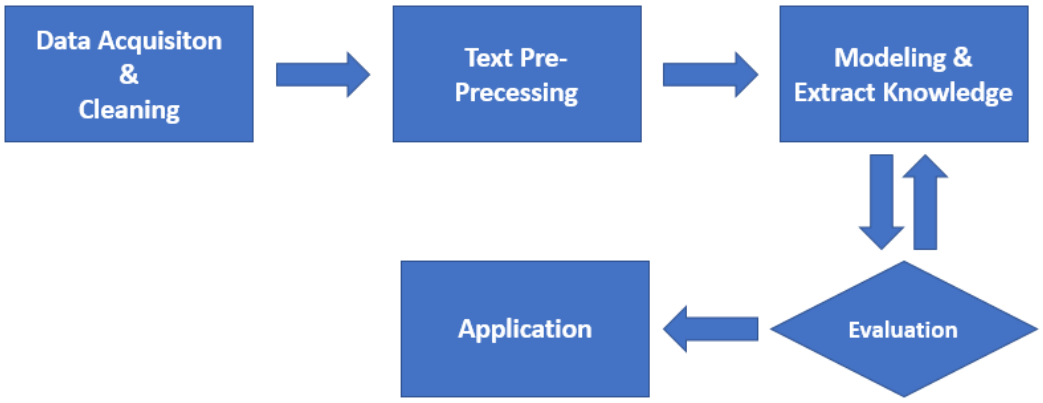

Figure 3.2 Overall text mining process

Text mining is not actually much different than data mining. The difference between them is that data mining is designed to process structured data. However, text mining can examine unstructured or semi-structured data. Html files or emails could be given as examples of unstructured data (Gupta, & Lehal, 2009). In text mining,

intermediate algorithms and text cleanup methods are used to remove patterns from texts in the pre-process stage.

The pre-process stage plays a very important role in text mining. At this stage, meaningless characters are removed from the text. Words that do not affect the result are removed from the text. The suffixes and prefixes in the words are removed and meaningful texts are tried to be extracted as much as possible. However, no matter how successful data mining and text mining practices are done, it doesn't do much good if it is not combined with a system that will help people. Decision support systems come into play at this stage.

3.3 Decision Support System

Companies and governments are getting more and more data stored every day. Therefore, extracting knowledge from these data is becoming harder and harder every day. Decision support systems can prevent errors at critical points. The concept of decision support system was evolved in 1960 by the work of the Carnegie Institute of Technology (Keen, 1978). Although the definition and scope of DSS has been defined as "computer-based system to help decision making as", it has been involved as "interactive computer-based auxiliary systems" at the end of the 1970s (Diasio, & Agell, 2009). In the 1980s, DSS became very popular, and in the late 80's, executive information systems (EIS), group decision support systems (GDSS) and organizational decision support systems (ODSS) were developed from a single user and model-oriented DSS. General structure of DSS is visualized in Figure 3.3.

