CLASSIFICATION OF MEDICAL DOCUMENTS ACCORDING TO DISEASES

Bekir PARLAK

MASTER THESIS

Computer Engineering Program

Supervisor: Asst. Prof. Dr. Alper Kürşat UYSAL

Eskişehir
Anadolu University
Graduate School of Sciences
June, 2016

This thesis is supported by Anadolu University, Fund of Scientific Research Projects under grant no. 1503F136.

JÜRİ VE ENSTİTÜ ONAYI
(APPROVAL OF JURY AND INSTITUTE)

Bekir PARLAK’ın **“Classification of Medical Documents According to Diseases”** başlıklı tezi 20/06/2016 tarihinde, aşağıdaki jüri tarafından değerlendirilerek “Anadolu Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliği”nin ilgili maddeleri uyarınca, **Bilgisayar Mühendisliği** Anabilim dalında Yüksek Lisans tezi olarak kabul edilmiştir.

		Adı Soyadı	**İmza**
Üye (Tez Danışmanı)	**:**	**Yrd. Doç. Dr. Alper Kürşat UYSAL**	**…………..**
Üye	**:**	**Yrd. Doç. Dr. İbrahim YAKUT**	**…………..**
Üye	**:**	**Yrd. Doç. Dr. Semih ERGİN**	**…………..**

………………..

Enstitü Müdürü

ÖZET

Yüksek Lisans Tezi

TIBBİ DOKÜMANLARIN HASTALIKLARA GÖRE SINIFLANDIRILMASI

Bekir PARLAK

Bilgisayar Mühendisliği Anabilim Dalı

Anadolu Üniversitesi, Fen Bilimleri Enstitüsü, Haziran, 2016

Danışman: Yrd. Doç. Dr. Alper Kürşat UYSAL

Bilgisayar kullanımının yaygınlaşmasından sonra, bilgisayar ortamında üretilen dokümanların sayısının her geçen sene ivmeli olarak arttığı görülmektedir. İnternet ortamında metinlerin üssel artışından dolayı otomatik metin sınıflandırma önemli hale gelmiştir. Metin sınıflandırmadaki önemli sorunlar öznitelik sayısının çok olması ve buna bağlı olarak yapılan hatalı sınıflandırmalardır. Bu tez çalışmasında, Türkçe makalelere ait tıbbi metin özetleri kullanılarak İngilizce ve Türkçe içerikli medikal alanda iki farklı veri kümesi oluşturulmuştur. Bu veri kümesi İngilizce tıbbi metin özetleri içeren Ohsumed isimli veri kümesine benzer yapıdadır. Literatürde akademik çalışmalarda kullanılmak üzere Türkçe kaynaklardan elde edilen Ohsumed benzeri bir veri kümesi bulunmamaktadır. Otomatik metin sınıflandırma aşamalarında çeşitli ön işlem, öznitelik seçim yöntemleri ve bu alanda başarılı sınıflandırıcılar kullanılmıştır. Ayrıca diller bazında farklılık gösteren ve ön işleme adımlarından biri olan kök bulma algoritmasının uygulanıp uygulanmamasına göre sınıflandırma başarımının nasıl etkilendiği diller bazında incelenmiştir. Bunun yanı sıra, farklı öznitelik seçim yöntemlerinin sınıflandırmadaki başarımı nasıl etkilediği incelenmiştir. Başarımı etkileyen bir diğer etken olan sınıflandırıcı performansları farklı sınıflandırıcıların uygulanması ile analiz edilmiştir. Son olarak ta, aynı yayınlara ait farklı dillerdeki tıbbi metin özetleri üzerinde en iyi başarımı sağlayan sınıflandırma şemaları belirlenmiştir.

Anahtar Sözcükler: Metin Sınıflandırma, Öznitelik Seçim Yöntemleri, Sınıflandırma Algoritmaları, Önişleme Adımları

ABSTRACT

Master of Science Thesis

CLASSIFICATION OF MEDICAL DOCUMENTS ACCORDING TO DISEASES

Bekir PARLAK

Department of Computer Engineering

Anadolu University, Graduate School of Sciences, June, 2016

Supervisor: Asst. Prof. Dr. Alper Kürşat UYSAL

The number of documents produced on computers has increased exponentially every year, after the spreading use of the computers. Automatic text classification has become an important due to the exponential growth of texts on the Internet. Significant problems in text classification are the great number of features and misclassification are made accordingly. In this thesis, it is constructed of two different datasets containing English and Turkish abstract belonging to Turkish articles in the medical field. This dataset is similar structure to namely Ohsumed which is containing English medical text summary. In the literature, there is no dataset like Ohsumed datasets obtained from Turkish datasets to be used in academic studies. Various preprocessing, feature selection and successful classifiers in this field are used in automatic text classification stages. It has been investigated in the basis of languages how influences the performance of the classification according to whether stemming which differs in languages and one of the preprocessing steps applied or not. And also, the classification performance of different feature selection method has been investigated. Classifier performance which is another factor affecting the performance was analyzed by applying different classifiers. Finally, classification schemes that provide the best performance on the medical text summary in the same publication and different languages is determined.

Keywords: Text Classification, Feature Selection Methods, Classification Algorithms, Preprocessing Steps

ACKNOWLEDGEMENTS

I would like to thank my supervisor Assist. Prof. Dr. Alper Kürşat UYSAL for his support and guidance to my research. I have taken his great amount of time and effort during research studies up-to-now.

I would also like to thank Assist. Prof. Dr. İbrahim YAKUT and Assist. Prof. Dr. Semih ERGİN for serving on my dissertation committee and for their valuable contributions.

Also, I am grateful to Res. Asst. Muhammet Yasin PAK for sharing the valuable information.

A heartful thanks to go to my parents. I am very grateful for their devotion and efforts on me.

Finally, I would like to thank my wife for her substantial support and patience during all my studies.

.

Bekir PARLAK
June, 2016

20.06.2016

STATEMENT OF COMPLIANCE WITH ETHICAL PRINCIPLES AND RULES

I hereby truthfully declare that this thesis is an original work prepared by me; that I have behaved in accordance with the scientific ethical principles and rules throughout the stages of preparation, data collection, analysis and presentation of my work; that I have cited the sources of all the data and information that could be obtained within the scope of this study, and included this sources in the references section; and that this study has been scanned for plagiarism with “scientific plagiarism detection program” used by Anadolu University, and that “it does not any plagiarism” whatsoever. I also declare that, if a case contrary to my declaration is detected in my work at any time, I hereby express my consent to all the ethical and legal consequences that are involved.

Bekir PARLAK

CONTENTS

Sayfa

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

BN	: Bayes Net
BoW	: Bag-of-Words
DFS	: Distinguishing Feature Selector
DR	: Dimension Reduction
DT	: Decision Tree
FS	: Feature Selection
GI	: Gini Index
IG	: Information Gain
MNB	: Multinomial Naive Bayes
RF	: Random Forest
TF	: Term Frequency
TF-IDF	: Term Frequency - Inverse Document Frequency

1. INTRODUCTION

Surprising development of internet technology and digital library has initiated a lot of research areas. Especially in recent years, owing to high availability of computing facilities, huge amount of data in electronic format is created. By virtue of the increasing exponentially number of documents in digital format, automated text classification has become more up-and-coming for a long time. Text classification that is also known as text categorization is widely used when organizing documents in a digital format. Text classification can be defined as assigning new documents to a set of annotated categories based on their contents (Al Zamil and Can 2011, Uysal and Gunal 2012). Text classification can be used to solve a miscellaneous of problems such as the filtering of spam e-mails (Kaya and Ertuğrul 2016, Sharma and Kaur 2016), SMS spam filtering (Uysal, Gunal et al. 2012, Goswami, Singh et al. 2016), topic detection (Chang, Hsieh et al. 2015), author identification (Narayanan, Paskov et al. 2012, Bay and Çelebi 2016), language identification (Takçı and Güngör 2012, Zhang, Clark et al. 2016), web page classification (Onan 2015, Belmouhcine and Benkhalifa 2016), medical document classification (Jindal and Taneja 2015, Parlak and Uysal 2015) and sentiment analysis (Agarwal and Mittal 2016, Pak and Gunal 2016). Majority of the researchers, conducted within the field of text classification, studied with text documents in English, German, French and Chinese. On the other hand, there are fewer researchers dealing with others such as Turkish language.

A typical text classification framework contains preprocessing, feature extraction, dimension reduction and document classification phases. Although dimension reduction can be accomplished either by feature selection or feature transformation, feature selection is widely preferred than feature transformation because of some concerns such as computational complexity. Structure of a text classification framework is shown in Figure 1.1. Also some short remarks are given about this framework in the following sentences.

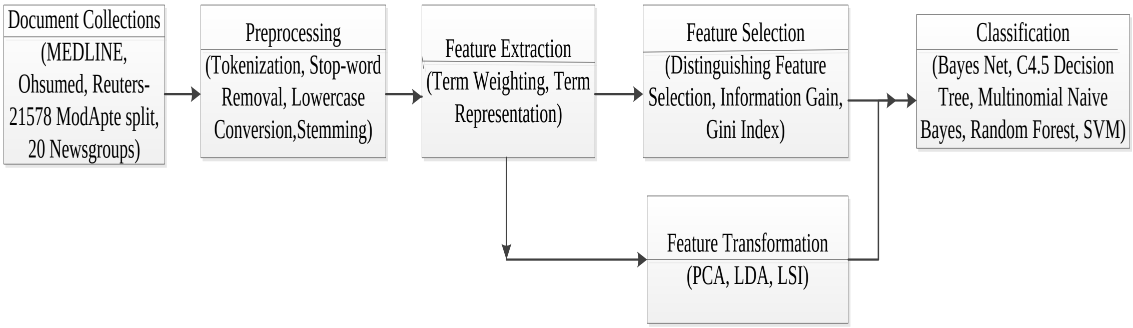
Document Collections
(MEDLINE, Ohsumed, Reuters-21578 ModApte split, 20 Newsgroups)
Preprocessing
(Tokenization, Stop-word Removal, Lowercase Conversion,Stemming)
Feature Extraction
(Term Weighting, Term Representation)
Feature Selection
(Distinguishing Feature Selection, Information Gain, Gini Index)
Classification
(Bayes Net, C4.5 Decision Tree, Multinomial Naive Bayes, Random Forest, SVM)
Feature Transformation
(PCA, LDA, LSI)

Figure 1.1 *Structure of text classification concept*

Preprocessing step, one of the most important stage in text classification concept, aims to prepare text collection for text classification process. Tokenization, stop-word removal and stemming are widely used preprocessing methods. Application of these methods differ according to the language the documents written. As an example, Turkish and English have different stop-word lists because of the fact that they have different vocabularies. Parameters or types of stemming algorithms may differ for various languages. While Porter (Porter 1980) stemming algorithm is known as a common solution for English language, Zemberek (Akın and Akın 2007) is an example to stemmers for Turkish language.

Feature extraction step aims to extract numerical information from raw text documents in the collections. At the end of this step, each text document is represented as a vector. Bag of words (BoW) technique and vector space model are used to realize this process. In the bag-of-words technique (Forman 2003), each distinct term in a document collection is regarded as an individual feature. Then, the order of terms within the document is ignored but frequencies of the terms are considered in order to represent documents using vector space model. Instead of directly weighting terms using their corresponding frequencies, some other term weighting methodologies may be utilized in this step.

Dimension reduction, which is an important step, objects to achieve greater efficiency by representing data in a lower dimensional feature space. Although there are two common ways in pattern recognition for dimension reduction, feature selection is the mostly preferred against feature transformation. Feature selection process aims to reduce dimension via selecting a subset from the original feature set. This process may decrease computational time and increase classification accuracy.

Classification step, aims to make classification using more appropriate algorithm. Although the classification algorithms are common for all pattern recognition tasks, the most efficient one may change due to the characteristics of the problem. Some of the classifiers used for text classification are Support Vector Machines, Naïve Bayes, and C4.5 decision trees.

1.1. Problems in Medical Text Classification

As mentioned in the introduction section, medical text classification is one of the application fields in text classification. Therefore, most of the problems are common or similar with the other text classification problems. The most important problem is the lack of public datasets and difficulty in construction of new datasets. Dataset construction is difficult because contribution of human experts is necessary for annotation. The most common way is to use medical paper abstracts in order to construct a dataset consisting of medical documents. Another problem about medical text classification is to detect efficient parameters of classification framework which is shown in Fig 1.1.

For information retrieval and classification studies on medical texts, electronic documents in MEDLINE database are frequently used. MEDLINE is a bibliographic database containing documents over 21 million belonging to approximately 5600 medical journals. MEDLINE database consist of medical article abstracts in English and medical subject headings (MeSH) as category of the article. This database can be queried through Pubmed search platform with certain parameters via the internet. In order to search articles, there are some available parameters such as publication language and medical topics. Besides, there are few resources on the internet for querying medical documents in other languages such as Turkish. ULAKBIM Medical Database in Turkey is an example to these types of resources which is created to facilitate access to information for experts working in the medical field. Both MEDLINE and ULAKBIM Medical Database are indexed through manually selecting MeSH category information by experts. Although automatic indexing policy is not available for MEDLINE database currently, there is some automatic classification studies conducted on MEDLINE documents in the literature. There exist some datasets constructed for medical document classification purposes consisting of MEDLINE documents especially in English. Benchmark datasets which are constructed using MEDLINE documents is utilized in most of the studies in the literature. The most important example of these types of datasets is Ohsumed which contains English medical text abstracts associated with 23 diseases. In contrast, there is no dataset available which can be used for research purposes containing medical documents especially in Turkish.

In the literature, there are a number of studies to improve the performance of classification accuracy. In a study(Yetisgen-Yildiz and Pratt 2005), classification was performed, using words, medical phrases and both of them was investigated the effect of classification accuracy. In the other study(Yi and Beheshti 2008), classification is performed with Hidden Markov models. Also, the impact of different representation of biomedical texts on the performance of classification are analyzed(Yepes, Plaza et al. 2015). There exist many researchers working on these issues yet.

