THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

TEXT MINING IN TURKISH RADIOLOGY REPORTS

Master's Thesis

TUĞBERK KOCATEKİN

ISTANBUL, 2013

THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
COMPUTER ENGINEERING

TEXT MINING IN TURKISH RADIOLOGY REPORTS

Master's Thesis

TUĞBERK KOCATEKİN

Supervisor: ASSIST. PROF. DR. DEVRİM ÜNAY

İSTANBUL, 2013

THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
COMPUTER ENGINEERING

Name of the thesis: Text Mining in Turkish Radiology Reports
Name/Last Name of the Student: Tuğberk Kocatekin
Date of the Defense of Thesis: 29.08.2013

The thesis has been approved by the Graduate School of Natural and Applied Sciences.

Assoc. Prof. Tunç BOZBURA
Graduate School Director
Signature

I certify that this thesis meets all the requirements as a thesis for the degree of Master of Arts.

Assist. Prof. Tarkan AYDIN
Program Coordinator
Signature

This is to certify that we have read this thesis and we find it fully adequate in scope, quality and content, as a thesis for the degree of Master of Arts.

Examining Comittee Members	Signature
Thesis Supervisor Assist. Prof., Devrim ÜNAY	----------
Member Assist. Prof. Tevfik AYTEKIN	----------
Member Dr. A. Kamuran KADIPAŞAOĞLU	----------

ACKNOWLEDGEMENTS

First and foremost I would like to thank my family, Belgin Kocatekin and M. Şahin Kocatekin for their constant support and motivation.

I would like to thank my advisor Assist. Prof. Dr. Devrim Ünay for his patience and belief in me, and our study group including Gökhan Gökay, İlkay Öksüz, Leonardo Iheme, Oguz Demir and Volkan Özdemir for their constant suggestions for improvement.

I would also like to thank Maltepe University Medical Faculty for supplying us with Turkish radiology reports.

I would also thank to Melih Arda Yalçıner, who is always there for me in need of help.

Tuğberk Kocatekin

ABSTRACT

TEXT MINING IN TURKISH RADIOLOGY REPORTS

Tuğberk Kocatekin

Computer Engineering

Thesis Supervisor: Assist. Prof. Dr. Devrim Ünay

September 2013, 47 Pages

Text mining and text classification is a popular area of machine learning and information retrieval. Automated categorization and analysis of medical documents may improve work flow, and aid in better diagnosis and therapy planning. There is already some research done on analysis and categorization of radiology reports. However, to the best of our knowledge there is no prior work on anatomical region based classification of Turkish radiology reports. In order to fill this gap, this thesis focuses on dictionary-based classification of Turkish radiology reports into anatomical regions.

The proposed solution is intented to automatize, speed up, and improve the accuracy of the task of classifying these documents, which is manually realized traditionally.

The proposed solution, implemented in Bash environment, consists of header-footer removal, Turkish character elimination, stemming, word frequency analysis, normalization, and scoring steps. Training (n=69) and performance evaluation (n=161) of the system is realized using a total of 230 Turkish radiology reports from 8 different anatomical regions acquired from routine clinical practice. F-score of the system is measured as 98,6%, and it is observed that the proposed system correctly identifies the actual classes of 7 reports that were previously misclassified by the radiology staff.

In order to improve the accuracy of the system one can increase the size of the training set, incorporate natural language processing solutions, or make use of ontologies that encode anatomical/pathological knowledge. In addition to that, the proposed system can be integrated with speech processing solutions to automatically create radiology reports from audio recordings of radiologists. Lastly, the system can be further improved by user feedback.

Keywords: Turkish, text mining, text classification, radiology reports, text categorization, frequency analysis, dictionary, stemming, normalization

ÖZET

TÜRKÇE RADYOLOJİ RAPORLARINDA METİN MADENCİLİĞİ

Tuğberk Kocatekin

Bilgisayar Mühendisliği

Tez Danışmanı: Yrd. Doç. Dr. Devrim Ünay

Eylül 2013, 47 Sayfa

Metin madenciliği ve sınıflandırma, makine öğrenmesi ve bilgi erişimi alanlarında popüler bir konudur. Tıbbi metinlerin otomatik analizi ve sınıflandırılması medikal veri akışında verimliğin artırılması, teşhis ve tedavinin iyileştirilmesi gibi konularda katkı sağlayabilir. Literatürde radyoloji raporlarının analizi ve sınıflandırılması konusunda çalışmalar mevcuttur. Ancak bahsedilen çalışmalar Türkçe raporların anatomik bölgeye göre sınıflanması problemine eğilmemiştir.

Dolayısıyla bu tez, metin madenciliği kullanarak sözlük temelli bir yöntemle Türkçe radyoloji raporlarını anatomik bölgelere göre sınıflandırmayı hedefleyerek literatürdeki eksiği kapatmayı amaç edinmiştir. Önerilen çözüm, radyoloji departmanlarında teknisyenler tarafından elle yapılan bu işin hızlandırılmasını, otomatikleştirilmesini ve doğruluğunun artırılmasını sağlayacaktır.

Raporlardaki alt ve üst bilgilerinin silinmesi, Türkçe karakterlerin elenmesi, kök bulma, kelime frekans analizi, normalizasyon ve skorlama aşamalarından oluşan önerilen yöntem Bash ortamında tasarlanmıştır. Yöntemin geliştirilmesi(n=69) ve başarımının ölçülmesi(n=161) için hastane ortamında rutin olarak hazırlanan 8 farklı anatomik bölgeye ait toplam 230 Türkçe radyoloji raporu kullanılmıştır. Önerilen yöntemin başarımı F-ölçütü kriterine göre %98,6 olarak ölçülmüştür. Ayrıca yöntemin elle sınıflamada hatalı sınıfa atanmış olan 7 adet raporu doğru sınıfladığı gözlenmiştir.

Önerilen yöntemin başarımının artırılması için öğrenme kümesinin büyütülmesi, doğal dil işleme çözümlerinden yararlanılması ve anatomik/patolojik bilgileri kodlayan ontolojilerin kullanılması gibi yollar denenebilir. Buna ek olarak bu yöntem konuşma tanıma çözümleri ile birlikte kullanılarak radyologların ses kayıtlarından raporların otomatik üretilmesi gerçekleştirilebilir. Son olarak, bu sistem kullanıcı geribildirimi yoluyla geliştirilebilir.