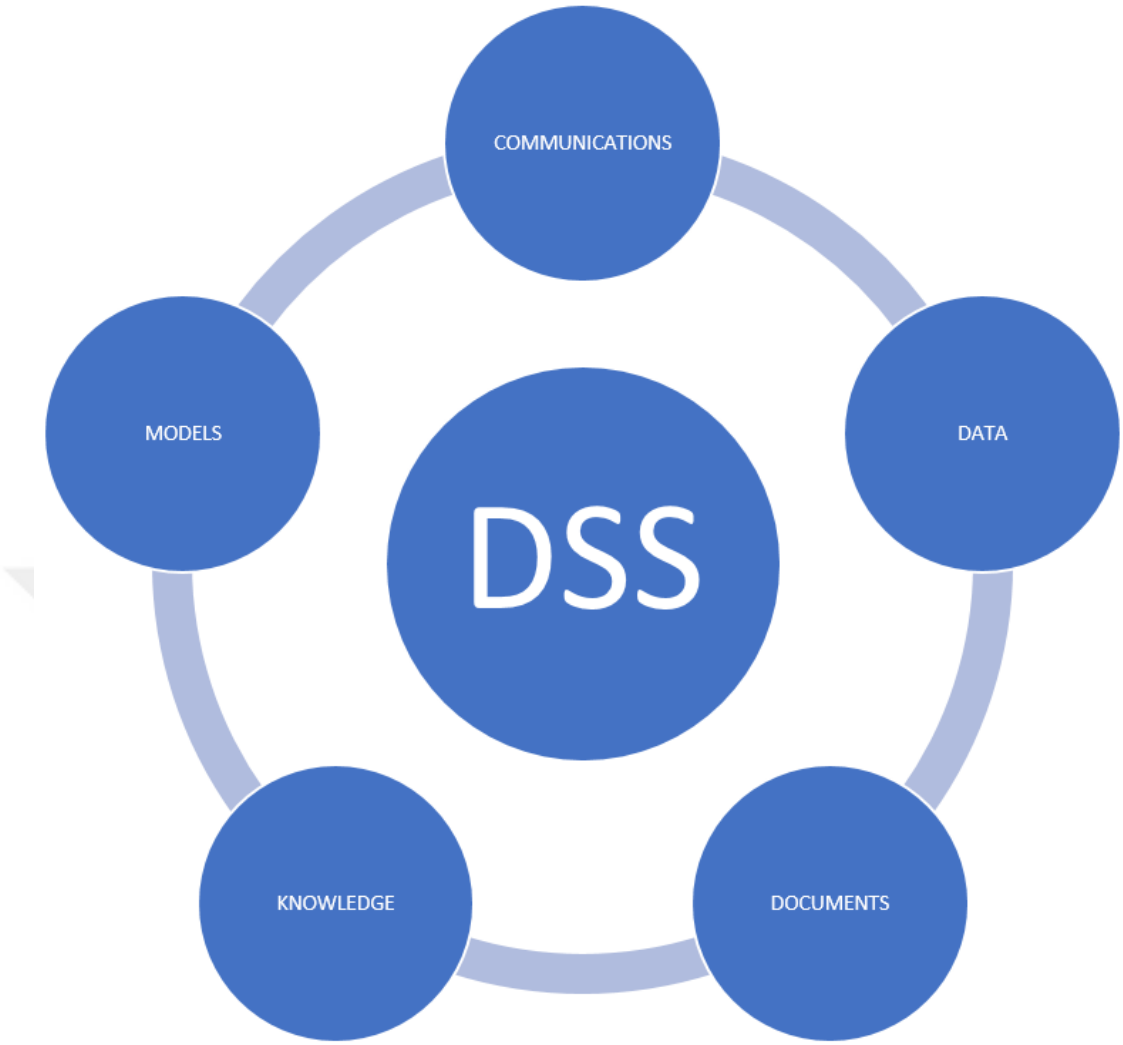

Figure 3.3 General structure of DSS

Decision support system is a system that supports institutional or personal decision making. Decision support system is often designed to help making decisions about issues related to management and planning that cannot be easily identified by the humans. decisions that are made automatically can be fully automated or human supported. In the current system, decisions made automatically can be improved with feedback from doctors. The technologies which are used in the proposed process and general operation of these technologies will be explained in detail in the following sections.

CHAPTER FOUR
SPINE SURGERY DECISION SUPPORT SYSTEM

Spine surgery decision support system consists of a combination of two applications: Text Convert and Diagnostic Support. Firstly, the text entered by professionals are run through the pre-process stage by the Text Convert application. Texts that are cleared from prefixes, suffixes and conjunctions are made ready for learning. After the pre-process phase, the data provided to the "Diagnostic Support System" is trained with 3 different algorithms and made ready for suggestions.

Text Convert application was developed using C# and MSSQL with the help of the Visual Studio. Diagnostic Support was developed using Java and Weka libraries supported by eclipse IDE. The compatibility of the Java with Weka library, which includes all popular learning algorithms are available, has an important role in choosing this language. The technologies used are described in detail below.

4.1 The Technologies

4.1.1 C#

C# is an object-oriented programming language which is become very popular lately. It has been developed and supported by Microsoft. Although it was first introduced to resolve inconsistencies in C ++ and visual basic languages, it has quickly become one of the most popular languages.

Windows and web-based applications have become quite easy to develop thanks to C# supported by .Net. However, .Net Framework is gradually being replaced by the .Net Core. Because of .Net core, developed programs are not depend on operation system. It allows you to develop applications quite quickly and easily on many computer systems with only C#.

Besides cross-platform and powerful frameworks, C# has a one more powerful support and that is Visual Studio. It has an advanced compiler and error checks thanks

to its combined power of Visual Studio. This provides us to ability to create error-safe applications.

4.1.2 Visual Studio 2015

Microsoft Visual Studio is a development environment developed by Microsoft. Visual Studio, which supports a wide range of platforms, from Windows forms to web and mobile application, is one of the best development environments. Visual Studio has become more popular with the release of free versions recently. It is chosen because it is the most suitable IDE for C# and .Net applications.

4.1.3 .Net

Previously, programs were compiled and translated into machine code. This approach changed with Java Virtual Machine. The compiled codes were first translated into intermediate codes and interpreted in real-time according to the operating system. Also .Net is a technology that works with this method too.

After the code is compiled on the .Net platform, it is converted to an intermediate language called Microsoft Intermediate Language. The Just in Time (JIT) compiler is activated when code from the .Net platform runs. The JIT compiler interprets these intermediate codes according to the operating system. The reason for the emergence of the concept of intermediate language is to ensure that the written code will work properly on different systems with minimum effort. The operation of the system is indicated in Figure 4.1.

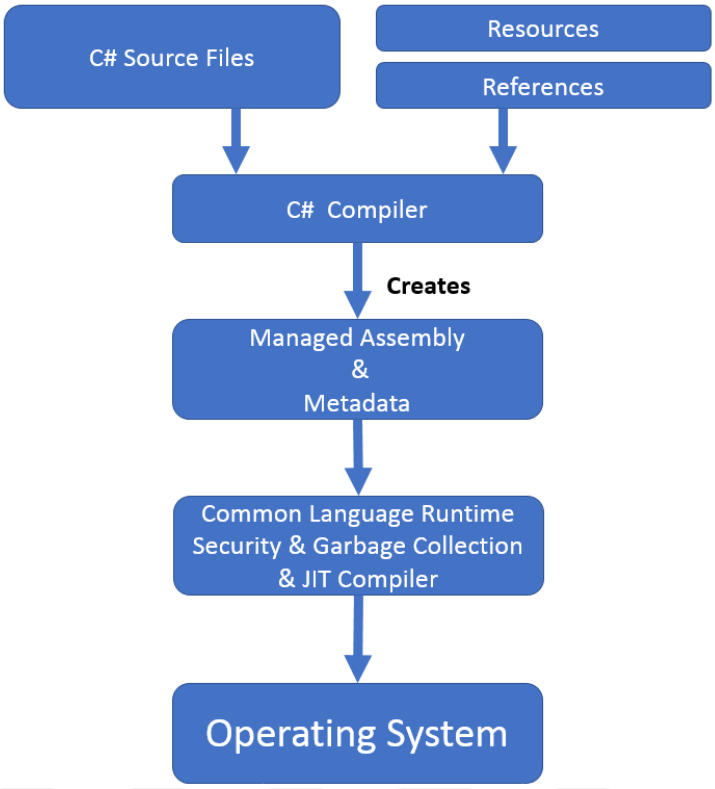

Figure 4.1 Overall of .Net framework

4.1.4 MSSQL – MSSQL Server Management Studio 2012

MSSQL is a database system developed by Microsoft. This system is very popular and can be used in many small and large scope applications. Server Management Studio is a management platform that includes many database managements features.