1.2. Our Contributions

In this thesis, various solutions are proposed to the problems mentioned in the previous subsection. The contributions are specifically the answers to the research questions below:

i. *What is the effect of some parameters such as preprocessing and feature selection on classification of medical documents especially in English?*

ii. *How do we construct a new annotated dataset including medical documents originated from Turkish resources?*

iii. *Which combination of text classification framework members will provide the most efficient performance for classification of Turkish originated medical documents?*

First of all, performances of the different classifier on English medical abstract from Ohsumed dataset are compared. The effect of stemming, as a text preprocessing step, is analyzed. Experiments are realized for two cases whether stemming is applied and not applied. Also, the impact of feature selection on classification of English medical abstract was investigated using two datasets. The performances of various feature selection methods namely GI, DFS and IG were analyzed using different classifiers. OHSUMED and self-constructed dataset were used for evaluation of feature selection methods.

Considering the second research question, two new datasets including Turkish and English medical abstracts of the same documents is constructed, respectively. In order to skip annotation requirements, English MEDLINE documents which are originated from Turkish resources are automatically retrieved with their corresponding labels via some queries on Pubmed search platform. As in Ohsumed dataset, 23 MeSH disease categories are used as labels for these purposes. After the construction of the first dataset which contains English medical documents, Turkish versions of these annotated medical abstracts are manually collected from the internet and a new Turkish dataset is constructed.

The last contribution is the analysis of better combinations providing the best performance on these two new datasets containing Turkish originated medical documents. Experiments were carried out on two self-constructed datasets and Ohsumed dataset. Experimental results show that the most successful one is BN classifier which obtained 0.68 F-Score the highest success rates in case of not applying the stemming preprocessing on Ohsumed dataset in the first experiment. It is followed by DT and RF classifiers. The second experimental results show that in most cases, DFS is superior to GI. In a small part of experiments, DFS and GI give similar results on both of the two datasets. BN classifier is more successful than DT classifier in most of the cases. As a result, the most successful setting is the combination of Bayesian Network classifier and Distinguishing Feature Selector. The third experimental results show that in most cases DFS is superior to IG and GI. It has been done with more successful classification on English document. It is observed that 0,79 F-Score the highest success rates in case of applying stemming and MNB classifier but generally the situation which not applying stemming is more successful in Turkish dataset. It is observed that 0,86 F-Score the highest success rates in case of applying stemming and the combination of DFS and MNB and also the situation which applying stemming is more successful in English dataset.

1.3. Organization of the Thesis

The next chapters of the thesis are organized as follows: Studies that conducted with Turkish and English text documents in the literature is described in Chapter 2. Text classification phases which is used in the thesis are described in Chapter 3. In Chapter 4, the characteristics of three datasets utilized within the scope of the thesis namely Ohsumed, self-constructed Turkish dataset and self-constructed English dataset are explained. Experimental work is presented in Chapter 5. In the last chapter, potential future works are discussed besides some concluding remarks about experimental results.

2. BACKGROUND AND RELATED WORKS

With the development of internet technology, a substantial increase was seen in the number of electronic documents. With this increase, automatic text classification has gained quite importance. The main task of automatic text classification approach is to assign the texts appropriate class according to contents (Uysal and Gunal 2012). Examining the previous studies in the literature, though there are many studies in text classification in other languages except Turkish, the number of text classification studies conducted with documents in Turkish language is very limited. The reason of this situation is that there are very few Turkish datasets.

There are many text classification studies in English. Under the topic of medical text classification the first that comes to mind is the classification of medical abstracts. Conducted on medical texts information retrieval and automatic text classification, it is generally used documents in MEDLINE database(MEDLINE). MEDLINE is a bibliographic database including documents over 21 million belonging to approximately 5600 medical journals. MEDLINE database contains medical article abstracts in English and medical subject headings (MeSH) as category of the article. This data can be retrieved thanks to search platform which is called PubMed(Pubmed) via the internet. MEDLINE are indexed by experts as the category information of related MeSH terms through selecting manual procedure. Although a system which is automated indexed in the MEDLINE database is not use, automatic text classification study conducted on MEDLINE data are available in the literature (Yetisgen-Yildiz and Pratt 2005, Camous, Blott et al. 2007, Rak, Kurgan et al. 2007, Spat, Cadonna et al. 2007, Poulter, Rubin et al. 2008, Yi and Beheshti 2008, Dollah and Aono 2011, Frunza, Inkpen et al. 2011, Fournier 2013, Uysal and Gunal 2014). In these studies, datasets which containing a particular portion MEDLINE documents and a smaller number of documents. The most important example in this issue is Ohsumed dataset that including English medical text summaries about 23 pieces disease. Ohsumed is multilabel due to the structure of the MEDLINE database and studies on all of the data is performed by multi-label classification approaches. In one of these studies, while classification was performed, using words, medical phrases and both of them was investigated the effect of classification accuracy, and the results showed that a small margin gives better results in case of two are used together(Yetisgen-Yildiz and Pratt 2005). In another study, multi-label classifiers based on associative classifier were examined classification performance on the medical articles(Rak, Kurgan et al. 2007). In the other study, classification is performed with hidden Markov models(Yi and Beheshti 2008). As well as, a number of studies to which the ontology-based classification approach is also available in the literature(Camous, Blott et al. 2007, Dollah and Aono 2011). In another study, an approach that the support vector machines and latent semantic indexing are used in combination were applied to datasets which among the medical texts(Uysal and Gunal 2014).

In the literature, there are a limited number of studies on the classification of Turkish medical texts(Ceylan, Alpkoçak et al. 2012, Arifoğlu, Deniz et al. 2014). However, there was not found studies which oriented classification of Turkish academic medical texts. The main reason is the lack of datasets including Turkish medical documents. Turkish academic texts in TUBITAK's National Medical Database are also available, it hasn't been seen a study using this data in the literature. In addition to the work done by MEDLINE documents, there is also medical text classification studies which is conducted a result of obtaining data from various clinical(Spat, Cadonna et al. 2007, Ceylan, Alpkoçak et al. 2012, Arifoğlu, Deniz et al. 2014). Some of these studies have been conducted classification studies in different languages such as German(Spat, Cadonna et al. 2007). In addition, Also, the impact of different representations of biomedical texts on the performance of classification are analyzed (Yepes, Plaza et al. 2015).

In a study (Kılınç, Özçift et al. 2015), a new Turkish dataset namely TTC-3600, which can be widely used in the studies of Text Classification about Turkish news and articles, is created. The other study(Uysal, Gunal et al. 2012), a new publicly available Turkish SMS message collection is constituted. In a study(Torunoğlu, Çakırman et al. 2011), the datasets Milliyet_9c_1k and Hürriyet_6c_1k are collected by using web crawler that they developed. Milliyet_9c_1k dataset includes text from the columns of Turkish newspaper Milliyet from years, 2002 to 2011. Hürriyet_6c_1k dataset includes news from 2010 to 2011 on Turkish newspaper Hürriyet. In a study(Toraman, Can et al. 2011), they constructed two different datasets called BilCat-MIL and BilCat-TRT by exploiting Bilkent News Portal. Finally, although Turkish public datasets are available in these domains, there is no available dataset in medical document.

3. MEDICAL DOCUMENT CLASSIFICATION STAGES

Although medical document classification frameworks have common structure in general, most of the stages differ in details according to the language the documents written. This study deals with classification of medical documents in Turkish and English. Before explaining the stages of medical document classification, characteristics of Turkish and English languages are briefly stated below. After this stage, it is mentioned preprocessing steps, feature extraction, feature selection, classification algorithms and performance metrics.

3.1. Overview of the Turkish Language Structure

Turkish is a member of the Oghuz group of languages, a subgroup of the Turkish language family. Turkish is native language of over 79 million people and belongs to the Altaic branch of the Ural-Altaic family of languages. The distinguishing characteristics of Turkish, such as vowel harmony, extensive agglutination and lack of grammatical gender, are universal within the Turkish family and the Altaic languages. For instance, in English "We will not get up" sentence is a single word in Turkish: "get up" is the stem, and the others "will", "not" and "we" are all suffixed to it: "kalkmayacağız". The Turkish is derived from Latin characters that has 29 letters consisting 8 vowels (a, e, ı, i, o, ö, u, ü) and 21 consonant (b, c, ç, d, f, g, ğ, h, j, k, l, m, n, p, r, s, ş, t, v, y, z) and 7 of them are modified from their original version in Latin alphabet (ç, ı, ş, ö, ü, ğ, İ). These seven characters are specific to Turkish alphabet and English alphabet doesn't include these characters. On the other hand, q, w, x characters are specific to English alphabet and Turkish language doesn't include these characters.

3.2. Overview of the English Language Structure

Among the different languages of the world, English is the most widely spoken and written languages of the world. English is a West Germanic language that was first spoken in early medieval England and is now a global lingua franca. English is either the official language or an official language in almost 60 sovereign states. Also, it is utilized by the largest number of the people of many nations in all the five continents in the world. Since the ninth century, English has been written in a Latin alphabet which also called Roman alphabet. The great majority of literary works in Old English that survive to today are written in the Roman alphabet. The modern English alphabet consists of 26 letters of the Latin script: a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z which also have capital forms: A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y, Z. Readers of English can generally rely on the correspondence between spelling and pronunciation to be fairly regular for letters or digraphs used to spell consonant sounds. The letters are b, d, f, h, j, k, l, m, n, p, r, s, t, v, w, y, z, respectively. For the vowel sounds of the English language, however, correspondences between spelling and pronunciation are more irregular. There are many more vowel phonemes in English than there are vowel letters a, e, i, o, u, w, y. The main characteristics for this stride of English language are: 1) Receptiveness, 2) Heterogeneousness, 3) Simplicity of Inflexion, 4) Fixed Word Order, 5) Use of Periphrasis, 6) Growth of Intonation.

3.3. Preprocessing

Generally, the preprocessing stage consists of 4 parts in text classification. These are tokenization, stop-word removal, lowercase conversion, and stemming. The preprocessing stage starts with tokenization step. In this step, a text document is converted into small parts known as words or terms. Afterwards, certain characters such as non-alphabetical ones are removed. The next step is lowercase conversion which is converted all of the tokens into lowercase. After this step, there are two steps performed: stopword removal and stemming the words. Each step is expressed in the following subsections.

3.3.1. Tokenization

Tokenization is the task to split a sentence into words, phrases or other meaningful parts which are expressed as tokens. Words or phrases are frequently separated from each other by blanks which are whitespace, tabs returns, semicolon, commas, and quotes as delimiters (Williams 2003). Typically, tokenization occurs at the word level. Firstly, a simple java tokenizer to tokenize the strings with delimiter set such as “\r”, “\n”, “\t” is applied and afterwards the punctuation list “ .,;:’’’()?!&-#0123456789+/<>$^%[]/=’’ ” is utilized to remove irrelevant tokens.

Tokenization may vary according to languages(Manning, Raghavan et al. 2008). As removing non-ASCII characters may be sufficient to tokenize text documents in English language, in Turkish language it may not be sufficient for text documents. Character sets of Turkish and English language are not same and some of the Turkish characters can not represent with ASCII characters. Table 3.1 shows an instance to tokenization of a Turkish and English sentence.

Table 3.1. *Sample tokenization*

Language	**Sentence**	**Tokens**
Turkish	Haftaya memlekete gidiyorum.	Haftaya, memlekete, gidiyorum
English	I will go to Malatya.	I, will, go, to, Malatya

3.3.2. Stop-word removal

Words that pronouns, conjuctions, adjectives, adverbs and prepositions are called stop-words. Stop-words are not related to the concept of the text and removed prior to the classification. It is important step to removing stop-words removal because of the increasing system accuracy in text classification. This process contains removing certain common words ‘a’, ‘able’, ‘about’,’above’, etc. in English language. In Turkish language, these words are ‘a’, ‘acaba’, ‘ama’, ‘ancak’, etc. Stop-words are specific according to the language being studied as in the event of stemming. Sample stop-words are showed in Table 3.2 in Turkish and English language.

Table 3.2. *Sample stopword list*

Language	**Stop-words**
Turkish	acaba, ama, bana, bazen, çok, çünkü, diğer, elbette, fakat, hangi
English	above, again, best, better, can, currently, definitely, every, has

3.3.3. Lowercase conversion

Lowercase conversion is one of the important preprocessing step in text classification. If we consider separately uppercase and lowercase forms of the words, we have used as different features for the same word. So, all uppercase characters are converted to their lowercase forms prior to stemming step. Lowercase conversion decreases the total number of features. Lowercase conversion can vary in some cases related with characteristic of Turkish and English language. An example of lowercase conversion of the same characters is represented for Turkish and English language in Table 3.3.

Table 3.3. *Lowercased forms of some characters*

Language	Original Form	Lowercased Form
Turkish	U, I	u, ı
English	U, I	u, i

3.3.4. Stemming

Stemming is implemented to obtain the stem of a word or term that is morphological root by removing the suffixes that present grammatical or lexical information about the word. Because of the fact that Turkish is an agglutinative language and hundreds or thousands of different words can be derived from a root word, stemming is a significant step before performing text classification. Stemming algorithms are varied according to the language being studied. In our studies, we utilized fixed prefix stemming (FPS)(Can, Kocberber et al. 2008) and a directory based Turkish stemmer called Zemberek(Akın and Akın 2007). On the other hand, the Porter stemming algorithm(Porter 1980) is widely used by researchers for English language. FPS is a pseudo stemming method which recognizes the first “n” character in the text documents. However, Zemberek is a general-purpose open source NLP toolkit and it contains a suffix dictionary created for stemming. An instance to stemming is represented in Table 3.4 for both Turkish and English languages, respectively.

Table 3.4. *Stemmed forms of some words*

Language	Original form	Stemmed form
Turkish	gözlemlemek	gözlem
English	investigating	investigate

3.3.5. Pruning

Because of the fact that the number of features are very large, pruning method is utilized. Pruning aims to discard terms which appearing rarely or too often. This terms do not contribute identify the topic of the document. In the first study, the pruning which is a simple method was applied for the purpose of dimension reduction. The performing style of the pruning method is as follows: Just passing term of more than ‘n’ times in documents used as an feature, the others were discarded.