Anahtar Kelimeler: Türkce, metin madenciliği, metin sınıflandırma, radyoloji raporları, frekans analizi, sözlük, kök bulma, normalizasyon

CONTENTS

APPENDICES

TABLES

FIGURES

ABBREVIATIONS

PACS	:	Picture Archiving and Communication System
SVM	:	Support Vector Machines
DT	:	Decision Tree
KNN	:	K-Nearest Neighbor
CT	:	Computed Tomography
MRI	:	Magnetic Resonance Imaging
MeSH	:	Medical Subject Headings
NLP	:	Natural Language Processing
HTK	:	Markov Model Toolkit
USG	:	Medical Ultrasonography
TUD	:	Take as Directed
ICD	:	International Classification of Diseases
WHO	:	World Health Organization
DF	:	Document Frequency
TF	:	Term Frequency
IDF	:	Inverse Document Frequency

SYMBOLS

Precision	:	π
Recall	:	ρ
P-Value	:	α
Null hypothesis	:	H_0

1. INTRODUCTION

It can be agreed upon that one of the easiest form of storing information is via text documents. Medical data is no different, and that is the reason for existing systems storing radiology data within their databases with certain attributes such as anatomical regions, modalities etc.

Increasingly large amount of data acquired in the medical field necessitate solutions for efficient storage, which lead to non-stop improvement in computer technology (e.g. distributed storage systems). These improvements affects the practice of radiology in a positive way (Thrall 2005). However, Reiner (2010) suggests that the systems used should not only be storing large amounts of data, but also should be able to provide efficient access. This efficient access can be provided by organizing the data into certain categories.

Ferris (2009) states that radiology reports are a major product of radiology departments, and the task of reporting should be more structured and standardized to become more patient-directed and to facilitate data mining applications and quality control. A radiology report is written after many steps. First, a medical doctor requests acquisition of a medical image. This image is taken, and further examined by radiologists. In many cases, radiologists create audio recordings of their observations, and an additional staff listens and redacts it. This process shows how radiology reports are formed. After this step, if there is any classification done, it is realized manually. Automating this task will be more time-efficient and more precise. Also, it would make acquisition of specific data from classified databases easier rather than un-classified databases. With number of reports daily increasing, this simplicity in acquisition would become more useful.

Aaron et al (2005) defines text classification as a process of determining whether a document or part of a document has particular characteristics of interest. Previously, various studies applied certain classification techniques on radiology reports, however these studies did not intend to classify reports based on anatomical regions.

To this end, the presented study aims to fulfill this need by successfully implementing an algorithm to automatically classify Turkish radiology reports into their respective anatomical regions.

1.1 RESEARCH OVERVIEW

The motivation of this thesis is to classify Turkish radiology reports into their respective anatomical regions by implementing a dictionary-based classification system.

The system is developed and its performance is evaluated using a total of 230 pre-labeled reports acquired from the routine clinical practice. After pre-processing and stemming the data, the system is trained by applying frequency analysis on the reports belonging to each of the eight anatomical regions separately. Frequency analysis results are then investigated to identify distinctive words and creating dictionaries for classification.

The system classifies test documents by searching every term in every dictonary in a given document. The number of repetitions are multiplied by the word's weight to compute a score for that class. After doing this for every class, the classifier assigns the class with the highest score as the class of the document under investigation.

1.2 CONTRIBUTIONS OF THE THESIS

Automated radiology report mining and categorization is a difficult, yet crucial task to increase the efficacy in radiology practice. As the literature review below suggests, automated solutions for classifying Turkish radiology reports are limited, and those limited solutions do not focus on report categorization into anatomical regions. In addition to that, to the best of our knowledge anatomical region based classification of radiology reports in languages other than Turkish is an unexplored area of research as well. Accordingly, *the proposed system is first ever to classify Turkish radiology reports into respective anatomical regions*. It employs a dictionary-based approach, and it is evaluated over Turkish radiology reports acquired from routine clinical practice.

1.3 THESIS OUTLINE

The remainder of this thesis is organized in 5 chapters.

Chapter 2 presents literature review and previous work on related subjects. It also includes a graph in order to better illustrate the gap this thesis attempts to fill.

Chapter 3 describes the data, the method used for report classification, and the implementation environment.

Chapter 4 details evaluation results of the system's performance and the related analyses.

Chapter 5 concludes the thesis and presents suggestions to improve the proposed system.

2. LITERATURE REVIEW

Text classification is defined as determining if a document has characteristics of a pre-defined class. Cohen *et al* (2005) claims that the need for effective and accurate text classification methods are strong with more biomedical information is being created in text form now more than ever. There are several studies on applying text classification methods on biomedical documents. Using NLP (Natural Language Processing) as a classification method is common in radiology reports (Hripcsak et al. (2002), Mamlin et al. (2003), Goldstein *et al.* (2007). Solutions make use of other approaches such as ad-hoc classification, rule-based classification, boolean classification and SVM are also explored. (Friedman et al. (1994), Aronow *et al.* (1999), Thomas et al. (2004), Maghsoodi *et al.* (2012), Lakhani *et al.* (2012))

Friedman *et al.* (1994), developed a general NLP system which aims to identify clinical information from radiology reports and create a formal model to represent this clinical information. The classifier has three phases of processing: 1) Parsing, which identifies the structure of the text by using a grammar that defines semantic patterns, 2) regularization that standardizes the terms by a compositional mapping of multi-word phrases, and 3) encoding, which maps the terms into a vocabulary. They did a preliminary study consisting of randomly selected 230 reports to evaluate the processor. Four diseases which the system is not previously trained are chosen for evaluation: neoplasm, congestive heart failure, acute bacterial pneumonia, and chronic obstructive pulmonary disease. The recall and precision of the system are found to be 70 and 87 percent, respectively.

Aronow *et al*. (1999) used an ad-hoc classification system to classify dictated mammography results. Ad-hoc classification is mainly used when a large number of documents are needed to be sorted in non-standard categories defined by the user. Their dataset consisted of 421 relevant and 256 irrelevant documents and taken from U.S. Naval medical centers. They chose 40 relevant and 40 irrelevant documents for training, and divided the remaining documents into 7 different test collections. They did classification into three classes: positive, uncertain and negative; and measured an F-score varying between 89.7 and 79.1 percent.

Hripscak et al. (2002), proposed a system in order to evaluate translation of chest radiographic reports by using NLP. Their dataset consisted of 889921 chest radiographic reports from 1989 to 1998. They excluded CT (computed tomography) and MR (magnetic resonance) reports in order to create a more homogeneous sample. They used MEDLEE to convert narrative text to a semantic structure. Their study assessed 24 clinical conditions, both common and uncommon conditions included. Because the accuracy of the processor had not been tested, they have enlisted a medical school graduate coder to verify the accuracy of the translation, which is further validated by six radiologists and seven internists. Coder reviewed 150 randomly chosen reports, and classified each condition as present or absent. They compared NLP coded reports with manual coded reports and reported that natural language processing can be as accurate as expert human coders for coding radiographic reports. On the 150 manually coded reports, the system's average sensitivity is found to be 0.81, and average specificity to be 0.95. They concluded that these findings were similar to earlier results for that system, and thus comparable to the results of previous reports of expert human coding.

Goldstein et al. (2007), described and evaluated three systems for predicting ICD-9-CM codes of radiology reports from short excerpts. Their first system uses Lucene, second system uses BoosTexter, and third system uses a set of hand-crafted rules which captures lexical elements from BoosTexter's n-grams.