4.1.5 Java

Java is an open-source, object-oriented, platform-independent, highly efficient, multi-functional, high-level, interpreted language developed by James Gosling, a Sun Microsystems engineer. When it first came out, Java was thought to be a common language designed for use on smaller devices. However, the platform independent feature and effective library support offer a far superior and secure software development and operating environment than C and C ++ and are now being used almost everywhere.

The written codes are converted to an intermediate form called Bytecode. This bytecode is run by the Java Virtual Machine (JVM). JVM executes java codes by interpreting each bytecode command one by one depending on platform.

4.1.6 Eclipse Oxygen

Eclipse is a non-profit and open source IDE. Eclipse, which could be enhanced with plug-ins, has a significant share in the market. Many technologies such as Java, C and android are developed using eclipse.

4.1.7 Weka

The Waikato Environment for Knowledge Analysis (WEKA) came about through the perceived need for a unified work-bench that would allow researchers easy access to state-of-the-art techniques in machine learning (Hall et al., 2009). Weka, which includes many machine learning algorithms, provides a great convenience when testing data mining applications. In addition, the library which is provided by Weka for java makes it easy to use data mining in applications.

4.2 System Architecture

In the studies conducted, it is aimed to develop a system to assist the doctor during diagnosis process. Patients' narratives were examined and tested to see if a successful decision support system could be implemented (Utku et al., 2010). Firstly, the relations between the words used in patients' narratives and diagnosis were examined. In this direction, patients' narrative texts were cleared from conjunctions and special characters. Then, it was aimed to remove the suffix in the words by determining the roots of them. Subsequently, the texts swere divided into classes according to the diagnostic codes. The most common 40 words for each diagnostic code were determined and the frequency of these words in the text was calculated. Subsequently, a demo decision support system study was conducted in the light of this information which could assist in diagnosis during the examination.

It is also aimed to develop a system to assist the doctor during treatment or diagnosis. This system, which can continue learning with new arrivals, will be able to make stronger predictions day by day. The number of algorithms used is more than one, so it is more likely to check the error and make a correct guess and to find better suggestions. The general operation of the system is shown in Figure 4.2.

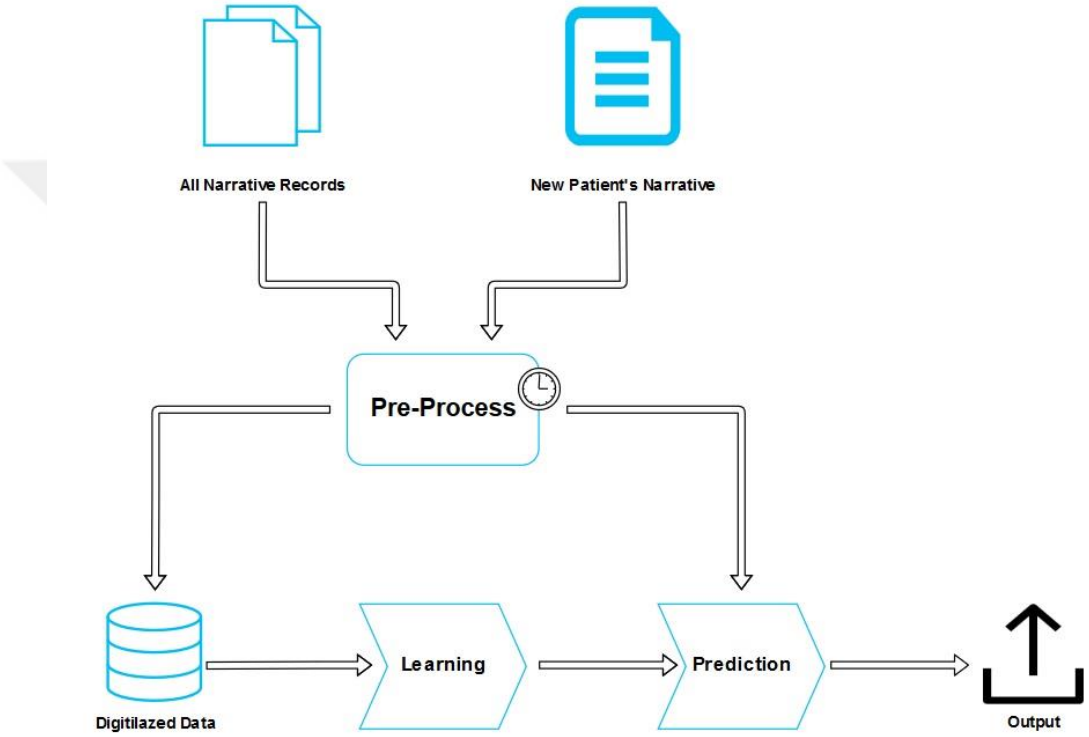

Figure 4.2 Overall operation of the system

The study can be examined under two headings. The first is the stage in which the raw texts entered by the professionals are made ready for learning. After this stage, texts are converted to numerical data and made ready for learning and then numerical data are recorded. Those operations are done automatically by an application named "Text Convert". The digitized data is then directed to the “Decision Support” application for learning and suggestion processing.

Models are created according to data that Decision Support application received, after they are trained with three preselected algorithms. New incoming data are going

to be digitized and predictions will be made according to those models. Among the results, the most common result is obtained and referred to the professionals.

4.2.1 Text Convert

Pre-processing in text mining could be defined as workflow that involves processes related to words that are the building blocks of text documents such as textual separation, finding semantic value of the words, separating prefixes and suffixes and determining the meaning of the words. As illustrated in Figure 4.3, patient narratives are digitized by pre-processing before being examined.