3.4. Feature Extraction

The type of representation is known as the “bag-of-words”. Words work well as representation units for classifying documents in information retrieval research domain (Lewis 1992). Each different word corresponds to a feature with a weight that is correlated to the number of times the word occurs in the document in the bag-of-words representation approach. As a result, a document is represented by a multi-dimensional feature vector, for instance vector space model (Salton, Wong et al. 1975).

3.4.1. TF

In text classification algorithms, the documents are represented as vectors. Term frequency of each word in a document is a weight that depends on the distribution of each word and expresses the importance of the word in the document(Diao and Diao 2000).

We suppose that K documents in the text corpus, k represent individual document, $f_{k,d}$ or TF is the number of times “k” term appears in a document “d”, t_k represent term k and occurs k_i of text documents.

$$\text{TF(k,d)} = \begin{cases} number\ of\ occurence\ of\ term\ k\ in\ document\ d, & \text{If k occurs in document d} \\ 0, & \text{otherwise state} \end{cases} \tag{3.1}$$

3.4.2. TF-IDF

TF-IDF weighting is a significant step which determines the success or failure of the classification system(Salton and Buckley 1988). TF factor and IDF factor effect the importance of a term in a document. Inverse document frequency of each word or term is a weight that depends on the distribution of each word in the document. TF-IDF technique uses both TF and IDF to determine the weight a term. TF-IDF term weighting technique is widely-used in text classification domain and the other term weighting schemes are variants of this scheme.

Heuristically, TF-IDF method determines how relevant a given terms are in a particular text document. Inverse document frequency formula is as follows:

$$IDF(\, t_k) \;=\; log\frac{K}{k_i} \tag{3.2}$$

TF-IDF formula is as follows:

$$\text{TF-IDF(k,d)} = \text{TF(k,d)} * \text{IDF}(t_k) = \; f_{k,d} * \log\frac{K}{k_i} \tag{3.3}$$

Consequently, TF-IDF is a vector that containing the various terms and terms weights. This formula is used to measure the relevant or significance value of a term in the text document (Zhang, Yoshida et al. 2011).

3.5. Feature Selection

Feature selection is very important step to reduce dimensionality and remove irrelevant features in text classification domain. This step selects a subset from the all feature set according to some certain rules of feature importance(Liu, Kang et al. 2005).

Feature selection techniques generally are divided into three categories: filters, wrappers and embedded methods. Filter methods are based specific characteristics of the training instances for selecting features without applying any learning algorithm. Filters methods are fast in terms of computation, but they generally do not take feature dependencies into account. Wrapper methods try to find features better appropriated to a predefined classifier. They are expensive according to the filters in terms of computation. Embedded methods combine feature selection to classifier training phase. So, these methods are specific to the used learning model. However, these methods are computationally less intensive than the wrappers(Saeys, Inza et al. 2007).

3.5.1. Distinguishing Feature Selector(DFS)

In text classification domain, an each different terms correspond to a feature. Ranking of terms can be implemented considering the following requirements:

1. If a feature often occurs in a single class and does not occur in the other class, it is distinctive; so it must be assigned a high score.
2. If a feature seldom occurs in a single class and does not occur in the other classes, it is indistinctive; so it must be assigned a low score.
3. If a feature often occurs in all classes, it is indistinctive; so it must be assigned a low score.
4. If a feature often occurs in some of the classes, it is partially distinctive; so it must be assigned a relatively high score.

DFS is one of the recent successful feature selection methods for text classification(Uysal and Gunal 2012) whose aim is to select distinctive features while eliminating uninformative ones considering some predetermined criteria. DFS can be expressed with the following formula:

$$DFS(t) = \sum_{i=1}^{M} \frac{P(C_i|t)}{P(\bar{t}|C_i) + P(t|\bar{C}_i) + 1} \quad \textbf{(3.4)}$$

where M is the total number of classes, $P(C_i|t)$ is the conditional probability of class C_i given presence of term t, $P(\bar{t}|C_i)$ is the conditional probability of absence of term t given class C_i, and $P(t|\bar{C_i})$ is the conditional probability of term t given all the classes except C_i. DFS scores of the features are between 0.5 and 1.0 according to their significance. The most distinctive features have a significance score that is about to 1.0 while the least distinctive features a low importance score that is about to 0.5.

3.5.2. Gini Index(GI)

GI is an improved version of the method originally used to find the best split of features in decision trees(Shang, Huang et al. 2007). It is an accurate and fast method. Its formula as below:

$$GI(t) = \sum_{i=1}^{M} P(t|C_i)^2 \cdot P(C_i|t)^2 \qquad \textbf{(3.5)}$$

where $P(t|C_i)$ is the probability of term t given presence of class C_i, $P(C_i|t)$ is the probability of class C_i given presence of term t, respectively.

3.5.3. Information Gain

IG is one of the popular feature selection methods which employed as a term significance criterion in the text document(Yang and Pedersen 1997). This approach is formulated as below:

$$IG(t) = -\sum_{i=1}^{M} P(C_i) logP(C_i) + P(t) \sum_{i=1}^{M} P(C_i|t) log\, P(C_i|t) + P(\bar{t}) \sum_{i=1}^{M} P(C_i|\bar{t}) \cdot \log P(C_i|\bar{t}) \qquad \textbf{(3.6)}$$

where M is the number of classes, P(C_i) is the probability of class C_i, P(t) and P($\bar{t}$) are the probabilities of presence and absence of term t, $P(C_i|t)$ and $P(C_i|\bar{t})$ are the conditional probabilities of class C_i given presence and absence of term t, respectively.

3.6. Classification Algorithms

Generally, text classification is object to classifying uncategorized documents into predefined categories. In terms of machine learning, the aim of text classification is to learn classifiers from labeled documents and complete classification on unlabeled text documents. Some of the commonly used classifiers in text classification domain are MultiNomial Naïve Bayes(MNB), Random Forest(RF), Bayesian Network(BN) And Decision Tree(DT) classifiers. Five well-known classification methods are proven to be substantially successful in text classification domain(Amasyalı and Diri 2006, Güran, Akyokuş et al. 2009, Torunoğlu, Çakırman et al. 2011, Tufekci and Uzun 2013). The detailed information about these classifiers are explained in the following subsections.

3.6.1. Multinomial Naïve Bayes

Multinomial Naive Bayes is a specialized version of Naïve Bayes that is designed more for text documents. Whereas simple Naïve Bayes would model a document as the presence and absence of particular words, multinomial naive bayes explicitly models the word counts and adjusts the underlying calculations to deal with in.

The set of classes be denoted by C and N repserents the size of our vocabulary. Multinomial Naive Bayes assigns a test document t_i to the class that has the highest probability P(c|t_i), which, using Bayes' rule, is given below:

$$P(c|t_i) = \frac{P(c) \cdot P(t_i|c)}{P(t_i)}, \quad c \in C \tag{3.7}$$

The class prior P(c) can be estimated by dividing the number of documents belonging to class c by the total number of documents. $P(t_i|c)$ is the probability of obtaining a document like t_i in class c and is calculated given below:

$$P(t_i|c) = (\sum_n f_{ni})! \prod_n \frac{P(w_n|c)^{f_{ni}}}{f_{ni}!} , \qquad \textbf{(3.8)}$$

f_{ni} is the count of word n in our test document t_i and P(w_n|c) the probability of Word n given class c. The latter probability is estimated from the training documents as:

$$\mathrm{P}(w_n|\mathrm{c}) = \frac{1 + F_{nc}}{N + \sum_{x=1}^{N} F_{xc}} , \qquad \textbf{(3.9)}$$

F_{xc} is the count of word x in all the training documents belonging to class c, and the Laplace estimator is used to prime each word's count with one to avoid the zero-frequency problem(McCallum and Nigam 1998). The normalization factor P(t_i) in Equation 1 can be computed using

$$\mathrm{P}(t_i) = \sum_{k=1}^{|C|} P(k) \cdot \mathrm{P}(t_i|\mathrm{k}) , \qquad \textbf{(3.10)}$$

Note that the computationally expensive terms $(\sum_n f_{ni}$)! and $\prod_n f_{ni}!$ İn Equation 2 can be deleted without any change in the results, because neither depends on the class c, and Equation 2 can be written as:

$$\mathrm{P}(t_i|\mathrm{c}) = \alpha \prod_n P(w_n|c)^{f_{ni}}, \qquad \textbf{(3.11)}$$

where a is constant that drops out because of the normalization step.

3.6.2. Random Forest

Random Forest which was developed by Leo Breiman and Adele Cutler is an ensemble learning method of decision trees. Initially, subset of features are randomly selected to construct branches of decision trees(Xu, Guo et al. 2012). Afterwards, training data is created to be used to generate each individual tree. Eventually, RF classification model is created by combining all individual trees. All input parameters are passed to each individual tree in the forest for text classification process. Classification label returns from all trees in the forest and the label with highest vote is selected as predicted outcome.

Random Forest is a quietly successful classifier which runs efficiently on large datasets and can solve thousands of input features without any deletion. RF has an effective method for estimating missing data and maintains accuracy when a very large proportion of data is missing. RF includes an experimental method for detecting feature interactions. Because of the fact that Random Forest are biased for the benefit of these features with more levels, RF classifier may not run efficiently in a dataset containing categorical variables with various number of values.

3.6.3. Bayesian Network

Bayesian Network which is belief network is one of the method which are used to denote modeling and state transitions(Witten and Frank 2005). BN is often used for modeling discrete and continuous variables of multinomial data. These networks encrypt the relationships between variables in the modeled data. In BN, the nodes are interconnected by arrows to indicate the direction of engagement with each other.

A Bayesian Network encodes the probability distribution p(x), where x = $(X_1,\ldots,X_d)$ is a vector of variables, and it can be seen as a pair (M, θ). M is a directed acyclic graph(DAG) where the nodes correspond to the variables and the arcs represent the conditional dependencies or independencies among the variables. By X_i, both the variable and the node corresponding to this variable is denoted. M will give the factorization of p(x):

$$\boldsymbol{p(x) = \prod_{i=1}^{d} p(\, x_i |\, \pi_i)}, \qquad \textbf{(3.12)}$$

where $\prod_\mathrm{i}$ is the set of parent variables that X_i has in M and $\boldsymbol{\pi_i}$ its possible instantiations. The number of states $\prod_\mathrm{i}$ will be denoted as $|\prod_\mathrm{i}| = q_i$ and the number of different values X_i as $|X_i| = r_i \cdot \theta = (\theta_{ijk})$ will represent the probability of X_i being in its *k*th state while $\prod_\mathrm{i}$ is in its *j*th instantiation. This factorization of the joint distribution can be used to generate new instances using the conditional probabilities in a modelled dataset.

Informally, an arc between two nodes relates the two nodes so that the value of the variable corresponding to the ending node of the arc depends on the value of the variable corresponding to the starting node. Every probability distribution can be defined by a Bayesian network. As a results, Bayesian Networks are widely used in problems where uncertainty is handled using probabilities.

3.6.4. Decision Trees

The decision tree is a widely-used machine learning classifier to automate the induction of classification trees based on training data(Quinlan 1986). The main purpose of the decision tree algorithms is to split the feature space into unique regions corresponding to the classes. An unknown feature vector is assigned to a class via a sequence of Yes/No decisions along a path of nodes of a decision tree. C4.5 is an algorithm used to generate a decision tree and it is known as one of the successful decision tree classification algorithms.

The aim of splitting feature space is to produce subsets that are more class homogeneous compared to former subsets. Entropy is well-known and used information to define impurity and it can be computed given below:

$$\mathrm{I(t)} = -\sum_{i=1}^{M} P(C_i|t) \cdot \log_2 P(C_i|t), \tag{3.13}$$

$P(C_i|t)$ symbolizes the probability that a vector in the subset X_t, attached to a node t, belong to class C, i = 1,2,…,M. Performing a split, N_{tY} points are sent into “Yes” node (X_{tY}) and N_{tN} into “No” node (X_{tN}). The decrease in node impurity is formulated as follows:

$$\Delta I(t) = \mathrm{I(t)} - \frac{N_{tY}}{N_t} \mathrm{I}(t_{YES}) - \frac{N_{tN}}{N_t} \mathrm{I}(t_{NO}), \tag{3.14}$$

$\mathrm{I}(t_{YES})$ and $\mathrm{I}(t_{NO})$ are the impurities of the t_{YES} and t_{NO} nodes, respectively.

In general, there are two phases that the first phase is tree growing and the second phase the overfitted branches of the tree are removed in decision tree training algorithm(Damerau, Zhang et al. 2004).

J48 classifier which is a Java implementation of the C4.5 algorithm utilizes divide-and-conquer approach for growing the decision tree. J48 is highly successful in the field of text classification and also has bring up the advantages such as having high performance on large datasets and shorter training duration. The main disadvantage of this classifier is that small variation in training data may cause dissimilar decision trees.

3.7. Performance Metrics

The F-measure, precision and recall are generally utilized to evaluate the accuracy of text classification results. These metrics are utilized to evaluate the accuracy of the result of Bayes Net, Multinomial Naïve Bayes, Random Forest and Decision Tree for text classification. Precision is inversely correlated to False Positive (FP) examples. Recall is inversely correlated to False Negative (FN) examples. The F-measure is a harmonic mean of the precision and recall values used in information retrieval (Özgür, Özgür et al. 2005).