Their dataset consisted of 978 pre-labeled reports which came from 2007 Computational Medicine Center challenge. They used 880 reports for training their system, and the remaining 98 reports were used for initial testing. They completely tested their system on 976 reports which were released by challenge organizers later on. Their first system, Lucene is measured to have an F-score of 66.9 percent; second system, BoosTexter 80.4 percent, and the rule-based system outperformed other systems by having an F-score of 88.5 percent.

Mamlin et al. (2003) evaluated the extension of a commercially available product to complete encoding of narrative cancer-related x-ray reports. They evaluated LifeCode

which combines NLP and and a medical coding expert system which can extract and normalize demographic and clinical information from free-text clinical reports. They used a dataset of 26778 reports generated in Wishard Memorial Hospital in Indianapolis, Indiana. After applying some filtering onto their dataset, only 3015 reports were left. Their training dataset consisted of randomly chosen 1400 documents. The system was trained by processing a set of reports, manually reviewing and correcting. They tested their system by using 500 randomly selected reports which were manually coded by a board-certified internist. They evaluated the system by comparing human-generated codes with computer-generated codes and linked all matches. Unlinked codes were tagged as false positives and false negatives. The recall and precision values range between 83.6 – 86.6, and 82.6 – 98.6 percent.

Thomas *et al*. (2004) proposed a system in order to create and validate an automated computerized method for categorizing narrative text reports by using Boolean language search strings. Their training dataset consisted of 512 ankle radiolography reports from a single clinical imaging center, and test dataset consisted of 750 spine and extremity reports produced at three different clinical imaging sites. They did classification in three classes: normal, neither normal nor fracture, and fracture. After expert review, their accuracy is 91,6 percent for normal, 87,8 percent for neither normal nor fracture, and 94,1 percent for fracture.

Prasad *et al*. (2010) proposed a text mining system consisting of three main modules: medical finding extractor, report and image retriever, and text-assisted image feature extractor. Extractor module extracts the medical findings of brain CT radiology reports and extracts medical findings and modifiers, and structures them. Retrieval modules analyzes user query and retrieves the matching reports and images. They used a term mapper which maps single and multiple-word terms to a medical lexicon. This lexicon is constructed by the authors from combining MeSH vocabulary, other radiology and anatomy thesaurus and actual CT radiology reports. MeSH, Medical Subject Headings is a large vocabulary developed by National Library of Medicine to be used in medical texts indexing. The parser is trained using labeled brain CT radiology reports.

Maghsoodi *et al*. (2012) created a pipeline for automatic sentence classification of narrative breast cancer radiology reports. Their dataset consisted of 353 reports including 8166 sentences. They did classification in seven classes: left, right, bilateral, mammogram, ultrasound, mri and recommendation. They chose classifiers from different learning paradigms: rule-based decision tree, support vector machines, probabilistic naive bayes, and instance-based k-nearest neighbor. After averaging over seven classes, SVM and DT outperformed the other classifiers with classification accuracy ranging between 92 and 98 percent.

Lakhani *et al*. (2012) developed a text-mining algorithm which automatically identify radiology reports containing critical results. They used a rule-based approach to design the algorithm and searched for common words in radiology reports that indicate critical results. Their initial test collection consisted of approximately 2.3 million diagnostic radiology reports ranging from 1997 to 2005. Their subsequent test collection consisted of approximately 10 million radiology reports ranging from 1988 to 2011. They chose nine results as classes: *acute pulmonary embolism*, *acute cholecystitis*, *acute appendicitis*, *ectopic pregnancy*, *scrotal torsion*, *tension or new/increasing large pneumoThorax*, *unexplained free intraperitoneal air*, *increasing or new intracranial hemorrhage*, and *mal-positioned nasogastric*, *feeding*, and *endotracheal tubes*. They used f-score to measure the accuracy of the system and it varied between 81 and 100 percent.

The number of studies done with Turkish radiology reports are limited. Oğuz *et al*. (2007) presented a survey on applying text mining techniques in medicine.

Arisoy *et al*. (2006) proposed a task-specific, Turkish radiology dictation system for radiology applications in order to ease the process of converting medical doctors' radiology speeches into report. Their system is composed of acoustic model training to generate a database of physical models, and language model to generate a recognition network with the help of HTK (Hidden Markov Model Toolkit). They performed experiments using recordings of 100 words in form of a radiology training corpus taken

from 15 different speakers consisting of 8 males and 7 females. Since the number of recordings are limited they did cross-validation (10 reports used to generate the pronunciation variants, and remaining 5 are used for testing purposes). By using words as recognition units, they achieved a success rate of 87.06 percent.

Soysal *et al*. (2010) described a system that processes free text radiology reports in order to extract the information and convert it into a structured information model. It uses NLP in combination with ontology as domain knowledge to transform verbal descriptions into an information model to be used for computational purposes. Their system only works with abdominal radiology reports, however they argue that this system can be used in other fields of medicine, by adapting the ontology and the rule set. They used a dataset consisting of 756 abdominal USG (Ultrasound-based diagnostic imaging technique) reports. Their system consists of 5 modules: morphological analysis, ontology, term analysis, rule set, and knowledge model. The morphological analysis module, works on morpheme, the smallest unit of grammar. This module uses its own dictionary constructed from radiology reports. Ontology is a dictionary which shows terminology of medical terms and hierarchical localization. Term analysis uses the terminology and context part of the ontology in order to determine the terms. Rule sets are determined with the specialists by reading the reports in training set. The system achieves a precision and recall of 97 and 92 percent respectively.

Ceylan *et al.* (2012) proposed a intelligent system to help the task of assigning ICD-10 codes to medical records. ICD-10 is the 10^{th} revision of the International Statistical Classification of Diseases and Related Health Problems (ICD), which is a medical classification list by the World Health Organization (WHO). System classifies free-form text fields and uses Terrier information retrieval engine to handle unstructured text data. Terrier is an open source framework for research and experimentation in information retrieval. (Ounis 2006). Their data consists of 57835 preprocessed and anonymized medical records, which are taken from Ankara University Hospital Information systems. They do classification for 40 diseases types and measured an accuracy of 76.5 percent.