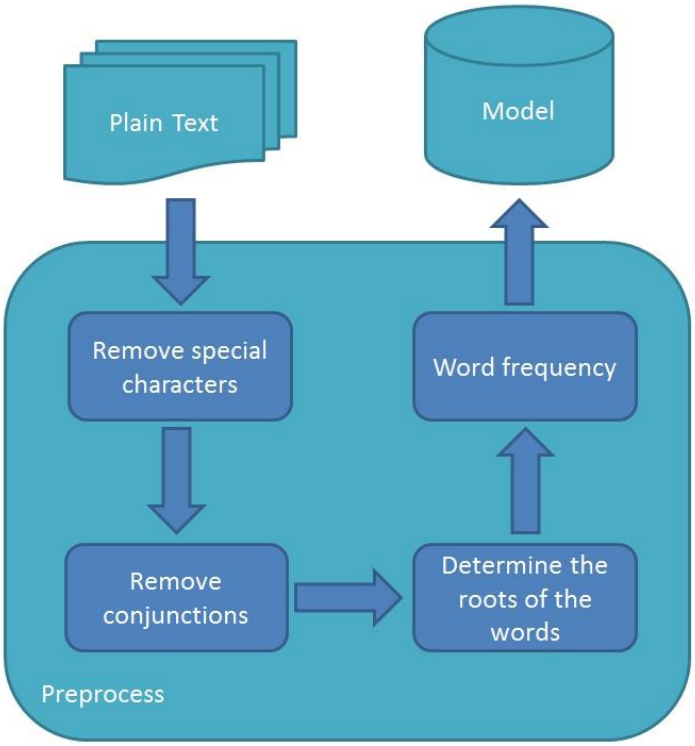

Figure 4.3 Preprocessing of narrative text

In that sense, as a first step, the text was cleared from conjunctions and special characters with this process. Therefore, the texts are scanned completely and any characters without numbers or letters are removed from the texts. Text that eliminates the special character is then cleared from the conjunctions too. In this context, the list of conjunctions obtained from the Turkish Language Institution (Türk Dil Kurumu,

n.d.) was utilized. According to information that been taken recently, there have been 35 conjunctions. Each word in the texts is checked one by one and if there is a match in this conjunctions list, the matched word is removed from the text. These operations are shown in Figure 4.4. After this process, the next step is to clear the words from the suffixes.

```
private string ClearSpecialCharactersAndConjunctions(string word)
{
    if (string.IsNullOrEmpty(word))
    {
        return "";
    }
    var value = "";
    for (int i = 0; i < word.Length; i++)
    {

        if (char.IsLetterOrDigit(word[i]))
        {
            value += word[i];
        }
    }
    if (Conjunctions.Contains(value.ToLower()))
    {
        value = "";
    }
    return value;
}
```

Figure 4.4 Remove special characters and conjunctions

After the texts are cleaned from special characters and conjunctions, the next stage of operation is to clear the words from any suffix or prefix. At this stage, the Zemberek library (Akın, & Akın, 2007) was used. Zemberek is a natural language processing library developed especially for Turkish. This library with support for Java and C # is very useful for Turkish language processing.

Separating words from suffixes and prefixes is a more complex task compared to other preprocess tasks. Firstly, the text separated from special characters and conjunctions is broken into pieces by the Zemberek library and it finds each word's root. For later calculations, this word root is used. For words with more than one

meaning, the first meaning in the dictionary was considered as valid. This algorithm is visualized in Figure 4.5.

```
private string GetRoots(string [] sentence)
{
    StringBuilder builder = new StringBuilder();
    Zemberek zem = new Zemberek(new TurkiyeTurkcesi());
    foreach (var item in sentence)
    {
        if (string.IsNullOrWhiteSpace(item))
        {
            continue;
        }
        var words = zem.kelimeCozumle(item);
        if (words == null || words.Length == 0)
        {
            builder.Append(item);
            builder.Append(" ");
        }
        else
        {
            builder.Append(words[0].kok().icerik());
            builder.Append(" ");
        }
    }
    return builder.ToString();
}
```

Figure 4.5 Getting roots algorithm

Simplification of texts is a very long process. Since there are 65606 stories in total, preprocess cannot be done often to examine and clear all texts with the algorithms mentioned above. Therefore, the texts are examined, and the information obtained is saved in the MSSQL database. This step which is prior to digitization was done to prevent any time loss that could occur in each trial. The database model in which the information is kept is shown in Figure 4.6.

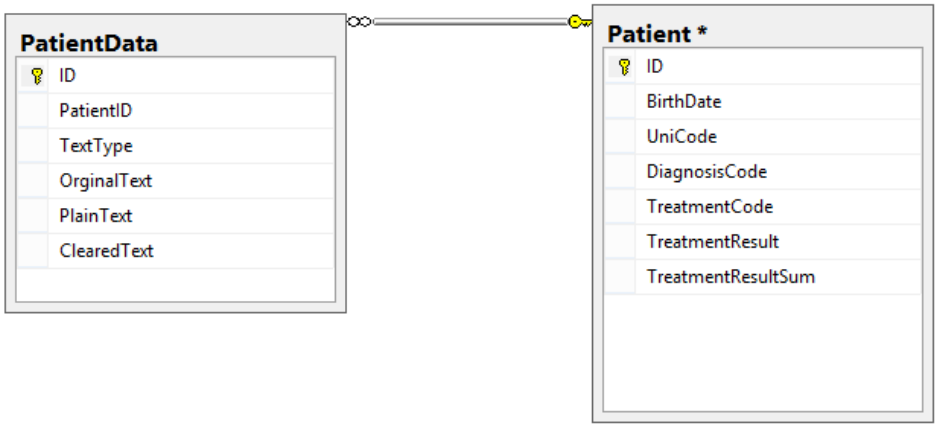

Figure 4.6 Text convert database model

"Patient" table contains information about patients. Patients' birth dates, diagnoses, treatments and results are kept in this table. The narratives about patients are kept in the "PatientData" table. All patients' narratives are kept in all three version. All patients' narratives are kept in all three version: Original narratives, allowed only versions of original narratives and narratives that are ready to be digitized. This approach has been adopted to check whether the text cleanup steps work successfully or not and to prevent data loss. Tables and fields' equivalents are explained detailed below.