True Positives, True Negatives, False Positives and False Negatives are four different prediction outcomes. Accuracy is the well-known performance evaluation metric that is the ratio of the total number of class files which are classified correctly. It is calculated by using equation given below.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}\,, \qquad \textbf{(3.15)}$$

True Positive (TP) is the count of the documents correctly assigned to the related category. FP is the count of the documents incorrectly assigned to the related category. FN is the count of the documents incorrectly rejected from the related category. According to this definition;

$$\text{Precision} = \frac{correct\ assignment\ by\ the\ system}{the\ total\ number\ of\ the\ system\ assignments} = \frac{TP}{TP+FP}, \qquad \textbf{(3.16)}$$

$$\text{Recall} = \frac{correct\ assignment\ by\ the\ system}{the\ total\ number\ of\ correct\ instances} = \frac{TP}{TP+FN}, \qquad \textbf{(3.17)}$$

Widely-known F-measure values are used as a success measures in our experiments. When parameters corresponding to optimization of F-measure are utilized, this model perform with higher precision and lower recall. Documents are classified to the correct category, but some categories will be ignored. F-measure is computed globally without class discrimination. In this situation, very few categories will be ignored, but it will increase the number of false predictions. Computation of F-measure can be formulated as

$$\text{F-measure} = \frac{2 \cdot p \cdot r}{p+r}, \qquad \textbf{(3.18)}$$

The pair of (p,r) corresponds to precision and recall values, respectively.

4. DATASETS

In the experimental studies, we used three different datasets. The first is Ohsumed dataset contains 18,302 documents. The second is a self-constructed English dataset which contains 1235 English documents. The third one is self-constructed Turkish dataset containing 1235 Turkish documents.

4.1. Ohsumed

Ohsumed is a bibliographical document collection, developed by William Hersh and colleagues at the Orgeon Health Scientces University. The Ohsumed test collection which is a subset of MEDLINE database is a set of 348,566 references from MEDLINE, the on-line medical information database, consisting of titles and abstracts from 270 medical journals covering the years 1987 to 1991. The available fields are title, abstract, MeSH indexing terms, author, source, and publication type. The documents were manually indexed using subject categories in the National Library of Medicine. This dataset show in Table 4.1.

Table 4.1. *Ohsumed dataset*

Disease Names	Class Names	Number of Documents
Bacterial Infections and Mycoses	C01	631
Virus Diseases	C02	249
Parasitic Diseases	C03	183
Neoplasms	C04	2513
Musculoskeletal Diseases	C05	505
Digestive System Diseases	C06	837
Stomatognathic Diseases	C07	132
Respiratory Tract Diseases	C08	634
Otorhinolaryngologic Diseases	C09	169
Nervous System Diseases	C10	1328

Table 4.1. (Continued) *Ohsumed dataset*

Disease Names	Class Names	Number of Documents
Eye Diseases	C11	337
Urologic and Male Genital Diseases	C12	842
Female Genital Diseases and Pregnancy Complications	C13	473
Cardiovascular Diseases	C14	2876
Hemic and Lymphatic Diseases	C15	307
Neonatal Diseases and Abnormalities	C16	356
Skin and Connective Tissue Diseases	C17	592
Nutritional and Metabolic Diseases	C18	815
Endocrine Diseases	C19	200
Immunologic Diseases	C20	1060
Disorders of Environmental Origin	C21	1283
Animal Diseases	C22	56
Pathological Conditions, Signs and Symptoms	C23	1924
All of the disease	C1-C23	18302

4.2. Self-Constructed English Dataset

In the literature, there is a dataset that existing frequently used in text classification working namely Ohsumed. Ohsumed dataset contains 23 disease related medical text summaries. For the purpose of the study, it is constructed a dataset like Ohsumed. However, the content in this data set consists of abstracts of the Turkey-based magazine. For this operation, it was necessary to be collected lots of data over the Internet. The aim of this thesis is compiling Turkish and English counterpart of the Turkey-based medical text. If Turkish counterpart of related documents is not found at the end of the second of the thesis, this documents would be removed from the data set. So, we have reduced the English dataset to the counterparts of the Turkish dataset. It is performed with more grouping in the creation of the Turkish dataset, each document has had to be sought from the Internet one by one. Technical details of constructing the English dataset are described below:

PubMed environment has primarily benefited from the creation of the dataset. We have reached English title of the Turkish article, the Turkish title, the English summary making queries from PubMed environment located within the MEDLINE database. As a result of that querying, we have obtained the data that XML format from this platform. While English title corresponds <ArticleTitle>, Turkish title corresponds <VernacularTitle> in this XML file. Then, through utilizing this XML files that using xml-parser method in Netbeans platform, we create a text file associated with each disease. The file names is named with numbers that corresponding to the section <ArticleId IdType="pubmed">. So, we assign the PubMed ID numbers of that disease to the name of each disease document. In this way, we have set all the documents according to their category. However, Ohsumed due to the nature of the MEDLINE database is multi-label and studies on all of the data is performed by multi-label classification approaches. But in our study we have carried out single-label classification approach. Therefore, we eliminated multi-label documents.

Table 4.2. *Self-Constructed English dataset*

Class Number	**Disesase Category**	***Number of Documents***
1	Bacterial Infections and Mycoses	284
2	Virus Diseases	44
3	Parasitic Diseases	116
4	Neoplasms	32
5	Musculoskeletal Diseases	140
6	Digestive System Diseases	28
7	Stomatognathic Diseases	39
8	Respiratory Tract Diseases	90
9	Otorhinolaryngologic Diseases	20
10	Nervous System Diseases	83
11	Eye Diseases	4
12	Urologic and Male Genital Diseases	2
13	Female Genital Diseases and Pregnancy Complications	11

Table 4.3. (Continued) *Self-Constructed English dataset*

Class Number	**Disesase Category**	***Number of Documents***
14	Cardiovascular diseases	231
15	Hemic and Lymphatic Diseases	5
16	Neonatal Diseases and Abnormalities	3
17	Skin and Connective Tissue Diseases	13
18	Nutritional and Metabolic Diseases	8
19	Endocrine Diseases	1
20	Immunologic Diseases	7
21	Disorders of Environmental Origin	0
22	Animal Diseases	1
23	Pathological Conditions, Signs and Symptoms	73
All of the disease	C1-C23	1235

MEDLINE, only because of hosting the English text summaries, Turkish summary compilation of these articles has created another problem. In this section, it was obtained that benefiting from electronic archive of the magazine sharing on the internet. However, some Turkish documents could not be found. Therefore, we had to eliminate English documents corresponding Turkish documents that can not be found. Because aim of our thesis is to construct English and Turkish datasets corresponding to each other, then doing classification working on them. We have eliminated the English documents from the dataset using the Java programming language in the Netbeans platform. And consequently, we have constructed 1235 piece document in both languages. Because it is less document number of certain categories, we have eliminated them. These categories are C6, C9, C11, C12, C13, C15, C16, C17,C18, C19, C20,C21, C22, C23. It is used categories that are C1, C2, C3, C4, C5, C7, C8, C10, C14, C23. So it was used 1160 piece document in both languages.

4.3. Self-Constructed Turkish Dataset

In this of the thesis, we have tried to construct Turkish medical text summaries relating to the 23 piece diseases. We benefited from the English dataset that constructed earlier, when we done this. We have saved this dataset that related disease in both Turkish and English titles according to separated Pubmed id. Then, the Turkish titles recorded investigated via the Internet (electronic journals etc.) are recorded. However, we found that some of the documents in pdf format on the Internet because of the documents were very old. The data in this documents have been converted Word through PDF Transformer Program. Later these data have been saved related documents. Some Turkish documents were not found even via the Internet. Here, Turkish documents have been eliminated the corresponding English documents as we mentioned earlier.

Table 4.3. *Self-Constructed Turkish dataset*

Class Number	**Disesase Category**	**Number of Documents**
1	Bacterial Infections and Mycoses	284
2	Virus Diseases	44
3	Parasitic Diseases	116
4	Neoplasms	32
5	Musculoskeletal Diseases	140
7	Stomatognathic Diseases	39
8	Respiratory Tract Diseases	90
10	Nervous System Diseases	83
14	Cardiovascular diseases	231
15	Hemic and Lymphatic Diseases	5
16	Neonatal Diseases and Abnormalities	3
17	Skin and Connective Tissue Diseases	13
18	Nutritional and Metabolic Diseases	8
19	Endocrine Diseases	1
20	Immunologic Diseases	7
21	Disorders of Environmental Origin	0
22	Animal Diseases	1
23	Pathological Conditions, Signs and Symptoms	73
All of the disease	C1-C23	1235

5. EXPERIMENTAL WORKS

Experimental works carried out within the scope of this thesis are explained in the following three subsections. In the first subsections, the effect of different classifiers and stemming algorithms to the classification performance are analyzed using Ohsumed dataset. In the second part, the impact of feature selection on medical document classification is analyzed using two datasets namely Ohsumed and self-constructed English dataset. In the third part, the effect of stemming on classification of Turkish originated medical documents is analyzed using two datasets. These two datasets contain Turkish and English counterpart of some MEDLINE documents.

5.1. Effect of Stemming on English Medical Document Classification

In this phase of the thesis, classification performance of the different classifier on medical text abstracts are analyzed. Firstly, classification studies were carried out on available Ohsumed dataset. Then, studies were conducted on new constructed dataset. Experiments are realized with two different settings whether stemming is applied and not applied. So, the effect of stemming on classification performance was analyzed. While these studies performed on existing Ohsumed dataset, Bayesian Network, C4.5 Decision Tree and Random Forest classification algorithms were used. Dataset used in the experiments of the first stage, it is a subset of Ohsumed dataset that transformed single label version containing the 10 largest classes. In the second phase, the English dataset that constructed new dataset was used. In this datasets, first experiments were carried out on a subset that including the largest 12 classes. While the experimental results obtained at this stage, It was used %50 for training and %50 for testing of the datasets. The following, the methods used in the experiments described respectively and then the experimental results obtained so far was presented in table form. Experimental results of the study showed classifier performance ratio and the effect of stemming algorithm on these ratio.

Classification of medical text summary is a multi-label text classification problem according to the type of data which can be considered as a single or multi-labeled. As in most text classification studies, it is possible to use bag-of-words approach in feature extraction step. Therefore, each of the different words in a text collection is considered to be a separate attribute. As a result, a text document is represented by a multi-dimensional feature vector that weighted by the frequency of the words in the documents. Common preprocessing steps is “stop-word removal” and stemming process which used during feature extraction from text documents. In our study, because of using a dataset consisting of English medical text summary, stop-words which are specific to the English language are removed.

In the first experimental study, it is used a subset of single label Ohsumed dataset. Ohsumed which has 23 categories is a multi-label dataset. Owing to the fact that this work be done on single label classifier analysis, firstly it was constructed a single label subset of this dataset. This subset is obtained by eliminating the documents that last more than one class. Thus, the same number of class with Ohsumed and a new dataset has been constructed in class where each document. But in this case it was observed that the number of documents belonging to some class is quite small compared to others. In order to have a more balanced distribution in class-based of data, data which belonging 10 classes with the most number of documents was used conducted experiments. The basis of class distribution of data used in the experiments are shown in Table 5.1. Class Number field in this table show the number as specified in the Ohsumed dataset. In performed experiments, half of the number of documents for training and the other half is used for testing.

Table 5.1. *The basis of class distribution of used data*

Class Number	Disease Category	Number of Documents
14	Cardiovascular diseases	2876
4	Neoplasms	2513
23	Pathological Conditions, Signs and Symptoms	1924
10	Nervous System Diseases	1328
21	Disorders of environmental origins	1283
20	Immune system disease	1060
12	Urology and male genital disease	842
6	Digestive system disease	837
18	Related nutritional and metabolic disease	815
8	Respiratory Tract Disease	634
	Total	**14112**

In experiments, stop-word removal is applied in each case, the stemming algorithm implementation was carried out to relevant situation. In this study, feature selection methods were not applied, but pruning method which is a simple method was applied in order to dimension reduction. The manner of applying the pruning method is as follows: Just terms that last more than 5 document used as a feature and the others were discarded. After this process step, in case of using the stemming algorithm 11712 piece, in case of not using the algorithm 15956 piece feature has emerged. In the next step, features which is obtained was classified using RF, DT and BN. F-scores which commonly used pattern recognition problems are use to measure the success of the classification process. Classification results in cases where the application of stemming preprocessing are shown in Table 5.2.

Table 5.2. *Results in cases stemming applied*

Disease Category	**F-Score RF**	**F-Score DT**	**F-Score BN**
Cardiovascular diseases	0,717	0,801	0,813
Neoplasms	0,695	0,759	0,785
Pathological Conditions, Signs and Symptoms	0,336	0,338	0,343
Nervous System Diseases	0,526	0,553	0,654
Disorders of environmental origins	0,651	0,669	0,655
Immune system disease	0,678	0,685	0,643
Urology and male genital disease	0,404	0,627	0,748
Digestive system disease	0,404	0,606	0,709
Related nutritional and metabolic disease	0,582	0,678	0,629
Respiratory Tract Disease	0,165	0,548	0,602
Weighted F-Score	**0,564**	**0,646**	**0,672**

It is observed that the most successful BN classifier in the case applying stemming preprocess. It is followed by DT and RF classifiers. The class which is classified with the highest success is “Cardiovascular Disease” class for all three classifiers on the basis of the classes. Classes which are classified with the lowest success on the basis of the class is “Respiratory Disease” for RF classifier, for DT and BN classifier is “pathological condition, signs and symptoms” class. The classification results applying of the stemming preprocessing are shown in Table 5.3.

Table 5.3. *Results in cases stemming not applied*

Disease Category	F-Score RF	F-Score DT	F-Score BN
Cardiovascular diseases	0,724	0,783	0,815
Neoplasms	0,707	0,752	0,794
Pathological Conditions, Signs and Symptoms	0,342	0,331	0,357
Nervous System Diseases	0,59	0,503	0,657
Enviromental induced disorders	0,68	0,633	0,678
Immune system disease	0,662	0,716	0,648
Urology and male genital disease	0,52	0,633	0,725
Digestive system disease	0,442	0,593	0,721
Related nutritional and metabolic disease	0,606	0,627	0,655
Respiratory Tract Disease	0,22	0,524	0,62
Weighted F-Score	**0,589**	**0,63**	**0,68**

It is also observed that the most successful BN classifier in case of not applying the stemming preprocessing. It is followed by DT and RF classifiers. The class which classified with the highest success is the “Cardiovascular Disease” in case applying such as stemming preprocessing on the bases of classes for each three classifiers. It is observed to be 0,68 F-Score the highest success rates in case of not applying stemming preprocessing and BN classifier when it is evaluating in the first stage experiments. While stemming preprocessing increased performance for DT, it leads to a decrease in success rates for BN and RF.