Table 2.1: Summary of prior art on computer-based radiology report mining

Author / Year	Classification Method	Language	Test Data	Classified Into	Accuracy (%)
Aranow et al., 1999	Adhoc	English	597	3 diagnostic decisions	79,1–89,7
Friedman et al., 1994	NLP	English	230	4 types of diseases	70-80
Maghsoodi et al., 2012	Mixed	English	353	7-arbitrary classes	92-98
Thomas et al., 2004	Boolean	English	750	3 diagnostic decisions	91,17
Lakhani et al., 2012	Rule-based	English	2.3 million	9 types of diseases	81-100
Ceylan et al., 2012	Terrier system	Turkish	57835	40 disease types	76,5
Present Work	Dictionary-based	Turkish	161	8 anatomical regions	98,6

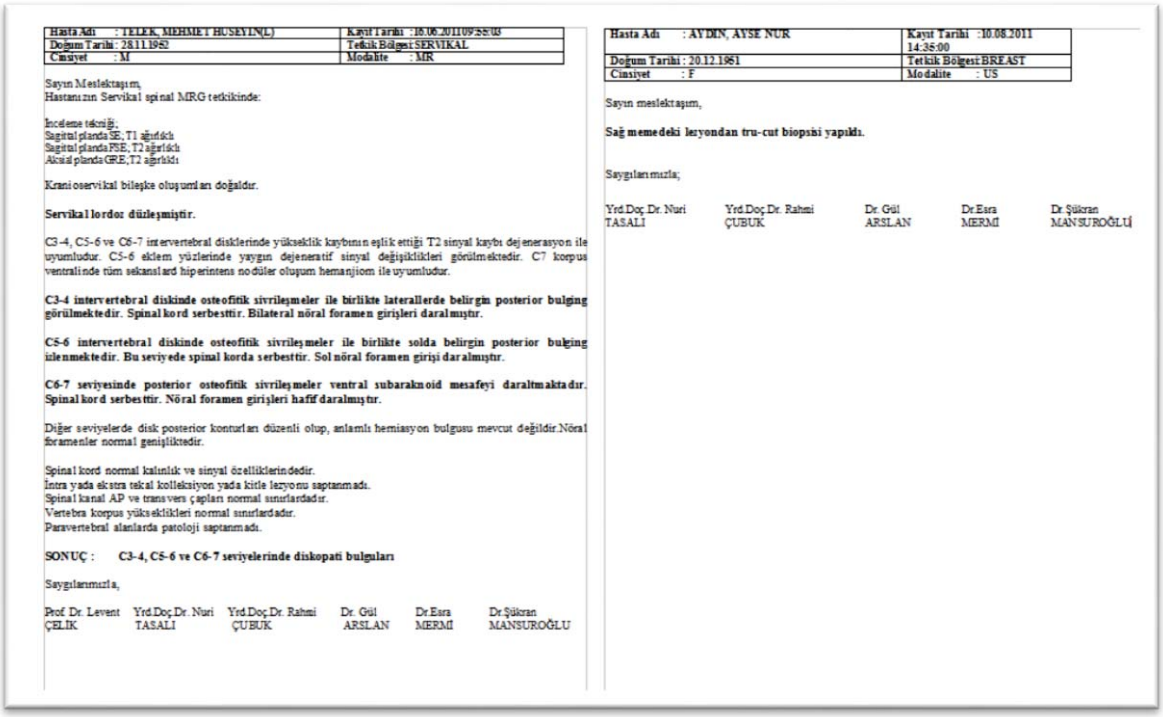
As Table 2.1 shows, computerized solutions for Turkish radiology reports are limited, and none of these works focus on categorizing these reports into their anatomical regions. In addition to that, this type of categorization is not applied to other languages as well. The proposed system is first ever to classify Turkish radiology reports into respective anatomical regions by employing a dictionary-based approach.

3. DATA AND METHOD

3.1 DATA AND ENVIRONMENT

In this project, a total of 230 radiology reports belonging to 8 different anatomical classes are used which are compiled from “Maltepe Tıp Fakültesi Hastanesi” via “RadyolojiOnline”. A total of 69 documents are used for training data, while the remaining 161 documents are used for testing data.. Each report is in written in Turkish language, and has the same structure consisting of 3 main parts: 1) Header, including a table with patient information such as name, sex and age; 2) free text, and 3) footer including the names of contributing doctors. Reports also include punctuation marks, and Turkish characters (ğ,ü,ş,ç,ı,ö). As seen in Figure 3.1, some reports are long, whereas others may include just a single line of text. All reports are distributed in Microsoft Word format. Figure 3.2 shows the distribution of modalities in radiology reports. There are four modalities: MR (*magnetic resonance*), CR(*computed radiography*), US (*ultrasound*) and CT (*computerized tomography*). Most common modality is MR with 126 reports where others are 17, 9, and 10 respectively.

Figure 3.1: Screenshots of a long and short report from the database



Hasta Adı	: TELEK, MEHMET HUSEYIN(L)	Kayıt Tarihi	:10.08.2011 09:58:03
Doğum Tarihi	: 28.11.1952	Tetkik Bölgesi	SERVIKAL
Cinsiyet	: M	Modalite	: MR

Sayın Meslektaşım,
Hastanızın Servikal spinal MRG tetkikinde:

İnceleme tekniği;
Sagital planda SE, T1 ağırlıklı
Sagital planda FSE, T2 ağırlıklı
Aksial planda GRE, T2 ağırlıklı

Kranioservikal bileşke oluşumları doğaldır.

Servikal lordoz düzleşmiştir.

C3-4, C5-6 ve C6-7 intervertebral disklerinde yükseklik kaybının eşlik ettiği T2 sinyal kaybı dejenerasyon ile uyumludur. C5-6 eklem yüzlerinde yaygın dejeneratif sinyal değişiklikleri görülmektedir. C7 korpus ventralinde tüm sekanslard hiperintens nodüler oluşum hemanjiom ile uyumludur.

C3-4 intervertebral diskinde osteofitik sivrileşmeler ile birlikte laterallerde belirgin posterior bulging görülmektedir. Spinal kord serbesttir. Bilateral nöral foramen girişleri daralmıştır.

C5-6 intervertebral diskinde osteofitik sivrileşmeler ile birlikte solda belirgin posterior bulging izlenmektedir. Bu seviyede spinal korda serbesttir. Sol nöral foramen girişi daralmıştır.

C6-7 seviyesinde posterior osteofitik sivrileşmeler ventral subaraknoid mesafeyi daraltmaktadır. Spinal kord serbesttir. Nöral foramen girişleri hafif daralmıştır.

Diğer seviyelerde disk posterior konturları düzenli olup, anlamlı herniasyon bulgusu mevcut değildir.Nöral foramenler normal genişliktedir.

Spinal kord normal kalınlık ve sinyal özelliklerindedir.
İntra yada ekstra tekal kolleksiyon yada kitle lezyonu saptanmadı.
Spinal kanal AP ve transvers çapları normal sınırlardadır.
Vertebra korpus yükseklikleri normal sınırlardadır.
Paravertebral alanlarda patoloji saptanmadı.

SONUÇ : C3-4, C5-6 ve C6-7 seviyelerinde diskopati bulguları

Saygılarımızla,

Prof. Dr. Levent ÇELIK | Yrd.Doç.Dr. Nuri TASALI | Yrd.Doç.Dr. Rahmi ÇUBUK | Dr. Gül ARSLAN | Dr.Esra MERMİ | Dr.Şükran MANSUROĞLU

Hasta Adı	: AYDIN, AYSE NUR	Kayıt Tarihi	:10.08.2011 14:36:00
Doğum Tarihi	: 20.12.1951	Tetkik Bölgesi	BREAST
Cinsiyet	: F	Modalite	: US

Sayın meslektaşım,

Sağ memedeki lezyondan tru-cut biopsisi yapıldı.