Patient:

ID: Unique identification number for patients.

BirthDate: The column where patients' birth dates are stored.

UniCode: The unique id of the database from which the data was received. Added to prevent the same patients from being transferred repeatedly.

DiagnosticCode: It is the diagnostic information code that is put on the patient. It is a unique code for each diagnosis.

ThreatmentCode: It is the code of treatment. It is unique for every kind treatment.

ThreatmentResult: It keeps whether the treatment is successful or not.

ThreatmentResultSum: Summary of treatment outcome.

PatientData:

ID: Unique code for all narratives.

PatientId: Unique number for all patients. The relationship to the Patient table is provided through this column.

TextType: Has no function currently. It is added that there may be a need for classification in the future.

OriginalText: This column keeps original version of patient narratives.

PlainText: This column keeps all characters that are not numbers or letters.

ClearedText: This column keeps cleaned narratives after preprocess.

Once the texts are simplified, they are ready to digitize. In this context, the words in the texts are examined and the frequency of each word in the text is subtracted. The frequency range is taken dynamically from user interface. First, all texts are scanned, and the frequency of each word is calculated. Obtained frequencies according to the desired attribute number is examined and the resulting matrix is obtained from most frequent words. The resulting matrix is combined with the diagnostic code information to obtain ready-to-learn files. The screenshot of the application is shown in Figure 4.7.

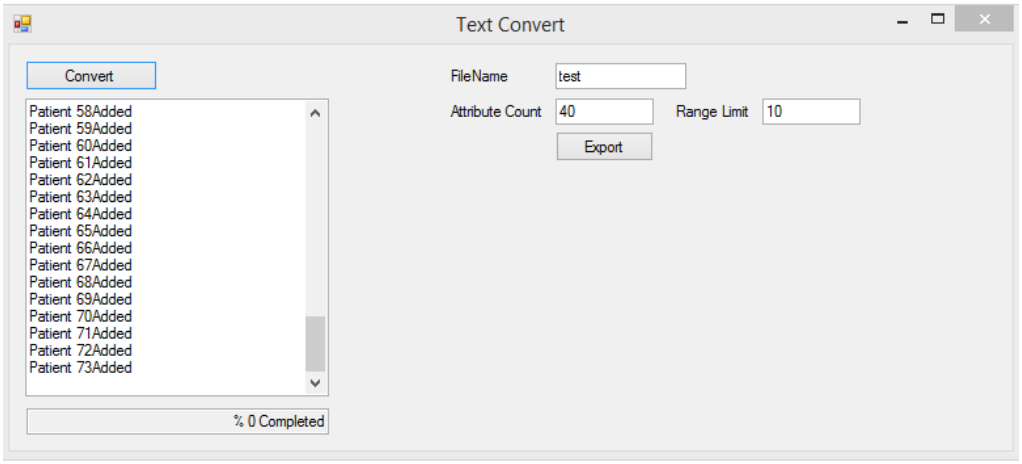

Figure 4.7 The screenshot of the text convert

This screenshot shows 2 different parts of the application. First, the “Convert” button takes the patient narratives and saves them to the database after the text is cleaned. At the same time, on the screen, we can see the log of the transferred patients and the percentage of transfer.

Another part of the application is seen on the right. In this section, there are parameters related to digitized data and Export button. After entering the file name to record the data, how many attributes to select from each subject and frequency max limit are specified. The "Attribute Count" field represents the number of words to select from the text. If 2 is entered, the most frequently used two words are selected and the file is created accordingly. "Range Limit" corresponds to the frequency limit of the words in the text. For example, if two is entered, the frequency will be left as two if the frequency is two or more.

In the scope of this study, patient narratives were divided into classes according to the diagnostic codes. The most common 40 words for each diagnostic code were determined and the frequency of these words set as 10 in the narratives were calculated. Our examples have shown that the most successful results will be obtained with these data for this data set.

After the patients' narratives were edited and digitized, the text is brought to the form that can be applied to learning. Subsequently, data is stored in database. These processes have been automated by the program "Text Convert" which is programmed for preprocessing of narrative texts.

The text convert application was developed using Winform and .Net frameworks. Also powered by MSSQL data, this application has the capacity to analyze millions of patient records and generate the necessary results. In addition, each step was tested to ensure the accuracy of the digitized data generated. After the preprocess levels, the data are now ready to train with. From this point on, the Decision Support Application manages the required tasks.

4.2.2 Decision Support Application

The main purpose of the study is to implement a system of suggestions that will help medical staff in real-time by processing the information held in the text. Algorithms learned from cleaned data were used to provide diagnostic predictions by processing the last entered text information.

All learning methods selected at the beginning of the application are trained with the data provided. Since the application is intended to serve the professionalist in real-time, models should be prepared before starting the estimation process. It is necessary to wait before the start of the application in order to make immediate and quick feedback to the professionalist.

Since the application is powered by Weka libraries, any learning algorithms can be used. The application has a support of more than one algorithm parametrically. In the development stages of the application, many different algorithms were used and tested.

In the studies performed on the digitized data, the three most successful algorithms were selected and learning processes were performed separately for each algorithm.