5.2. Effect of Feature Selection on English Medical Document Classification

In this phase of the thesis, the impact of feature selection on medical document classification is analyzed using two datasets containing MEDLINE documents. The performances of two different feature selection methods namely Gini Index and Distinguishing Feature Selector are analyzed using two pattern classifiers. These pattern classifiers are Bayesian network and C4.5 decision tree. As this study deals with single-label classification, a subset of documents inside OHSUMED and a self-constructed dataset is used for assessment of feature selection methods. Due to having low amount of documents for some categories in self-compiled dataset, only documents belonging to 10 different disease categories are used in the experiments for both datasets.

The first dataset is a subset of well-known OHSUMED dataset. It consists of medical abstracts collected in 1991 related to 23 cardiovascular disease categories. As this study deals with single-label text classification, the documents belonging to multiple categories are eliminated. Also, only 10 classes are used for classification in order to make the class distribution same with the second dataset. The second one is a self-constructed dataset whose data is retrieved programmatically with querying Pubmed search platform. This dataset differs from the first one. It consists of MEDLINE documents originated from medical journals in Turkey. However, it has smaller amount of data than the first dataset. However, it has same categories with smaller amount of data than the first one. In this dataset, 10 categories having enough number of documents were used for the evaluation. This categories are C1, C2, C3, C4, C5, C7, C8, C10, C14, C23 in Turkish and English datasets. In the experiments, seventy percent of documents in each class was used training and the rest was used for testing.

In experimental studies, we used bag-of-words approach for feature extraction process. In this approach, the order of terms within documents is ignored and their occurence frequencies are used. Therefore each of the unique word in a text collection is considered as a different feature. Consequently, a document is represented by a multi-dimensional feature vector. In a feature vector, each dimension corresponds to a value which is weighted by term-frequency(TF).

It should be also noted that it is necessary to apply some preprocessing steps during feature extraction from text documents. Widely-used preprocessing steps are “stopword removal” and “stemming”. In this study, both of these steps were applied. Porter stemming algorithm was used for stemming and term frequency was used as weighting approach.

In this study, two classifiers in Weka package were used programatically. These are Bayesian Networks and C4.5 Decision Tree classifiers.

Varying numbers of the features, which are selected by each selection method, were fed into DT and BN classifiers. In the experiments, stopword removal and stemming were applied. Widely-known Porter stemmer was carried out as stemming algorithm. In this study, GI and DFS are used as feature selection methods. Dimension reduction was carried out by constructing feature sets consisting of 300, 500, 1000, and

2000 features. Also, F-score was used as success measure. This score is presented as both class specific and weighted averaged. Resulting F-Scores are listed in Table 5.5. and Table 5.6. The best ones in the results are shown as bolded.

Table 5.4. *Results on Ohsumed dataset*

Number of Features			**Options**		
	DFS+DT	**DFS+BN**	**GI+BN**	**GI+DT**	**Classes**
300	0,57	**0,65**	0,63	0,46	C1
	0,62	0,56	0,50	0,55	C2
	0,69	**0,77**	0,76	0,62	C3
	0,83	**0,85**	0,83	0,81	C4
	0,50	**0,58**	0,50	0,42	C5
	0,35	**0,59**	0,58	0,17	C7
	0,59	**0,62**	0,61	0,52	C8
	0,65	**0,67**	0,65	0,57	C10
	0,86	**0,86**	0,84	0,84	C14
	0,45	**0,47**	0,44	0,38	C23
Weighted Average	0,69	**0,71**	0,68	0,64	
500	0,55	**0,67**	0,66	0,51	C1
	0,58	0,52	0,53	0,50	C2
	0,69	0,74	**0,78**	0,70	C3
	0,84	**0,84**	0,82	0,80	C4
	0,46	**0,57**	**0,57**	0,44	C5
	0,24	0,56	**0,57**	0,32	C7
	0,62	**0,62**	0,60	0,48	C8
	0,66	**0,66**	0,65	0,58	C10
	0,85	**0,86**	0,84	0,82	C14
	0,44	**0,45**	**0,45**	0,41	C23
Weighted Average	0,69	**0,70**	0,69	0,64	
1000	0,55	**0,72**	0,68	0,50	C1
	0,58	0,52	0,51	0,50	C2
	0,71	**0,73**	0,70	0,68	C3
	0,83	**0,83**	0,82	0,82	C4
	0,47	**0,58**	**0,58**	0,46	C5
	0,27	**0,55**	0,51	0,24	C7
	0,61	**0,63**	0,62	0,54	C8
	0,63	**0,7**	0,68	0,58	C10
	0,84	**0,86**	0,85	0,81	C14
	0,43	**0,47**	0,46	0,41	C23
Weighted Average	0,68	**0,71**	0,70	0,64	
2000	0,51	**0,72**	**0,72**	0,50	C1
	0,61	0,5	0,5	0,56	C2
	0,67	**0,74**	0,73	0,65	C3
	0,82	**0,84**	0,83	0,81	C4
	0,46	0,57	**0,58**	0,46	C5
	0,14	0,51	**0,52**	0,24	C7
	0,61	0,62	**0,63**	0,53	C8
	0,63	**0,71**	0,7	0,64	C10
	0,84	**0,86**	0,85	0,83	C14
	0,42	**0,47**	**0,47**	0,40	C23
Weighted Average	0,67	**0,72**	0,71	0,65	

Table 4.5. *Results on Self-Constructed dataset*

Number of Features			**Options**		
	DFS+DT	**DFS+BN**	**GI+BN**	**GI+DT**	**Classes**
300	**0,81**	**0,81**	0,79	0,8	C1
	0,67	0,42	0,44	0,57	C2
	0,86	0,72	0,72	0,84	C3
	0,63	0,31	0,31	0,57	C4
	0,62	0,76	**0,77**	0,68	C5
	0,67	**0,7**	**0,7**	0,67	C7
	0,74	0,55	0,59	0,6	C8
	0,27	0,43	**0,39**	0,3	C10
	0,72	**0,88**	0,87	0,69	C14
	0,53	0,49	0,52	**0,56**	C23
Weighted Average	**0,70**	**0,70**	0,69	0,68	
500	0,8	**0,81**	**0,81**	**0,81**	C1
	0,59	0,31	0,31	**0,62**	C2
	0,84	0,77	0,77	**0,88**	C3
	0,63	0,31	0,31	**0,63**	C4
	0,67	**0,77**	**0,77**	0,71	C5
	0,67	**0,75**	**0,75**	0,67	C7
	0,63	0,58	0,57	0,56	C8
	0,37	0,45	**0,46**	0,36	C10
	0,67	**0,89**	0,88	0,7	C14
	0,57	0,53	0,53	**0,59**	C23
Weighted Average	0,68	**0,71**	**0,71**	0,70	
1000	**0,82**	0,81	0,81	0,8	C1
	0,58	0,31	0,31	0,54	C2
	0,83	0,77	0,77	**0,85**	C3
	0,5	0,31	0,31	**0,57**	C4
	0,73	0,77	**0,77**	0,71	C5
	0,67	**0,75**	**0,75**	0,6	C7
	0,67	0,58	0,58	0,65	C8
	0,51	0,45	0,45	**0,51**	C10
	0,7	**0,89**	**0,89**	0,73	C14
	0,59	0,53	0,53	0,56	C23
Weighted Average	**0,71**	**0,71**	**0,71**	**0,71**	
2000	**0,82**	0,81	0,81	0,8	C1
	0,54	0,31	0,31	**0,56**	C2
	0,87	0,77	0,77	**0,93**	C3
	0,67	0,43	0,31	0,63	C4
	0,69	**0,78**	0,77	0,74	C5
	0,63	**0,78**	0,75	0,67	C7
	0,58	0,57	0,58	**0,59**	C8
	0,58	0,46	0,45	0,46	C10
	0,73	**0,89**	**0,89**	0,68	C14
	0,39	**0,53**	**0,53**	0,5	C23
Weighted Average	0,70	**0,71**	**0,71**	0,70	

Considering the highest weighted averaged F-scores, in most cases, DFS is superior to GI. In a small part of experiments, DFS and GI give similar results on both of the two datasets. It should be noted that DFS seems more successful when the feature size is low. Besides, in spite of originated from different sources and having different class-based distributions, the maximum classification performances obtained on these two datasets are similar. BN classifier is more successful than DT classifier in most of the cases.

Considering class based F-scores, classification performances obtained on neoplasms (C4) and cardiovascular diseases (C14) categories are generally higher than the others for the first dataset. This may be due to having high amount of training instances for these two categories. For self-constructed dataset, classification performances obtained on parasitic diseases (C3) and cardiovascular diseases (C14) categories are generally higher than the others. In this case, these are not the classes with maximum number of documents. This situation may be caused by having small amount of data for most of the categories. Also, for most of the class-based F-scores, combination of DFS and BN seems better than the other ones.

Experimental results show that the most successful setting is the combination of Bayesian Network classifier and Distinguishing Feature Selector.

5.3. Analysis of Classification Frameworks on English and Turkish Counterparts

In this phase of the thesis, the effect of stemming which is one of the preprocessing step on medical document classification is analyzed using two datasets in Turkish and English language containing MEDLINE documents. The success measures of three different feature selection methods namely Gini Index(GI), Information Gain(IG) and Distinguishing Feature Selector(DFS) are analyzed using four pattern classifiers. These pattern classifiers are MultiNomial Naïve Bayes(MNB), Bayesian Network(BN), Random Forest(RF) and C4.5 Decision Tree(DT).

As this study deals with single-label classification, a self-constructed Turkish dataset and a self-constructed English dataset is used for assessment of feature selection methods. Due to having low amount of documents for some categories in self-compiled dataset, only documents belonging to 10 different disease categories which are C1, C2, C3, C4, C5, C7, C8, C10, C14 and C23 are used in the experiments for both datasets.

The first dataset is a self-constructed English dataset whose data is retrieved programmatically with querying Pubmed search platform. The documents having multiple categories are removed from this dataset because of concerning single-label classification of medical documents. The second dataset is a self-constructed Turkish dataset whose data retrieved programatically with querying Pubmed platform and Internet(electronic journals etc.). So, we compiled Turkish and English counterpart of the Turkey-based medical texts. In these datasets, 10 categories which are C1, C2, C3, C4, C5, C7, C8, C10, C14 and C23 having enogh number of documents were used the evaluation. The detailed information regarding these datasets is provided in Table 4.2 and Table 4.3. In the experimental studies, seventy percent of medical text documents in each class was used training and the rest was used for testing.

In studies, we used bag-of-words approach as a feature extraction process. While the order of terms within text documents is ignored and their occurrence frequencies are used in this approach. So each of the unique word in a text is seen as a different feature. As a result, each document is represented by a multi-dimensional feature vector. Each of the feature vector, each dimension corresponds to a value which is weighted by term frequency (TF).

We applied some preprocessing steps during feature extraction from text documents. Extensively used preprocessing steps are “tokenization”, “lowercase conversion”, “stopword removal” and “stemming”. In our experiments, all of these steps were applied. Porter stemming algorithm and Zemberek was used for stemming in Turkish and English language, respectively. TF-IDF was used for term weighting.

In this experiment, four classifiers in Weka package were used programmatically. These are Bayesian Network, C4.5 Decision Tree, MultiNomial Naïve Bayes and Random Forest.

Varying numbers of the features, which are selected by each selection method, were fed into DT, MNB, RF and BN classifiers. In the experiments, stopword removal, tokenization, lowercase conversion and stemming were applied. It is compared two cases where stemming is applied and not applied. Widely-known Porter stemmer and Zemberek were carried out as stemming algorithm. In this study, IG, GI and DFS are used as feature selection methods. Dimension reduction was carried out by constructing feature sets consisting of 300, 500, 1000, and 2000 features. Also, F-score was used as success measure. This score is presented as both class specific and weighted averaged.

Resulting F-Scores are listed in Table 5.7.-5.18. The best ones in the results are shown as bolded.