Saygılarımızla;

Yrd.Doç.Dr. Nuri TASALI | Yrd.Doç.Dr. Rahmi ÇUBUK | Dr. Gül ARSLAN | Dr.Esra MERMİ | Dr.Şükran MANSUROĞLU

Figure 3.2: Distribution of imaging modalities from which the reports are created

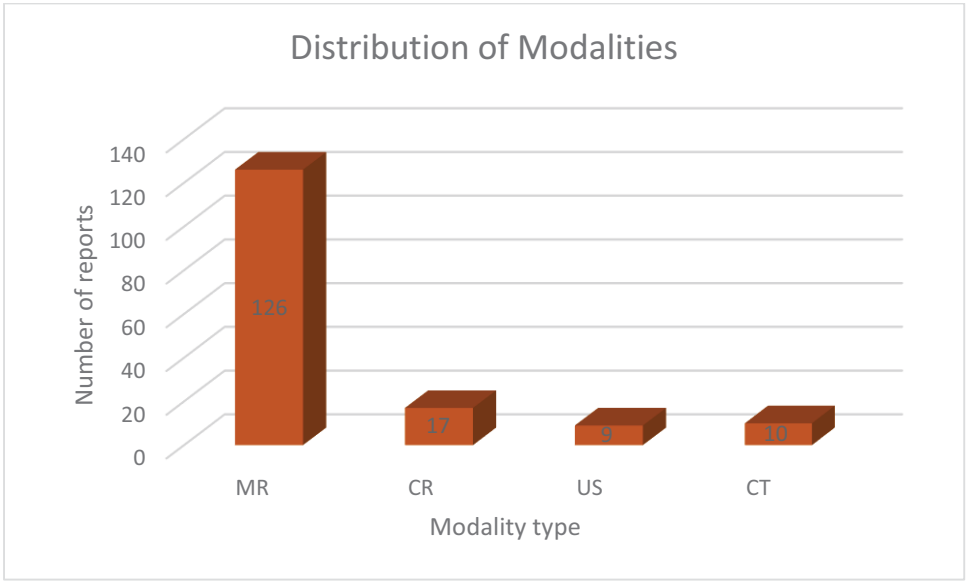


Table 3.1: Number of documents included in training and test datasets

Anatomical Region	Training Dataset	Test Dataset
Breast	10	13
Foot	10	14
Elbow	10	20
Head	6	43
Shoulder	6	19
Pelvis	16	23
Spine	6	26
Thorax	5	3
Total:	**69**	**161**

The system works on Ubuntu which is a GNU/Linux based operating system. Ubuntu is a free operating system, which enables any user to download and use it without any copyright issues.

The most important reason for the using this operating system is the availability of Bash. Bash, short for Bourne Again Shell, is a Unix shell written by programmer Brian Fox for the GNU Project as a replacement for Bourne Shell. Bash has its own scripting language called Bash Scripting, which enables the user to write and execute codes in a very simple way. The most important reason to use this scripting language in this project is because of the hard-coded scripts such as "*sed*" and "*grep*", which are very useful in text manipulation and processing.

Since the system works on Bash, a software is needed in order to read these *.doc documents. There is freeware program named *Antiword*, which can open these documents in Bash.

Student t-test calculations are done by *R-Project,* which is an open-source statistics application.

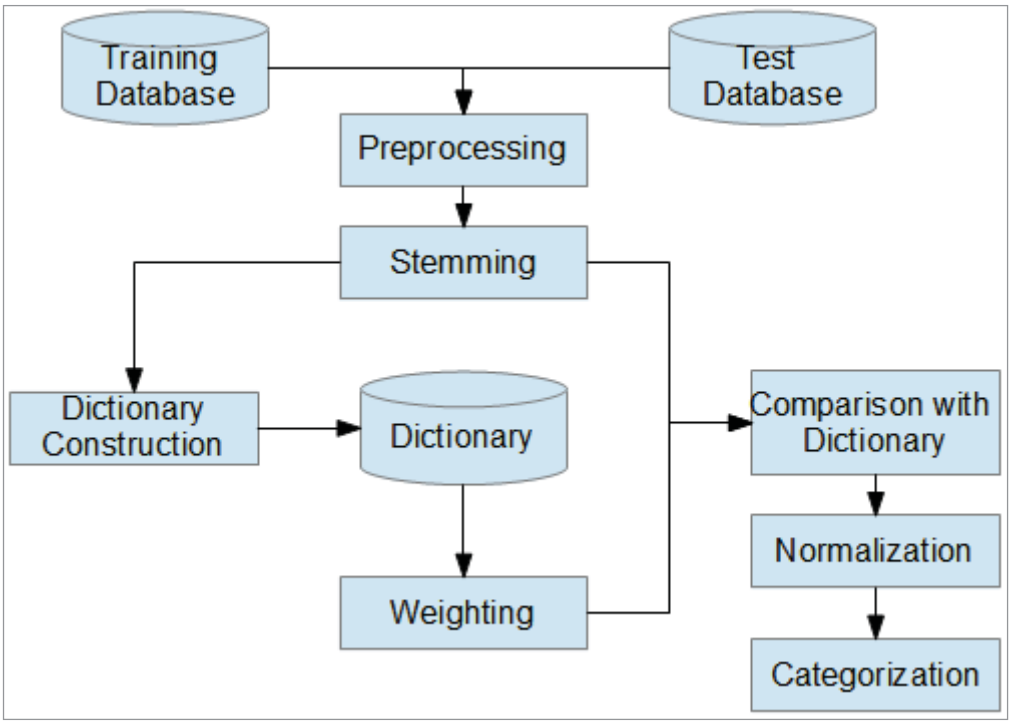
3.2 METHOD

Figure 3.3 shows the schematic view of the system and provides information about how the system works.

Both training and test data are collected from the same clinical site (Maltepe University Hospital), and thus share the same format.

Training the system includes three parts: frequency analysis, dictionary construction and weighting. Frequency analysis is applied onto all classes, and outputs a summary file presenting the frequency of every word present in the document list. Dictionaries are constructed by choosing distinctive words from these output files. This task is repeated for every class, thus providing a dictionary for every anatomical region. After constructing the dictionaries, the system calculates weights of words in dictionaries and updates them.