The top three algorithms which are selected are as follows;

- Random Committee
- Sequential Minimal Optimization (SMO)
- Iterative Classifier Optimizer

These algorithms were selected by comparing the obtained data according to the results of the tests applied separately. Many algorithms suitable to the data set were tried and the results were compared. As a result, the best three algorithms have been chosen.

In the random committee classifier, the seeds generating different random numbers build various base classifiers. The individual base classifier determines the final prediction using a straight average of the predictions for every base classifier (Huang, Hung, & Jiau, 2006). Sequential Minimal Optimization (SMO) is a simple algorithm that can quickly solve the SVM QP (Quatratic Programming) problem without any extra matrix storage and without using numerical QP optimization steps at all (Platt, 1998). In iterative classification, a model is built using a variety of static and dynamic attributes. Classifier that includes dynamic attributes relies on the previous classification of related objects. When training the model, the class labels of all objects are known and consequently the values of all dynamic attributes are also known (Neville, 2000).

The reason which the diagnostic proposal system does not use a single algorithm but merge three algorithms is that more successful and more homogeneous results were obtained when more than one algorithms were used. However, tests have shown that when more than three algorithms were used, the success rate started to fall again. For this reason, these three algorithms were used together in the decision support system developed.

```
Begin
    Initialize ClassifierArray; //array for classifiers with 3 elements
    Initialize resultArray array as length ClassifierArray;
    Read newInput; //New patients' narrative
    newInput = pre-process(newInput);

    foreach classifer in ClassifierArray
        resultArray[index] = classifer.classify(newInput);
    End of loop

    return resultArray.getPopularElement();
End
```

Figure 4.8 The algorithm used for decision making

After the data entered by the medical personnel are passed through the text cleaning methods mentioned above, a separate prediction process is performed for each of the algorithms. After performing the forecasting by the algorithms that mentioned above, the most common estimation among them is considered correct. The result with the highest probability of success is presented to the doctor. Algorithm of decision making shown in Figure 4.8.

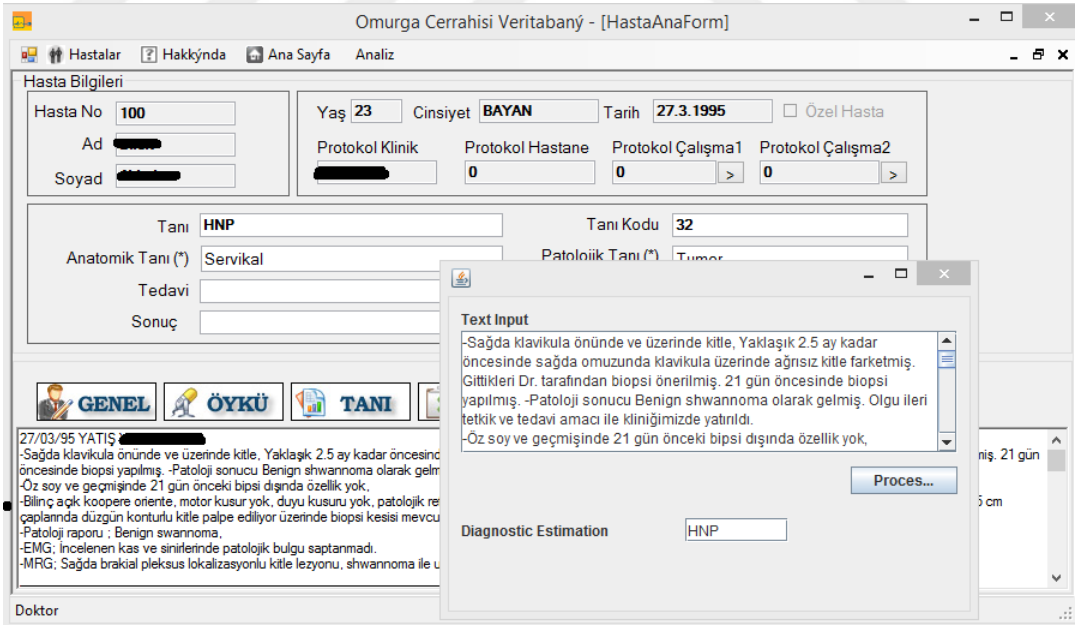

Figure 4.9 The screenshot of Diagnostic Decision Support System

If the diagnostic information checked by the professional appears incorrect, the diagnosis is expected from the professional to be corrected. Corrected diagnostic

information is added to pre-formed models so that the system can continue learning. Thanks to the model that will develop over time, the application would make more successful predictions and can achieve a much higher success than the information currently obtained. Suggestion dialog is shown in Figure 4.9.

CHAPTER FIVE
SYSTEM TESTING AND EXPERIMENTAL RESULTS

The data used for system testing is texts containing patient narratives that were entered by doctors which are specialists in their fields. The number of patients registered in the system is 31,830 and there are 79 different diagnostic groups. However, due to large differences in the number of patients in the diagnostic groups, not all data was used during the test phase.

In order to ensure a more systematic testing of the system, only diagnostic groups with 500 or more patients were taken into consideration. 500 patients were randomly selected from each diagnostic group and used for system testing. The diagnostic groups that meet these requirements are listed below:

- Herniated Nucleus Pulposus (HNP)
- Spinal Stenosis (SS)
- Cervical Disc Herniation (CDH)
- Sciatica Pain(SP)
- Facet Joint(FJ)
- Trigeminal Neuralgia (TN)
- Cervical Arthrosis (CA)

500 narratives selected randomly from each of the diagnostic groups mentioned above were divided into training and trial classes. 80% of the data was obtained from each diagnosis group, which means 400 for each group (2,800 in total) were used in system learning. The remaining 20% were used for system testing (700 stories in total).