Table 5.6. *Results on Self-Constructed Turkish dataset with not stemming for DFS*

Number of Features	Options				
	DFS+DT	**DFS+BN**	**DFS+RF**	**DFS+NB**	**Classes**
300	0,57	0,65	0,63	0,46	C1
	0,62	0,56	0,50	0,55	C2
	0,69	0,77	0,76	0,62	C3
	0,83	0,85	**0,83**	0,81	C4
	0,50	0,58	0,50	0,42	C5
	0,35	0,59	0,58	0,17	C7
	0,59	0,62	0,61	0,52	C8
	0,65	0,67	0,65	0,57	C10
	0,86	**0,86**	0,84	**0,84**	C14
	0,45	0,47	0,44	0,38	C23
Weighted Average	0,69	**0,71**	0,68	0,64	
500	0,55	0,67	0,66	0,51	C1
	0,58	0,52	0,53	0,50	C2
	0,69	0,74	0,78	0,70	C3
	0,84	0,84	0,82	**0,80**	C4
	0,46	0,57	0,57	0,44	C5
	0,24	0,56	0,57	0,32	C7
	0,62	0,62	0,60	0,48	C8
	0,66	0,66	0,65	0,58	C10
	0,85	**0,86**	**0,84**	0,82	C14
	0,44	0,45	0,45	0,41	C23
Weighted Average	0,69	**0,70**	0,69	0,64	
1000	0,55	0,72	0,68	0,50	C1
	0,58	0,52	0,51	0,50	C2
	0,71	0,73	0,70	0,68	C3
	0,83	0,83	0,82	**0,82**	C4
	0,47	0,58	0,58	0,46	C5
	0,27	0,55	0,51	0,24	C7
	0,61	0,63	0,62	0,54	C8
	0,63	0,70	0,68	0,58	C10
	0,84	**0,86**	**0,85**	0,81	C14
	0,43	0,47	0,46	0,41	C23
Weighted Average	0,68	**0,71**	0,70	0,64	
2000	0,51	0,72	0,72	0,50	C1
	0,61	0,50	0,50	0,56	C2
	0,67	0,74	0,73	0,65	C3
	0,82	0,84	0,83	0,81	C4
	0,46	0,57	0,58	0,46	C5
	0,14	0,51	0,52	0,24	C7
	0,61	0,62	0,63	0,53	C8
	0,63	0,71	0,70	0,64	C10
	0,84	**0,86**	**0,85**	**0,83**	C14
	0,42	0,47	0,47	0,40	C23
Weighted Average	0,67	**0,72**	0,71	0,65	

Table 5.7. *Results on Self-Constructed Turkish dataset with stemming for DFS*

Number of Features	Options				
	DFS+DT	**DFS+BN**	**DFS+RF**	**DFS+NB**	**Classes**
300	0,66	0,69	0,74	0,78	C1
	0,41	0,32	0,35	0,59	C2
	0,73	0,67	0,76	**0,84**	C3
	0	0,18	0,14	0,14	C4
	0,47	**0,81**	**0,81**	0,80	C5
	0,34	0,70	0,57	0,63	C7
	0,53	0,58	0,50	0,56	C8
	0,44	0,50	0,55	0,51	C10
	0,61	0,72	0,75	**0,84**	C14
	0,42	0,63	0,56	0,68	C23
Weighted Average	0,55	0,66	0,69	**0,74**	
500	0,65	0,70	0,75	0,83	C1
	0,35	0,35	0,13	0,76	C2
	0,75	0,66	0,70	0,85	C3
	0	0,18	0,17	0,29	C4
	0,56	**0,79**	**0,83**	**0,86**	C5
	0,40	0,42	0,29	0,63	C7
	0,61	0,52	0,55	0,67	C8
	0,48	0,52	0,39	0,67	C10
	0,49	0,76	0,72	0,83	C14
	0,54	0,58	0,56	0,70	C23
Weighted Average	0,55	0,65	0,67	**0,78**	
1000	0,72	0,75	0,75	0,84	C1
	0,48	0,27	0,25	0,69	C2
	0,76	0,67	0,77	0,80	C3
	0	0,18	0	0,40	C4
	0,59	**0,80**	**0,82**	0,84	C5
	0,26	0,44	0,29	0,63	C7
	0,54	0,55	0,59	0,75	C8
	0,51	0,56	0,56	0,69	C10
	0,60	0,74	0,64	**0,86**	C14
	0,51	0,56	0,59	0,67	C23
Weighted Average	0,61	0.67	0,67	**0,79**	
2000	0,72	0,79	0,69	0,84	C1
	0,50	0,42	0	0,67	C2
	0,76	0,74	0,75	0,78	C3
	0,40	0,18	0	0,50	C4
	0,66	**0,84**	**0,89**	0,83	C5
	0,24	0,43	0,15	0,74	C7
	0,55	0,64	0,53	0,76	C8
	0,52	0,59	0,47	0,71	C10
	0,54	0,70	0,69	**0,87**	C14
	0,48	0,60	0,62	0,65	C23
Weighted Average	0,60	0,70	0,66	**0,79**	

Table 5.8. *Results on Self-Constructed Turkish dataset with not stemming for IG*

Number of Features	**Options**				
	IG+DT	**IG+BN**	**IG+RF**	**IG+NB**	**Classes**
300	**0,69**	0,73	0,71	**0,79**	C1
	0,45	0,23	0,13	0,52	C2
	0,65	0,66	0,65	**0,79**	C3
	0,09	0,31	0	0,14	C4
	0,57	**0,80**	**0,85**	0,76	C5
	0,46	0,67	0,50	0,74	C7
	0,45	0,48	0,46	0,41	C8
	0,39	0,54	0,44	0,43	C10
	0,60	0,70	0,72	0,73	C14
	0,40	0,54	0,59	0,56	C23
Weighted Average	0,55	0,65	0,66	**0,68**	
500	0,69	0,72	0,72	0,78	C1
	0,52	0,33	0,24	0,54	C2
	0,73	0,71	0,67	**0,82**	C3
	0,27	0,31	0	0,59	C4
	0,69	**0,82**	**0,79**	0,78	C5
	0,19	0,60	0,59	0,70	C7
	0,42	0,44	0,39	0,51	C8
	0,32	0,57	0,25	0,39	C10
	0,57	0,75	0,67	**0,82**	C14
	0,51	0,49	0,67	0,60	C23
Weighted Average	0,57	0,66	0,64	**0,72**	
1000	**0,79**	0,76	0,74	0,81	C1
	0,50	0,48	0	0,55	C2
	0,78	0,73	0,81	**0,84**	C3
	0,29	0	0	0,53	C4
	0,72	**0,82**	**0,83**	0,80	C5
	0,43	0,58	0,29	0,70	C7
	0,56	0,52	0,35	0,64	C8
	0,54	0,52	0,38	0,67	C10
	0,60	0,77	0,66	**0,84**	C14
	0,42	0,56	0,52	0,56	C23
Weighted Average	0,65	0,69	0,65	**0,76**	
2000	0,69	0,73	0,71	**0,90**	C1
	0,50	0,46	0	0,71	C2
	0,87	0,74	0,72	0,88	C3
	0,11	0	0	0,62	C4
	0,56	**0,84**	**0,78**	0,81	C5
	0,39	0,48	0	0,74	C7
	0,48	0,55	0,24	0,70	C8
	0,49	0,61	0,33	0,67	C10
	0,58	0,75	0,60	0,88	C14
	0,47	0,47	0,37	0,59	C23
Weighted Average	0,59	0,67	0,60	**0,81**	

Table 5.9. *Results on Self-Constructed Turkish dataset with stemming for IG*

Number of Features	**Options**				
	IG+DT	**IG+BN**	**IG+RF**	**IG+NB**	**Classes**
300	0,61	0,70	0,72	**0,80**	C1
	0,40	0,23	0	0,38	C2
	0,63	0,61	0,67	0,71	C3
	0	0,18	0,15	0,15	C4
	0,54	**0,80**	**0,81**	0,75	C5
	0,45	0,40	0,38	0,47	C7
	0,39	0,55	0,52	0,56	C8
	0,38	0,46	0,31	0,40	C10
	0,46	0,70	0,67	**0,80**	C14
	0,46	0,61	0,61	0,70	C23
Weighted Average	0,50	0,63	0,64	**0,69**	
500	0,66	0,72	0,70	0,80	C1
	0,50	0,33	0,13	0,65	C2
	0,70	0,68	0,69	**0,82**	C3
	0	0,17	0	0,31	C4
	0,47	**0,82**	**0,80**	0,81	C5
	0,45	0,34	0,15	0,67	C7
	0,35	0,53	0,44	0,58	C8
	0,39	0,41	0,26	0,39	C10
	0,52	0,76	0,65	0,80	C14
	0,50	0,56	0,59	0,67	C23
Weighted Average	0,53	0,65	0,63	**0,73**	
1000	0,71	0,74	0,70	0,84	C1
	0,57	0,35	0	0,73	C2
	0,84	0,68	0,74	0,78	C3
	0,09	0,14	0	0,59	C4
	0,60	**0,78**	**0,82**	0,82	C5
	0,22	0,38	0,15	0,74	C7
	0,52	0,56	0,44	0,73	C8
	0,53	0,56	0,31	0,67	C10
	0,57	0,76	0,63	**0,86**	C14
	0,50	0,62	0,56	0,62	C23
Weighted Average	0,60	0,67	0,63	**0,79**	
2000	0,72	0,73	0,70	0,89	C1
	0,50	0,32	0	0,71	C2
	0,81	**0,75**	0,71	0,84	C3
	0,12	0,14	0	0,67	C4
	0,59	0,78	**0,78**	0,83	C5
	0,30	0,34	0	0,74	C7
	0,57	0,70	0,49	0,78	C8
	0,49	0,51	0,21	0,83	C10
	0,55	0,73	0,61	**0,91**	C14
	0,57	0,60	0,52	0,60	C23
Weighted Average	0,61	0,67	0,61	**0,83**	

Table 5.10. *Results on Self-Constructed Turkish dataset with not stemming for GI*

Number of Features	**Options**				
	GI+DT	**GI+BN**	**GI+RF**	**GI+NB**	**Classes**
300	0,64	0,72	0,70	0,76	C1
	0,50	0,18	0,24	0,62	C2
	0,73	0,75	**0,78**	**0,79**	C3
	0,11	0,15	0	0,15	C4
	0,64	**0,80**	0,72	0,72	C5
	0,32	0,56	0,56	0,67	C7
	0,47	0,43	0,47	0,43	C8
	0,30	0,45	0,42	0,42	C10
	0,65	0,68	0,69	0,72	C14
	0,47	0,51	0,59	0,58	C23
Weighted Average	0,57	0,63	0,64	**0,67**	
500	**0,71**	0,72	0,71	0,77	C1
	0,50	0,26	0,13	0,79	C2
	0,69	0,72	**0,84**	**0,88**	C3
	0,26	0,17	0,18	0,56	C4
	0,71	**0,81**	0,83	0,79	C5
	0,29	0,53	0,67	0,70	C7
	0,53	0,47	0,44	0,54	C8
	0,34	0,52	0,39	0,51	C10
	0,56	0,75	0,67	0,80	C14
	0,44	0,48	0,63	0,62	C23
Weighted Average	0,58	0,65	0,67	**0,74**	
1000	0,69	0,69	0,71	0,82	C1
	0,40	0,38	0,13	0,71	C2
	0,85	0,69	0,78	**0,85**	C3
	0,33	0,17	0,18	0,53	C4
	0,73	**0,80**	**0,86**	0,82	C5
	0,40	0,58	0,59	0,74	C7
	0,46	0,49	0,37	0,64	C8
	0,42	0,59	0,32	0,67	C10
	0,48	0,77	0,66	0,82	C14
	0,56	0,52	0,45	0,67	C23
Weighted Average	0,59	0,66	0,65	**0,78**	
2000	0,78	0,75	0,73	**0,86**	C1
	0,56	0,46	0	0,62	C2
	0,81	0,74	0,77	0,84	C3
	0,40	0	0	0,50	C4
	0,76	**0,81**	**0,83**	0,82	C5
	0,31	0,48	0,15	0,74	C7
	0,53	0,52	0,25	0,70	C8
	0,56	0,61	0,36	0,77	C10
	0,54	0,80	0,64	0,83	C14
	0,48	0,49	0,48	0,62	C23
Weighted Average	0,63	0,68	0,63	**0,79**	

Table 5.11. *Results on Self-Constructed Turkish dataset with stemming for GI*

Number of Features	**Options**				
	GI+DT	**GI+BN**	**GI+RF**	**GI+NB**	**Classes**
300	0,66	0,69	0,71	0,80	C1
	0,40	0,18	0	0,74	C2
	0,78	0,69	0,78	**0,85**	C3
	0	0,18	0,12	0,17	C4
	0,53	**0,81**	**0,80**	0,68	C5
	0,47	0,29	0,42	0,67	C7
	0,47	0,52	0,39	0,53	C8
	0,42	0,43	0,30	0,43	C10
	0,49	0,71	0,73	0,81	C14
	0,47	0,63	0,59	0,63	C23
Weighted Average	0,55	0,63	0,65	**0,72**	
500	0,67	0,71	0,67	0,81	C1
	0,37	0,35	0	0,71	C2
	0,75	0,67	0,76	**0,88**	C3
	0	0,18	0,17	0,15	C4
	0,55	**0,80**	**0,82**	0,82	C5
	0,50	0,40	0,38	0,63	C7
	0,51	0,52	0,41	0,62	C8
	0,47	0,52	0,44	0,62	C10
	0,51	0,75	0,61	0,82	C14
	0,50	0,61	0,55	0,65	C23
Weighted Average	0,56	0,66	0,62	**0,76**	
1000	0,71	0,70	0,73	**0,88**	C1
	0,50	0,33	0	0,69	C2
	0,82	0,65	0,73	0,84	C3
	0	0	0	0,40	C4
	0,65	**0,80**	**0,80**	0,81	C5
	0,47	0,34	0,15	0,67	C7
	0,50	0,56	0,44	0,72	C8
	0,44	0,57	0,32	0,64	C10
	0,53	0,75	0,64	0,82	C14
	0,48	0,56	0,61	0,68	C23
Weighted Average	0,60	0,65	0,64	**0,79**	
2000	0,72	0,75	0,73	0,85	C1
	0,50	0,33	0	0,77	C2
	0,81	0,70	0,73	0,81	C3
	0,45	0,17	0	0,50	C4
	0,62	**0,78**	**0,88**	0,84	C5
	0,35	0,32	0,15	0,67	C7
	0,65	0,65	0,38	0,77	C8
	0,44	0,56	0,27	0,74	C10
	0,48	0,75	0,64	**0,86**	C14
	0,47	0,60	0,58	0,65	C23
Weighted Average	0,60	0,67	0,64	**0,81**	

Table 5.12. *Results on Self-Constructed English dataset with not stemming for DFS*