Figure 3.3: Schematic view of the system



3.2.1 Document Representation

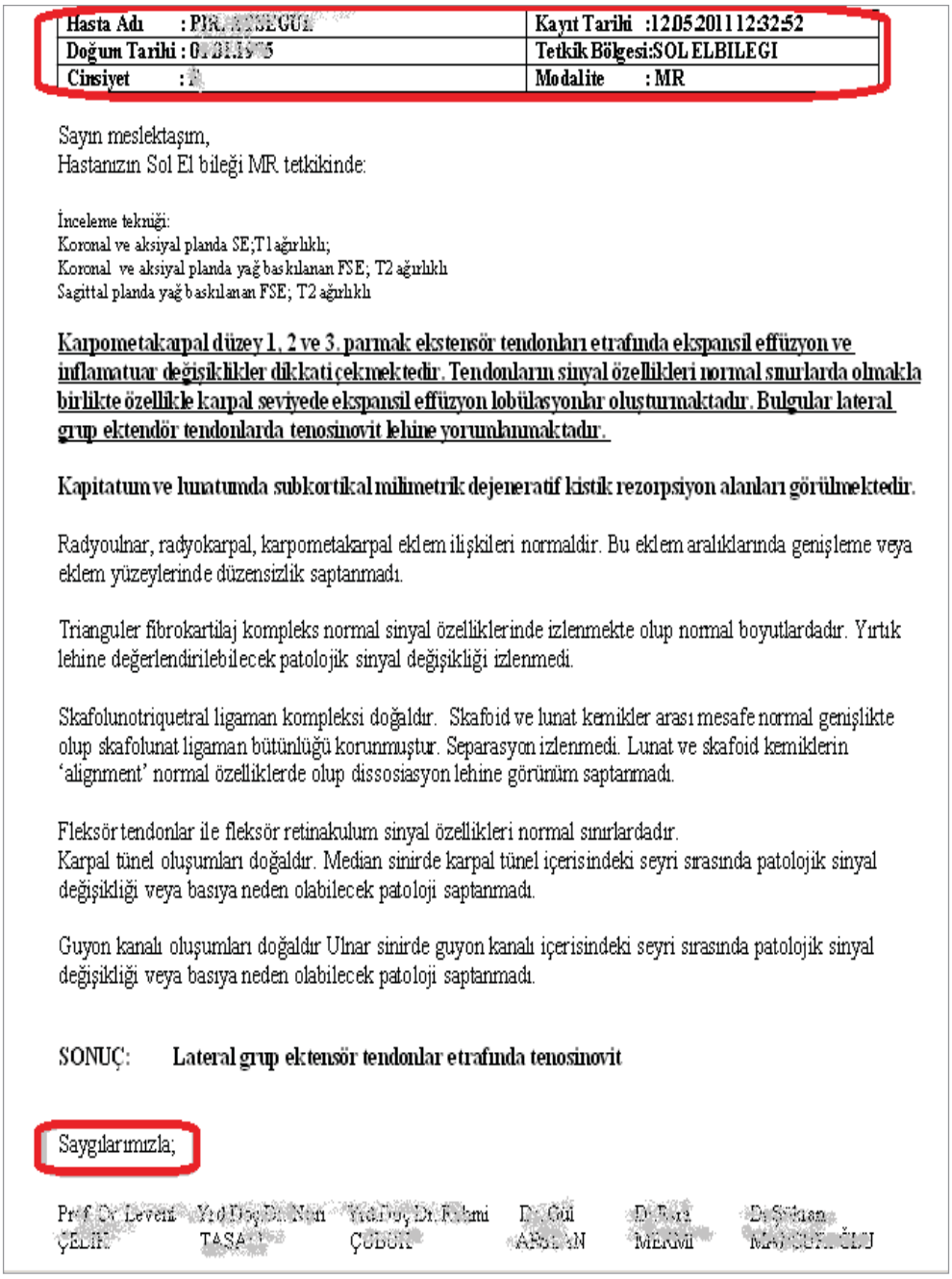
3.2.1.1 Preprocessing

Every report share the same format, and therefore, the same processing steps are applied to each report in order to obtain the report as a free-text.

i. Identification and removal of header and footer
ii. Removal of punctuation marks
iii. Conversion of characters to lower-case
iv. Stemming
v. Conversion of Turkish characters to English characters
vi. Replacing multi-spaces by single space
vii. Replacing spaces with lines

Originally, header is a table in the document, but *Antiword* changes and shows the table with the character " | ". Thus, the header is identified by this character, and the footer is identified with the polite ending "Saygılarımla". An example report of mentioned format is in Figure 3.4.

Figure 3.4: Header and footer highlighted in an example report



Hasta Adı : PIR. [illegible]SEGÜL	Kayıt Tarihi :12.05.2011 12:32:52
Doğum Tarihi : 0[illegible]5	Tetkik Bölgesi:SOL ELBILEGI
Cinsiyet : [illegible]	Modalite : MR

Sayın meslektaşım,
Hastanızın Sol El bileği MR tetkikinde:

İnceleme tekniği:
Koronal ve aksiyal planda SE;T1 ağırlıklı;
Koronal ve aksiyal planda yağ baskılanan FSE; T2 ağırlıklı
Sagittal planda yağ baskılanan FSE; T2 ağırlıklı

Karpometakarpal düzey 1, 2 ve 3. parmak ekstensör tendonları etrafında ekspansil effüzyon ve inflamatuar değişiklikler dikkati çekmektedir. Tendonların sinyal özellikleri normal sınırlarda olmakla birlikte özellikle karpal seviyede ekspansil effüzyon lobülasyonlar oluşturmaktadır. Bulgular lateral grup ektendör tendonlarda tenosinovit lehine yorumlanmaktadır.

Kapitatum ve lunatumda subkortikal milimetrik dejeneratif kistik rezorpsiyon alanları görülmektedir.

Radyoulnar, radyokarpal, karpometakarpal eklem ilişkileri normaldir. Bu eklem aralıklarında genişleme veya eklem yüzeylerinde düzensizlik saptanmadı.

Trianguler fibrokartilaj kompleks normal sinyal özelliklerinde izlenmekte olup normal boyutlardadır. Yırtık lehine değerlendirilebilecek patolojik sinyal değişikliği izlenmedi.

Skafolunotriquetral ligaman kompleksi doğaldır. Skafoid ve lunat kemikler arası mesafe normal genişlikte olup skafolunat ligaman bütünlüğü korunmuştur. Separasyon izlenmedi. Lunat ve skafoid kemiklerin 'alignment' normal özelliklerde olup dissosiasyon lehine görünüm saptanmadı.

Fleksör tendonlar ile fleksör retinakulum sinyal özellikleri normal sınırlardadır.
Karpal tünel oluşumları doğaldır. Median sinirde karpal tünel içerisindeki seyri sırasında patolojik sinyal değişikliği veya basıya neden olabilecek patoloji saptanmadı.

Guyon kanalı oluşumları doğaldır Ulnar sinirde guyon kanalı içerisindeki seyri sırasında patolojik sinyal değişikliği veya basıya neden olabilecek patoloji saptanmadı.