Table 5.1 Test results

	HNP	**SS**	**CDH**	**SP**	**FJ**	**TN**	**CA**	**Success Rate**
HNP	64	10	5	6	13	0	2	%64
SS	7	78	1	7	5	0	2	%78
CDH	15	2	75	0	0	0	8	%75
SP	6	3	1	80	8	0	2	%80
FJ	13	5	0	23	58	0	1	%58
TN	0	0	1	1	1	97	0	%97
CA	6	1	5	4	1	0	83	%83

The proposed system success rates are shared in Table 5.1. For each diagnostic group, 100 control groups were selected and diagnostic predictions were made. The test results obtained are as shown in Table 5.1. The actual values of the diagnostic groups are shown in the rows, and the predicted values are indicated in the columns.

As the results show, Average success rate of Random Committee, SMO and Iterative Classifier Optimizer combination is calculated as %77. Minimum success rate is 58 and the maximum success rate is %97. The results of the studies show that successful suggestions can be made by the help of the proposed system, and it has been observed that the system may have significant contributions in increasing the quality of the treatment of patients.

In the spinal surgery, almost all of the patients' symptoms and treatment stages are kept in the text format (Symptoms, therapeutic approaches and factors affecting treatment outcome). In our proposed system, it is aimed to reveal this information and extract useful information patterns. Once this information is received and grouped, the data was processed with text mining methods. Furthermore, by the use of the information obtained, suggestion system has been developed.

Different text mining methods were used to extract patterns and knowledge from patient narratives recorded by the specialist doctors. The results obtained were compared and the most successful algorithms were determined. Subsequently, success rates of using multiple algorithms were examined to increase the accuracy of the proposed system, and it was determined that the most successful method was the use of the three algorithms together. The suggestion system was developed accordingly.

The suggestion system may offer real-time suggestions to the doctor. Patients' narratives that lastly entered are compared with the previous narratives and the result with the highest probability of success is presented to the doctor as a suggestion. Also, feedbacks from doctors will be used as contributions to the system. Thus, it was aimed that future predictions will be more successful.

When the results are examined, the fact that the success rate of the proposed system is quite high will be seen. The proposed system can be very useful to help a final diagnosis. In medical specialty such as spinal cord surgery where there are very few experienced doctors, the proposed system may provide important contributions. Patient satisfaction may be increased to affect the life quality.

These results show that text mining, which can be successfully used in many areas, has achieved considerable success in the field of spinal surgery. The information that has not been studied for years has been examined and important knowledge has been obtained. The recommended system is a vital contribution to the existing structure when it is assumed that the false diagnosis and treatment possibilities are very high and vital in the medical field.

CHAPTER SIX
CONCLUSION

Studies shows that majority of the hospital records are in text form. When those records are examined, it was seen that they contain viral information about diseases and patients. In important field which is there are a few experienced specialists, it is a serious issue that such an information stays idle. This study focusses on resolving this issue.

In this study, patient narratives obtained from spinal cord surgery patients were examined with text mining methods and a successful decision support system was was created. The developed system provides assistance in diagnosis prediction with patient narratives. Developed decision support system improves itself with new incoming data and gives better decisions at a time.

In order to develop a successful decision support system, thousands of patients' narratives were examined and possible patterns were determined. After all the narratives cleaned with preprocess methods, they were studied with datamining methods. Patterns were extracted from the information and relationships between the patterns were examined. The relationship between extracted knowledge and diagnoses was examined.

The main focus of the study is to develop a decision support system. With the developed decision support system, it is aimed to prevent making false diagnosis for specialist. The system details are explained and the results are shared. The usage of more than one algorithm at the same time were examined in order to implement a successful decision support system and the system is developed accordingly.

The decision support system provides real-time support to the specialist. In addition, in case of false prediction, feedback given by the doctor was able to use for system learning. In this way, it was aimed to made more accurately predictions by the time.

Text mining via information extraction and data mining can provide a real advantage when used with a normalized spinal data to describe health care, adding new medical material and to help to integrate the electronic health record system into the staff work environment. The results of the studies show that successful suggestions can be made with the proposed system and it has been observed that the system may have significant contributions to increase the quality of the treatment of patients.

Test results show that the suggested system have achieved a remarkable success. The proposed system could avert many errors and misdiagnosis. In addition, patient dissatisfaction can be eliminated and complications that are results of false treatment can be avoided by increasing the quality of the treatment. These measures are essential in an area such as spinal surgery which directly affects patients' health and quality of their lives.

The system can be supported by an automation system that will be created for the hospital and can reduce workload by reducing non-digitized records within the hospital. In this way, a more organized system is created and the recommendations of the implemented system can be used to help more accurate treatment. Therefore, the data that will be collected by the system will found a basis for the future applications.

Appropriate treatment for diagnosis can be suggested by providing to learn with the treatment methods of the patients. In addition, by adding patient information such as gender and age to the system, the relationship between these parameters can be examined. Diagnosis and treatment suggestion system based on the groups of incoming patients can be developed.

Studies have shown that the present system contains many deficiencies. It seems that Spine Surgery Decision Support System will be able to provide a significant effect to a current situation. When the deficiencies and errors in the current system are considered, it will be seen how this effect can be. Considering the additional features that can be added to the application, the study will contribute paperless hospitals and will also provide idle knowledge and experience to use properly in this field.

REFERENCES

Akın, A. A., & Akın, M. D. (2007). Zemberek, an open source nlp framework for turkic languages. *Structure*, *10*, 1-5.

Alex, B., Grover, C., Tobin, R., Sudlow, C., Mair, G., & Whiteley, W. (2019). Text mining brain imaging reports. *Journal of Biomedical Semantics, 10*(S1), 23.