Number of Features	**Options**				
	DFS+DT	**DFS+BN**	**DFS+RF**	**DFS+NB**	**Classes**
300	0,63	**0,74**	0,72	0,72	C1
	0,50	0,18	0,42	0,56	C2
	0,66	0,72	**0,79**	0,75	C3
	0,76	0,62	0,56	0,62	C4
	0,60	**0,74**	0,74	0,67	C5
	0,53	0,70	0,33	**0,76**	C7
	0,39	0,51	0,42	0,49	C8
	0,26	0,24	0,32	0,33	C10
	0,56	0,61	0,64	0,75	C14
	0,47	0,45	0,50	0,62	C23
Weighted Average	0,56	0,61	0,63	**0,67**	
500	0,75	**0,79**	0,77	0,78	C1
	0,46	0,17	0,35	0,56	C2
	0,83	0,73	0,71	0,78	C3
	0,67	0,62	0,59	0,67	C4
	0,68	0,75	**0,78**	0,72	C5
	0,48	0,67	0,44	0,73	C7
	0,52	0,49	0,42	0,57	C8
	0,32	0,28	0,43	0,35	C10
	0,53	0,69	0,70	**0,80**	C14
	0,45	0,50	0,63	0,62	C23
Weighted Average	0,61	0,63	0,67	**0,72**	
1000	0,79	**0,80**	**0,80**	**0,81**	C1
	0,46	0,17	0,25	0,44	C2
	0,80	0,73	**0,80**	0,79	C3
	0,42	0,53	0,33	0,63	C4
	0,67	0,75	0,78	0,80	C5
	0,55	0,67	0,38	0,76	C7
	0,50	0,48	0,45	0,71	C8
	0,24	0,31	0,39	0,49	C10
	0,60	0,68	0,67	**0,81**	C14
	0,54	0,47	0,56	0,55	C23
Weighted Average	0,63	0,63	0,67	**0,75**	
2000	0,76	**0,79**	0,78	**0,87**	C1
	0,44	0,17	0,25	0,69	C2
	0,81	0,72	0,75	0,81	C3
	0,50	0,46	0	0,78	C4
	0,69	0,69	**0,80**	0,83	C5
	0,56	0,64	0,29	0,63	C7
	0,55	0,44	0,35	0,72	C8
	0,44	0,18	0,32	0,54	C10
	0,56	0,69	0,67	0,84	C14
	0,45	0,43	0,56	0,59	C23
Weighted Average	0,62	0,61	0,66	**0,79**	

Table 5.13. *Results on Self-Constructed English dataset with stemming for DFS*

Number of Features			**Options**		
	DFS+DT	**DFS+BN**	**DFS+RF**	**DFS+NB**	**Classes**
300	0,79	0,78	0,80	0,83	C1
	0,64	0,46	0,13	0,69	C2
	0,84	0,72	0,75	0,86	C3
	0,60	0,47	0	0,59	C4
	0,69	0,79	**0,86**	0,84	C5
	0,67	0,76	0,56	**0,92**	C7
	0,55	0,57	0,67	0,60	C8
	0,27	0,39	0,33	0,47	C10
	0,69	**0,89**	0,73	0,89	C14
	0,45	0,49	0,70	0,55	C23
Weighted Average	0,67	0,70	0,71	**0,78**	
500	0,80	0,78	**0,80**	0,83	C1
	0,59	0,25	0	0,59	C2
	0,84	0,70	0,77	0,87	C3
	0,44	0,37	0,29	0,55	C4
	0,72	0,78	0,77	0,81	C5
	0,67	0,76	0,67	**0,96**	C7
	0,58	0,57	0,68	0,70	C8
	0,37	0,39	0,31	0,67	C10
	0,69	**0,89**	0,67	0,88	C14
	0,54	0,52	0,65	0,62	C23
Weighted Average	0,69	0,69	0,69	**0,79**	
1000	0,82	0,78	0,79	0,87	C1
	0,64	0,25	0	0,63	C2
	0,89	0,70	0,74	0,86	C3
	0,50	0,40	0,18	0,67	C4
	0,62	0,79	**0,84**	0,86	C5
	0,67	0,76	0,59	**0,96**	C7
	0,67	0,56	0,59	0,71	C8
	0,47	0,41	0,26	0,70	C10
	0,68	**0,90**	0,71	0,91	C14
	0,53	0,53	0,62	0,67	C23
Weighted Average	0,71	0,70	0,70	**0,82**	
2000	0,82	0,78	0,77	0,89	C1
	0,54	0,25	0	0,85	C2
	0,84	0,70	**0,81**	0,85	C3
	0,67	0,40	0	0,67	C4
	0,71	0,79	**0,81**	0,85	C5
	0,67	0,76	0,50	1	C7
	0,59	0,56	0,51	0,80	C8
	0,54	0,41	0,21	0,74	C10
	0,69	**0,90**	0,69	**0,92**	C14
	0,38	0,53	0,56	0,68	C23
Weighted Average	0,70	0,70	0,67	**0,86**	

Table 5.14. *Results on Self-Constructed English dataset with not stemming for IG*

Number of Features	**Options**				
	IG+DT	**IG+BN**	**IG+RF**	**IG+NB**	**Classes**
300	0,73	**0,74**	0,71	0,70	C1
	0,45	0,17	0,25	0,47	C2
	0,63	0,69	0,56	0,70	C3
	0,70	0,71	0,31	**0,71**	C4
	0,79	0,72	**0,74**	0,64	C5
	0,48	0,67	0,33	0,67	C7
	0,52	0,50	0,42	0,49	C8
	0,26	0,24	0,32	0,19	C10
	0,64	0,60	0,67	0,73	C14
	0,48	0,49	0,51	0,52	C23
Weighted Average	**0,63**	0,60	0,61	**0,63**	
500	0,78	**0,78**	0,74	0,75	C1
	0,52	0,18	0,25	0,38	C2
	0,79	0,73	0,75	0,74	C3
	0,50	0,62	0,53	0,67	C4
	0,63	0,75	**0,77**	0,72	C5
	0,55	0,67	0,40	0,76	C7
	0,39	0,49	0,38	0,51	C8
	0,23	0,28	0,43	0,41	C10
	0,62	0,69	0,64	**0,77**	C14
	0,53	0,50	0,56	0,55	C23
Weighted Average	0,61	0,63	0,65	**0,68**	
1000	0,78	**0,79**	0,76	**0,85**	C1
	0,46	0,17	0,25	0,52	C2
	0,79	0,74	**0,78**	0,78	C3
	0,50	0,46	0,31	0,63	C4
	0,62	0,77	0,77	0,82	C5
	0,56	0,61	0,50	0,73	C7
	0,53	0,48	0,37	0,64	C8
	0,33	0,14	0,23	0,52	C10
	0,59	0,65	0,69	0,79	C14
	0,35	0,43	0,56	0,51	C23
Weighted Average	0,61	0,61	0,66	**0,75**	
2000	**0,77**	**0,79**	0,75	0,84	C1
	0,46	0,17	0,13	0,57	C2
	0,76	0,74	**0,85**	0,83	C3
	0,45	0,46	0,17	0,78	C4
	0,59	0,77	0,78	0,81	C5
	0,57	0,61	0,40	0,63	C7
	0,48	0,48	0,31	0,73	C8
	0,30	0,17	0,28	0,55	C10
	0,61	0,68	0,70	**0,85**	C14
	0,44	0,43	0,48	0,54	C23
Weighted Average	0,61	0,62	0,66	**0,78**	

Table 5.15. *Results on Self-Constructed English dataset with stemming for IG*

Number of Features	**Options**				
	IG+DT	**IG+BN**	**IG+RF**	**IG+NB**	**Classes**
300	0,77	0,80	0,81	0,81	C1
	0,59	0,27	0,25	0,51	C2
	0,85	0,76	0,70	0,84	C3
	0,57	0,62	0,46	0,55	C4
	0,72	0,74	**0,88**	0,81	C5
	0,67	0,75	0,63	**0,96**	C7
	0,62	0,54	0,62	0,61	C8
	0,34	0,48	0,38	0,55	C10
	0,67	**0,89**	0,76	0,89	C14
	0,50	0,50	0,65	0,58	C23
Weighted Average	0,67	0,71	0,72	**0,76**	
500	0,82	0,79	0,80	0,84	C1
	0,55	0,31	0	0,63	C2
	0,89	0,76	0,74	0,86	C3
	0,47	0,31	0,31	0,67	C4
	0,65	0,77	**0,81**	0,79	C5
	0,63	0,75	0,59	**0,96**	C7
	0,64	0,58	0,59	0,63	C8
	0,37	0,44	0,28	0,67	C10
	0,67	**0,89**	0,73	0,90	C14
	0,56	0,53	0,59	0,64	C23
Weighted Average	0,69	0,71	0,70	**0,79**	
1000	0,83	0,81	0,77	0,88	C1
	0,52	0,31	0	0,63	C2
	0,86	0,77	0,81	0,85	C3
	0,50	0,31	0,31	0,67	C4
	0,68	0,77	**0,82**	0,87	C5
	0,57	0,75	0,59	**0,96**	C7
	0,68	0,58	0,56	0,79	C8
	0,50	0,45	0,38	0,77	C10
	0,72	**0,89**	0,76	0,93	C14
	0,40	0,53	0,63	0,67	C23
Weighted Average	0,71	0,71	0,71	**0,84**	
2000	0,81	0,81	**0,80**	0,89	C1
	0,54	0,31	0	0,70	C2
	0,82	0,77	0,74	**0,92**	C3
	0,63	0,43	0	0,67	C4
	0,73	0,78	0,77	0,85	C5
	0,67	0,78	0,50	**0,92**	C7
	0,62	0,57	0,49	0,75	C8
	0,49	0,46	0,19	0,76	C10
	0,68	**0,89**	0,70	**0,92**	C14
	0,42	0,53	0,61	0,62	C23
Weighted Average	0,70	0,72	0,67	**0,84**	

Table 5.16. *Results on Self-Constructed English dataset with not stemming for GI*

Number of Features	**Options**				
	GI+DT	**GI+BN**	**GI+RF**	**GI+NB**	**Classes**
300	**0,66**	**0,74**	**0,69**	0,68	C1
	0,56	0,17	0,25	0,56	C2
	0,58	0,73	0,68	0,76	C3
	0,64	0,62	0,31	**0,78**	C4
	0,42	0,68	0,64	0,54	C5
	0,56	0,63	0,40	0,70	C7
	0,34	0,47	0,43	0,49	C8
	0,32	0,29	0,27	0,21	C10
	0,63	0,62	0,66	0,75	C14
	0,42	0,43	0,59	0,49	C23
Weighted Average	0,55	0,60	0,60	**0,63**	
500	**0,73**	**0,78**	0,77	0,74	C1
	0,55	0,17	0,25	0,48	C2
	0,67	0,74	**0,78**	**0,77**	C3
	0,59	0,62	0,59	0,67	C4
	0,53	0,74	0,72	0,63	C5
	0,55	0,67	0,40	0,76	C7
	0,38	0,50	0,36	0,45	C8
	0,30	0,27	0,32	0,24	C10
	0,57	0,67	0,65	0,73	C14
	0,56	0,45	0,55	0,60	C23
Weighted Average	0,58	0,62	0,65	**0,66**	
1000	0,76	**0,80**	**0,79**	**0,80**	C1
	0,52	0,17	0,35	0,52	C2
	0,82	0,73	0,79	0,79	C3
	0,45	0,53	0,17	0,67	C4
	0,67	0,75	0,78	**0,80**	C5
	0,56	0,67	0,50	0,76	C7
	0,49	0,48	0,34	0,63	C8
	0,20	0,31	0,19	0,49	C10
	0,62	0,68	0,64	0,79	C14
	0,50	0,47	0,55	0,54	C23
Weighted Average	0,63	0,63	0,65	**0,74**	
2000	**0,78**	**0,79**	0,77	**0,85**	C1
	0,50	0,17	0,25	0,69	C2
	0,78	0,72	0,69	0,83	C3
	0,45	0,46	0,31	0,71	C4
	0,59	0,74	**0,82**	0,82	C5
	0,56	0,64	0,50	0,63	C7
	0,51	0,45	0,48	0,71	C8
	0,46	0,17	0,27	0,54	C10
	0,57	0,66	0,68	0,81	C14
	0,50	0,43	0,56	0,58	C23
Weighted Average	0,62	0,61	0,67	**0,78**	

Table 5.17. *Results on Self-Constructed English dataset with stemming for GI*

Number of Features	Options				
	GI+DT	**GI+BN**	**GI+RF**	**GI+NB**	**Classes**
300	0,77	**0,78**	**0,75**	0,82	C1
	0,56	0,48	0	0,69	C2
	0,86	0,70	0,67	0,85	C3
	0,46	0,53	0	0,56	C4
	0,68	**0,78**	0,74	0,81	C5
	0,67	0,67	0,29	0,87	C7
	0,52	0,46	0,54	0,58	C8
	0,16	0,24	0,24	0,38	C10
	0,64	0,85	0,65	**0,89**	C14
	0,44	0,43	0,67	0,54	C23
Weighted Average	0,64	0,67	0,64	**0,75**	
500	0,79	0,78	0,75	0,84	C1
	0,54	0,37	0	0,67	C2
	0,88	0,68	0,66	0,88	C3
	0,46	0,50	0	0,53	C4
	0,70	0,79	**0,83**	0,82	C5
	0,63	0,76	0,50	**0,91**	C7
	0,63	0,54	0,53	0,66	C8
	0,26	0,38	0,33	0,60	C10
	0,66	**0,86**	0,68	0,88	C14
	0,40	0,46	0,62	0,58	C23
Weighted Average	0,67	0,69	0,67	**0,79**	
1000	**0,80**	0,78	0,70	0,87	C1
	0,50	0,32	0	0,73	C2
	0,86	0,70	0,67	0,85	C3
	0,44	0,43	0	0,56	C4
	0,71	0,81	**0,78**	0,86	C5
	0,67	0,76	0,29	**0,96**	C7
	0,70	0,57	0,44	0,73	C8
	0,52	0,39	0,21	0,70	C10
	0,73	**0,88**	0,65	0,92	C14
	0,55	0,50	0,63	0,67	C23
Weighted Average	0,71	0,70	0,62	**0,83**	
2000	0,82	0,78	**0,74**	**0,92**	C1
	0,58	0,32	0	0,88	C2
	0,87	0,70	0,59	0,86	C3
	0,52	0,43	0	0,59	C4
	0,74	0,81	**0,74**	0,86	C5
	0,67	0,76	0,29	1	C7
	0,56	0,57	0,35	0,74	C8
	0,39	0,39	0,14	0,73	C10
	0,67	**0,88**	0,65	0,91	C14
	0,50	0,50	0,47	0,67	C23
Weighted Average	0,70	0,70	0,61	**0,86**	

In terms of languages, English dataset is more successful than Turkish dataset. The cause of this condition, it could be different stemming algorithm for Turkish and English language. While F-scores (weighted average) is ranged from 0,50 to 0,80 in Turkish dataset, it is ranged from 0,55 to 0,86 in English dataset. The situation which not applying stemming is more successful than applying stemming algorithm in Turkish dataset. On the contrary, the situation which applying stemming is more successful than not applying stemming algorithm in English dataset.