SONUÇ: Lateral grup ektensör tendonlar etrafında tenosinovit

Saygılarımızla;

[illegible]

Punctuation marks are deleted, and every word is converted into lower-case. Stemming is performed before changing Turkish characters into English characters. The reason of this transformation is to eliminate the possibility of any mistake done at the recording of the reports. Simple formatting issues are corrected such as shortening every empty

space between two words into one space. Each space is replaced with a new line, and thus creating a list.

3.2.1.2 Stemming

Stemming is the process of finding the stem of every word. Turkish is a highly agglutinative language having many grammatical suffixes which can change the meaning of a word. This reason makes stemming an important part of analyzing the reports.

Porter's stemmer is a well-known stemming algorithm designed for stemming English words. (Porter 1980) Since our system works with Turkish reports, Porter's algorithm proves uneffective. However, M.F. Porter, founder of Porter's algorithm, designed a scripting language named Snowball, which has a simple structure where people can create stemmers. Snowball also includes a stemmer for Turkish language which is designed by Cilden (2006). This stemmer is also tested, but lacked some critical stemming needs such as stemming the word “memede” into “meme”, therefore it is not used in our system.

Zemberek is an open-source, platform independent, general purpose Natural Language Processing library designed for Turkic languages. It is also used as a stemmer in different projects, and proved effective. It is also used in many other projects successfully (Pala and Cicekli (2007), Cataltepe *et al.* (2007), Yildiz *et al*. (2007)). Our preliminary analyses revealed that Zemberek is more accurate and robust for our problem, hence it is used in our system.

For experimental purposes, another stemming approach is used which cuts every word into first 4-5-6 characters. This approach tries to eliminate the suffixes and leave the word with only the base.

3.2.2 Training

3.2.2.1 Frequency Analysis

Frequency analysis is the study of finding the frequency of every word in a document. Frequency of a word is a number which indicates the repetition of that word in the document.

The main purpose of frequency analysis is to understand which words are used more than others, for a given anatomical region. (See Appendix-1, pg. 31)

3.2.2.2 Dictionary Construction

Figure 3.5 displays an exemplary frequency analysis result with 8 different summary files, including the repetition of every word for each anatomical region.

Figure 3.5: Example of a frequency analysis output

```
24 mevcuttur
20 yag
20 t2
20 dejeneratif
19 tendonunda
19 supraspinatus
19 mesafesi
18 miktarda
17 normaldir
16 tendon
16 ozelliklerindedir
16 bursada
12 ruptur
12 humerus
11 tendonu
11 hafif
11 akromioklavikuler
```

As it can be seen in Figure 3.5, there are both distinctive and non-distinctive words present in a frequency analysis output. Since our system works with medical reports, there are meaningful but undistinctive words which are not relevant for categorizing the

reports.

That is why instead of creating a list of stopwords, the meaningful words are manually selected and put into a dictionary (shown in Table 3.2). Since there are no pre-constructed dictionaries including distinctive words for each anatomical region, these dictionaries are constructed manually by examining the training dataset.

Table 3.2: Dictionary terms manually selected from the training set using frequency analysis results

Region Name	Discriminative Words
Breast	*Meme, fibrokistik*
Foot	*Ayak, metatars, sesamoid, metatarsofalengeal*
Elbow	*Karpal, fleksor, skafoid, radyokarpal, karpometakarpal, dirsek*
Head	*Ventrikul, korpus, kranial, sisternalar, serebral, kallosum, serebellar, mastoid*
Shoulder	*Subakromial, glenoid, glenohumeral, omuz, akromion, subdeltoid, supraspinatus*
Pelvis	*Sakroiliak, femur, koksofemoral, sakral, pelvis, pubis, femoris, suprapubik, koksiks*
Spine	*Intervertebral, noral, dural, lomber, vertebra, spinal, disk, herniasyon*
Thorax	*Akciger, toraks, pankreas, trakea, kardiak, abdomen*

3.2.2.3 Weighting

The frequency of each word is an important, but an unreliable feature for classification, because it can largely effect (bias) the system's performance if the doctors mistakenly repeat some words. A good but malevolent example for this is as follows, in world wide web, it is common to use a specific search term with same color as the background, to increase the frequency of that word in the source code. It is highly possible that a word with high frequency is included in less number of documents but has a high frequency because it is repeated in those documents. The frequency of the word does not imply

that word is rare or common in the document list. That is why, the number of repetitions of a single word in a single document is not as important as the number of repetitions in the document collection.

The weight of a word is defined as in Equation 3.1. W_{word} is the weight of the word, N_{word} is the number of documents including the word, and Z is the total number of documents in that class.

$$W_{word} = \frac{N_{word}}{Z} \tag{3.1}$$

By this equation, we incorporate the significance of each word for each class in the classification. Weighting also helps in case of having identical words for different classes. By applying this, it can be identified in which class the word has more significance. (See Appendix-2 and Appendix-9, Figure 2)

TF-IDF score is a popular weighting scheme used in information retrieval. Experiments with this weighting scheme are also conducted and compared with our weighting scheme. As it can be seen in Equation 3.2, TF-IDF equals to the multiplication of term frequency with inverse document frequency, where DF equals to the number of documents with mention of term t and N is the total number of documents.

$$TFIDF_t = TF_t \; x \; log \frac{N}{DF_t} \tag{3.2}$$

3.2.3. Test

3.2.3.1 Normalization

As it is seen in Table 3.2, the number of words chosen for each class is not the same. This inequality can create a bias because the number of words to compare is different. To neutralize this bias, a normalization is needed.

There are different normalization techniques available in the literature. In this project, the normalization of wordlists are done by using a unit vector logic. Unit vector is a vector whose length is 1. This unit vector is obtained by multiplying the length of the vector with a specific number to make it 1, which is simply dividing itself by the length of itself.

Thus, dividing the score of that class by the number of words included in the wordlist produces the normalized score.

Normalization does not change the outcome drastically, but it is a needed step to eliminate the bias. Also, lack of normalization can result in cases where multiple classes have the same score. In this scenario, the system would need to choose the class with less number of words, and normalization will help the system in this manner.

3.2.3.3 Scoring

After each comparison, the software gives a score by counting the number of repetition for each word. M stands for number of matches where every match is scored by using Equation 3.3 which states that a each match has a score of M times the weight of the matching word. Later, maximum class score is chosen (Equation 3.4). The weighting scheme is explained in detail in the following sections. (See Appendix-3 and Appendix-8, Figure 1)

$$Score_{class} = \sum M\ X\ W_{word} \tag{3.3}$$

$$Score = \max(Score_{class}) \tag{3.4}$$

3.2.4 Evaluation Metrics

Performance measures are needed in text classification to evaluate the success of the proposed system. It is logical to assume that an ideal retrieval system would have a precision and recall percentage of 100.