Bernstein, M., Hebert, P. C., & Etchells, E. (2003). Patient safety in neurosurgery: detection of errors, prevention of errors, and disclosure of errors. *Neurosurgery Quarterly, 13*(2), 125-137.

Carroll, J., Koeling, R., & Puri, S. (2012). Lexical acquisition for clinical text mining using distributional similarity. *International Conference on Intelligent Text Processing and Computational Linguistics,7182* ,232-246.

Delespierre, T., Denormandie, P., Bar-Hen, A., & Josseran, L. (2017). Empirical advances with text mining of electronic health records. *BMC Medical Informatics and Decision Making*, *17*(1), 127.

Densen, P. (2011). Challenges and opportunities facing medical education. *Transactions of the American Clinical and Climatological Association*, *122*, 48.

Diasio, S. R., & Agell, N. (2009). The evolution of expertise in decision support technologies: A challenge for organizations. *13th International Conference on Computer Supported Cooperative Work in Design,* 692-697.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine, 17*(3), 37-37.

Ghosh, S., Roy, S., & Bandyopadhyay, S. (2012). A tutorial review on text mining algorithms. *International Journal of Advanced Research in Computer and Communication Engineering*, *1*(4), 7.

Gönenç, E. Ö. (2007). İletişimin tarihsel süreci. *Istanbul University Faculty of Communication Journal*, (0)28 87-102.

Gupta, S., Ross, K. E., Tudor, C. O., Wu, C. H., Schmidt, C. J., & Vijay-Shanker, K. (2016). miriad: A text mining tool for detecting associations of micrornas with diseases. *Journal of Biomedical Semantics, 7*(1), 9.

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence*, *1*(1), 60-76.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10-18.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, *44*(1), 1-12.

Health,(2009).*Adverse Health Events in Minnesota Annual Report. Minnesota Department.* Retrieved December 30, 2018, from https://www.health.state.mn.us/facilities/patientsafety/adverseevents/docs/2019a h ereport.pdf

Hearst, M. A. (1999). Untangling text data mining. *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, 3-10.

Henneman, P. L., Blank, F. S., Smithline, H. A., Li, H., Santoro, J. S., Schmidt, J., Henneman, E. A., et al. (2005). Voluntarily reported emergency department errors. *Journal of Patient Safety*, *1*(3), 126-132.

Huang, Y. M., Hung, C. M., & Jiau, H. C. (2006). Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem. *Nonlinear Analysis: Real World Applications*, *7*(4), 720-747.

Jen, C. H., Wang, C. C., Jiang, B. C., Chu, Y. H., & Chen, M. S. (2012). Application of classification techniques on development an early-warning system for chronic illnesses. *Expert Systems with Applications*, *39*(10), 8852-8858.

Jonnagaddala, J., Liaw, S. T., Ray, P., Kumar, M., Chang, N. W., & Dai, H. J. (2015). Coronary artery disease risk assessment from unstructured electronic health records using text mining. *Journal of Biomedical Informatics*, *58*, *203-210.*

Keen, P. G. (1978). *Decisión support systems; an organizational perspective* Boston: Addison-Wesley.

Lovell, M. C. (1983). Data Mining. *The Review of Economics and Statistics 65*(1), 1-12.

Luhn, H. P. (1958). A business intelligence system. *IBM Journal of research and development*, *2*(4), 314-319.

Marwaha, R. (2014). Data mining techniques and applications in telecommunication industry. *International Journal of Advanced Research in Computer Science and Software Engineering 4*(9), 430-433.

Michelson, J. D., Pariseau, J. S., & Paganelli, W. C. (2014). Assessing surgical site infection risk factors using electronic medical records and text mining. *American Journal of Infection Control*, *42*(3), 333-336.

Neville, J. (2000). *Iterative classificatio: Applying bayesian classifiers in relational data.* Retrieved December 25, 2018, from https://pdfs.semanticscholar.org/7074/b3209de8f30e26083596a217fd9e148b99c8.pdf.

Palmieri, P. A., DeLucia, P. R., Peterson, L. T., Ott, T. E., & Green, A. (2008). The anatomy and physiology of error in adverse health care events. *Patient Safety and Health Care Management*. 33-68

Platt, J. C. (1998). *Sequential minimal optimization: A fast algorithm for training support vector machines.* Retrieved December 29, 2018, from http://www.bradblock.com/Sequential_Minimal_Optimization_A_Fast_Algorithm_for_Training_Support_Vector_Machine.pdf.

Rupnik, R., Kukar, M., Bajec, M., & Krisper, M. (2006). DMDSS: Data mining based decision support system to integrate data mining and decision support. In *28th International Conference on Information Technology Interfaces*, 225-230.

Schulman, P., Castellon, C., & Seligman, M. E. (1989). Assessing explanatory style: The content analysis of verbatim explanations and the attributional style questionnaire. *Behaviour Research and Therapy*, *27*(5), 505-509.

Srinivas, K., Rani, B. K., & Govrdhan, A. (2010). Applications of data mining techniques in healthcare and prediction of heart attacks. *International Journal on Computer Science and Engineering*, *2*(2), 250-255.

The Joint Commission, (n.d.). Retrieved November 16, 2018, from https://www.jointcommission.org/issues/article.aspx?Article=AVw4k9l0%2FHrWYnT7qVs9NrWBCgOfdlZemMC0iyFsJUc%3D

Türk Dil Kurumu, (n.d.). Retrieved February 16, 2019 from http://www.tdk.gov.tr/

Utku, S., Baysal, H., & Zileli, M. (2010). Spine surgery database: A turkish registry for spinal disorders. *Turkish Neurosurgery, 20*(2), 223-230.

Weingart, N. S., Wilson, R. M., Gibberd, R. W., & Harrison, B. (2000). Epidemiology of medical error.*BMJ, 320*(7237), 774-777.