In terms of feature selection, in most cases DFS is superior to IG and GI both of the datasets. Classification on English document with DFS is more successful than classification on Turkish documents. The combination of DFS and MNB classifier is the most successful stemmed English dataset. IG is more successful in small size according to DFS and GI in English dataset but in Turkish dataset DFS is more successful in which feature size is 300, 500, 1000, 2000.

In terms of feature size, classification performance generally increase as increasing the feature size. So, the situation that the number of features are 2000 is the most successful performance in most cases.

In terms of classification algorithms, MNB is more successful than BN, DT and RF. The combination of MNB and GI is more successful than the combination of MNB + IG and MNB + DFS in some cases on Turkish and English dataset.

In terms of applying or not stemming algorithm, in Turkish dataset the situation which is not applying the stemming algorithm is more successful the applying stemming algorithm. On the contrary, in English dataset the cases where stemming is applied is more successful than the others in terms of classification performance.

The third experimental results show that in most cases DFS is superior to IG and GI. It has been done with more successful classification on English document. It is observed that 0,79 F-Score the highest success rates in case of applying stemming and MNB classifier but generally the situation which not applying stemming is more successful in Turkish dataset. DFS is more successful than GI and IG except from MNB classifier. DFS is being well combined with DT, BN and RF when applying or not applying stemming in Turkish dataset. In English dataset, the most successful feature selection method is generally DFS. It is observed that 0,86 F-Score the highest success rates in case of applying stemming and the combination of DFS and MNB and also the situation which applying stemming is more successful in English dataset. DFS is more successful than GI and IG with DT, BN, RF and MNB classifiers. DFS is not more successful in case of the number of features is 300.

6. CONCLUSIONS AND FUTURE WORK

In this thesis, we constructed two different datasets which are containing English and Turkish abstract belonging to Turkish articles in the medical domain.

In the experiments, the result of English dataset is more successful than Turkish dataset. The cause of this situation, it could be different stemming algorithm or the effect of feature selection methods or different classification algorithms for Turkish and English dataset.

A novel feature selection method, namely DFS, was utilized in experiments as well as GI and IG. In Turkish and English dataset experiments, in most cases DFS is superior to IG and GI. In terms of feature size, classification performance generally increase in case of increasing the feature size.

The impact of stemming, which is a kind of preprocessing step, on classification of medical abstracts was investigated. In Turkish dataset, the situation which is not applying the stemming algorithm is more successful the situation which is applying the stemming algorithm. In English dataset, this situation is exactly opposite.

Generally, stemming algorithm improve the classification accuracy, but in here this situation is not the case. The cause of this condition may be foreign terminology in Turkish medical documents. Zemberek was not successful for finding stemming of the medical words.

In terms of the impact of classification algorithms, MNB is more successful than BN, DT and RF. The combination of MNB and GI is more successful than the combination of MNB + IG and MNB + DFS in some cases on Turkish and English dataset.

As future works, we may apply various dimension reduction methods such as LSI, PCA for improving the performance. Also, we will try to applying new feature selection methods for text documents. Additionally, it will be tried new stemming algorithm such as Zemberek because of the decreasing classification performance on medical document records.

REFERENCES

Agarwal, B. and N. Mittal (2016). Machine Learning Approach for Sentiment Analysis. *Prominent Feature Extraction for Sentiment Analysis, Springer***:** 21-45.

Akın, A. A. and M. D. Akın (2007). "Zemberek, an open source NLP framework for Turkic Languages." *Structure* **10**: 1-5.

Al Zamil, M. G. and A. B. Can (2011). "ROLEX-SP: Rules of lexical syntactic patterns for free text categorization." *Knowledge-Based Systems* **24**(1): 58-65.

Alparslan, E., Karahoca, A., and Bahşi, H. (2011). "Classification of confidential documents by using adaptive neurofuzzy inference systems." *Procedia Computer Science* **3**: 1412-1417.

Amasyalı, M. F. and B. Diri (2006). Automatic turkish text categorization in terms of author, genre and gender. *Natural Language Processing and Information Systems*, Springer**:** 221-226.

Arifoğlu, D., Deniz, O., Aleçakır, K., and Yöndem, M. (2014). CodeMagic: Semi-Automatic Assignment of ICD-10-AM Codes to Patient Records. *Information Sciences and Systems*, Springer**:** 259-268.

Bay, Y. and E. Çelebi (2016). Feature Selection for Enhanced Author Identification of Turkish Text. *Information Sciences and Systems*, Springer**:** 371-379.

Belmouhcine, A. and M. Benkhalifa (2016). Implicit Links-Based Techniques to Enrich K-Nearest Neighbors and Naive Bayes Algorithms for Web Page Classification. *Proceedings of the 9th International Conference on Computer Recognition Systems CORES 2015*, Springer.

Camous, F., Blott, S., and Smeaton, A. F. (2007). Ontology-based MEDLINE document classification. *Bioinformatics Research and Development*, Springer**:** 439-452.

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

Ceylan, N. M., Alpkoçak, A., and Esatoğlu, A. E. (2012). "Tıbbi Kayıtlara ICD-10 Hastalık Kodlarının Atanmasına Yardımcı Akıllı Bir Sistem."

Chang, Y. C., Hsieh, Y. L., Chen, C. C., and Hsu, W. L. (2015). "A semantic frame-based intelligent agent for topic detection." *Soft Computing*: 1-11.

Damerau, F. J., Zhang, T., Weiss, S. M., and Indurkhya, N. (2004). "Text categorization for a comprehensive time-dependent benchmark." *Information Processing & Management* **40**(2): 209-221.

Diao, Q. and H. Diao (2000). Three term weighting and classification algorithms in text automatic classification. In *Proceedings Fourth International Conference/Exhibition on High Performance Computing in the Asia-Pacific Region*, IEEE.

Dollah, R. B. and M. Aono (2011). "Ontology based approach for classifying biomedical text abstracts." *Int. J. Data Eng* **2**(1): 1-15.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of machine learning research, 3*(Mar), 1289-1305.

Fournier, S. A. B. E. S. (2013). "Semantic Enrichments in Text Supervised Classification: Application to Medical Domain."

Frunza, O., Inkpen, D., Matwin, S., Klement, W., and O'blenis, P. (2011). "Exploiting the systematic review protocol for classification of medical abstracts." *Artificial intelligence in medicine* **51**(1): 17-25.

Goswami, G., Singh, R., and Vatsa, M. (2016). Automated Spam Detection in Short Text Messages. *Machine Intelligence and Signal Processing*, Springer**:** 85-98.

Günal, S. (2012). "Hybrid feature selection for text classification." *Turkish Journal of Electrical Engineering & Computer Sciences* **20**(Sup. 2): 1296-1311.

Güran, A., Akyokuş, S., Bayazıt, N. G., and Gürbüz, M. Z. (2009). Turkish text categorization using N-gram words. *Proceedings of the International Symposium on Innovations in Intelligent Systems and Applications (INISTA 2009).*

Jindal, R. and S. Taneja (2015). "A Lexical Approach for Text Categorization of Medical Documents." *Procedia Computer Science* **46**: 314-320.

Kaya, Y. and Ö. F. Ertuğrul (2016). "A novel approach for spam email detection based on shifted binary patterns." *Security and Communication Networks*.

Lewis, D. D. (1992). An evaluation of phrasal and clustered representations on a text categorization task. *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, ACM.*

Li, Y., Hsu, D. F., and Chung, S. M. (2009). Combining multiple feature selection methods for text categorization by using rank-score characteristics. *Tools with Artificial Intelligence, 2009. ICTAI'09. 21st International Conference on, IEEE.*

Liu, L., Kang, J., Yu, J., and Wang, Z. (2005). A comparative study on unsupervised feature selection methods for text clustering. N*atural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference on, IEEE.*

Manning, C. D., and Raghavan, P. S. (2008). Introduction to information retrieval, *Cambridge university press Cambridge.*

McCallum, A. and K. Nigam (1998). A comparison of event models for naive bayes text classification. *AAAI-98 workshop on learning for text categorization*, Citeseer.

MEDLINE "http://www.nlm.nih.gov/databases/databases_medline.html." (Retrieved January 2015)

Narayanan, A., Paskov, H., Gong, N. Z., Bethencourt, J., Stefanov, E., Shin, E. C. R., & Song, D. (2012). On the feasibility of internet-scale author identification. Security and Privacy (SP), *2012 IEEE Symposium on, IEEE.*

Onan, A. (2015). "Classifier and feature set ensembles for web page classification." *Journal of Information Science*: 0165551515591724.

Özgür, A., Özgür, L., and Güngör, T. (2005). Text categorization with class-based and corpus-based keyword selection. *Computer and Information Sciences-ISCIS 2005,* Springer**:** 606-615.

Pak, M. Y. and S. Gunal (2016). "Sentiment Classification based on Domain Prediction." *Elektronika ir Elektrotechnika* **22**(2): 96-99.

Parlak, B. and A. K. Uysal (2015). Classification of medical documents according to diseases. *Signal Processing and Communications Applications Conference (SIU), 2015 23th, IEEE.*

Porter, M. F. (1980). "An algorithm for suffix stripping." *Program* **14**(3): 130-137.

Poulter, G. L., Rubin, D. L., Altman, R. B., and Seoighe, C. (2008). "MScanner: a classifier for retrieving Medline citations." *BMC bioinformatics* **9**(1): 108.

Pubmed "http://www.ncbi.nlm.nih.gov/pubmed." (Retrieved January 2015)

Quinlan, J. R. (1986). "Induction of decision trees." *Machine learning* **1**(1): 81-106.

Rak, R., Kurgan, L. A., and Reformat, M. (2007). "Multilabel associative classification categorization of MEDLINE articles into MeSH keywords." *IEEE engineering in medicine and biology magazine* **26**(2): 47.

Saeys, Y., Inza, I., and Larrañaga, P. (2007). "A review of feature selection techniques in bioinformatics." *Bioinformatics* **23**(19): 2507-2517.

Salton, G. and C. Buckley (1988). "Term-weighting approaches in automatic text retrieval." *Information Processing & Management* **24**(5): 513-523.

Salton, G., Wong, A., and Yang, C. S. (1975). "A vector space model for automatic indexing." *Communications of the ACM* **18**(11): 613-620.

Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y., and Wang, Z. (2007). "A novel feature selection algorithm for text categorization." *Expert Systems with Applications* **33**(1): 1-5.

Sharma, R. and G. Kaur (2016). "E-Mail Spam Detection Using SVM and RBF." *International Journal of Modern Education & Computer Science* **8**(4).

Spat, S., Cadonna, B., Rakovac, I., Gutl, C., Leitner, H., Stark, G., and Beck, P. (2007). Multi-label text classification of German language medical documents. *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems, IOS Press*.

Takçı, H. and T. Güngör (2012). "A high performance centroid-based classification approach for language identification." *Pattern Recognition Letters* **33**(16): 2077-2084.

Torunoğlu, D., Çakirman, E., Ganiz, M. C., Akyokuş, S., and Gürbüz, M. Z. (2011). Analysis of preprocessing methods on classification of Turkish texts. I*nnovations in Intelligent Systems and Applications (INISTA), 2011 International Symposium on, IEEE*.

Tufekci, P. and E. Uzun (2013). Author detection by using different term weighting schemes. *Signal Processing and Communications Applications Conference (SIU), 2013 21st, IEEE*.

Uysal, A. K. and S. Gunal (2012a). "A novel probabilistic feature selection method for text classification." *Knowledge-Based Systems* **36**: 226-235.

Uysal, A. K. and S. Gunal (2014a). "The impact of preprocessing on text classification." *Information Processing & Management* **50**(1): 104-112.

Uysal, A. K. and S. Gunal (2014b). "Text classification using genetic algorithm oriented latent semantic features." *Expert Systems with Applications* **41**(13): 5938-5947.

Uysal, A. K., Gunal, S., Ergin, S., and Gunal, E. S. (2012b). A novel framework for SMS spam filtering. *Innovations in Intelligent Systems and Applications (INISTA), 2012 International Symposium on, IEEE*.

Uysal, A. K., Gunal, S., Ergin, S., and Sora Gunal, E. (2012c). "The impact of feature extraction and selection on SMS spam filtering." *Elektronika ir Elektrotechnika* **19**(5): 67-72.

Williams, K. (2003). A framework for text categorization, *Doctoral dissertation*, The University of Sydney.

Witten, I. H. and E. Frank (2005). Data Mining: Practical machine learning tools and techniques, *Morgan Kaufmann*.

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

Yang, Y. and J. O. Pedersen (1997). A comparative study on feature selection in text categorization. *ICML*.

Yepes, A. J. J., Plaza, L., Carrillo-de-Albornoz, J., Mork, J. G., and Aronson, A. R. (2015). "Feature engineering for MEDLINE citation categorization with MeSH." *BMC bioinformatics* **16**(1): 1.

Yetisgen-Yildiz, M. and W. Pratt (2005). The effect of feature representation on MEDLINE document classification. *AMIA*.

Yi, K. and J. Beheshti (2008). "A hidden Markov model-based text classification of medical documents." *Journal of Information Science*.

Yu, B., Xu, Z. B., and Li, C. H. (2008). "Latent semantic analysis for text categorization using neural network." *Knowledge-Based Systems* **21**(8): 900-904.

Zhang, W., Clark, R. A., Wang, Y., and Li, W. (2016) "Unsupervised language identification based on Latent Dirichlet Allocation." *Computer Speech & Language* **39**: 47-66.

Zhang, W., Yoshida, T., and Tang, X. (2011) "A comparative study of TF* IDF, LSI and multi-words for text classification." *Expert Systems with Applications* **38**(3): 2758-2765.