3.2.4.1 Success Scores

In order to evaluate the performance of the system, the popular F-score is used, which is defined as a weighted combination of Precision and Recall (Makhoul *et al.* 2004). Recall (ρ) is defined as the percentage of relevant documents which are retrieved and precision (π) is defined as the percentage of retrieved document that are relevant.

$$\pi_i = \frac{TP_i}{TP_i + FP_i} \tag{3.5}$$

$$\rho_i = \frac{TP_i}{TP_i + FN_i} \tag{3.6}$$

Equations above show the equations of precision and recall respectively. Here, TP_i (True Positive) shows the number of documents classified correctly to class *i*, FP_i (False Positive) shows the number of documents that do not belong to class *i*, but classified as such, and FN_i (False Negative) shows the number of documents that are not assigned to

class i, but belongs to class *i*.

F-score values range between 0 and 1, and larger f-score values indicates high classification quality (Özgür et al. 2005)
As it can be seen in the Equation below, F-score is the harmonic mean of precision and recall, and the reason for this is to minimize the impact of large outliers and maximizing the impact of small value (Nadeau and Sekine 2007).

F-score is computed globally over all categories. This approach gives equal weight to each class, thus is an average over all categories.

$$F = \frac{2\pi\rho}{\pi + \rho} \tag{3.7}$$

Accuracy is also used to evaluate the performance of the system, and it is defined as the proportion of true positive results across all population.

$$A = \frac{TP_i}{TP_i + TN_i + FP_i + FN_i} \tag{3.8}$$

3.2.4.2 Statistical Significance

Hull (1993) claimed that an evaluation study is not complete without measuring the significance of the differences between retrieval methods. Statistical significance tests provide these results and they are useful because they can show if the difference in results are meaningful or by chance.

There are many testing methods for statistical significance, but t-test (often referred to as students t-test) is one of the most widely used. It is argued that t-test is more reliable

than just showing a percentage difference (Sanderson and Zobel 2005). Also, Cormack and Lynam (2007) compared t-test with Wilcoxon test and sign test and determined that t-test proves superior.

In testing, there are null and alternative hypotheses. Null hypothesis ($\mathbf{H_0}$) states that all methods are equivalent in terms of performance, whereas alternative hypothesis defines the opposite. Before the test, a p-value $\boldsymbol{\alpha}$ is chosen, and if the outcome of the test yields a smaller value than that $\boldsymbol{\alpha}$, it can be said that tested methods are statistically different than each other.

4. FINDINGS

Table 4.1 shows accuracy and F-scores for both training and test documents. Here an overall accuracy of 97,10 percent is measured for training documents, and 96,32 percent is measured for test documents. Also, f-scores are measured 98,07 percent for training documents, and 98,68 percent for test documents.

Table 4.1: Accuracy of the system on training and test documents displayed for every region

	Training Documents		Test Documents	
Name of the Region	**Accuracy (%)**	**F-Score (%)**	**Accuracy (%)**	**F-Score (%)**
Breast	100	100	100	100
Foot	100	100	100	100
Elbow	100	100	94,1	97
Head	100	100	95,4	97,6
Shoulder	83,3	90,9	100	100
Pelvis	93,8	92,3	95,7	97,8
Spine	100	100	93,8	96,8
Thorax	100	100	100	100
Overall	**97,1**	**98,1**	**96,3**	**98,7**

4.1 EXPERIMENTS

In order to observe the importance of terms present in the dictionaries, a series of experiments are conducted. To this end four different dictionaries with increasing number of words are created: *one-keyword*, *two-keywords*, *three-keywords* and *full dictionary*.

One-keyword dictionary only consists of the first (most frequent) element of the dictionary, whereas *two-keywords* includes the most frequent two elements, and *three-keywords* includes the most frequent three elements.

Second experiment is done on stemming algorithm. Along Zemberek, a word-cutting algorithm is used where the system only takes first 4-5-6 characters of every word and combines the same bases.

Third experiment aims to compare weighting techniques. Originally, weights are document frequencies (DF) of every word, but in this experiment these values are changed with TF-IDF scores.

Fourth experiment aims to see if TF-IDF is a good indicator of distinctive words. Here, the dictionary is constructed automatically by choosing first 5 words based on their TF-IDF scores.

Fifth experiment aims to see if TF-DF is a good indicator of distinctive words. Here, as in fourth experiment, the dictionary is constructed automatically by choosing first 5 words per category based on their TF-DF scores.

4.2 RESULTS

In Table 4.2, the importance of having more distinctive words in class dictionaries are shown. It can be observed that precision, recall, F-Score and accuracy are all increased with increasing number of terms. Since the possible bias from different number of elements are eliminated by normalization, having more distinctive words in dictionaries are effective. (See Appendix 4-7, pg. 37-40)

P-values being smaller than 0.05 show that the difference in all metrics is not by chance, and it is statistically significant.

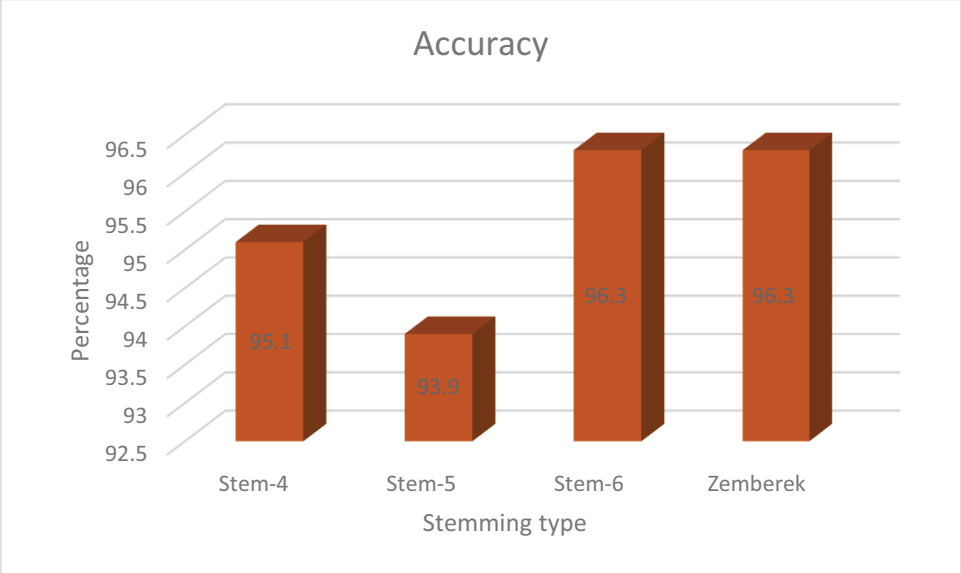
A system with an F-score higher than 90% is considered highly accurate for text classification (Taira and Soderland 1999). Accordingly, the proposed system is highly accurate even with the *two-keyword* dictionary.

Table 4.2: Effect of dictionary size on system's accuracy

Dictionary	Precision (%)	Recall (%)	F-Score (%)	Accuracy (%)	P-value
One-keyword	96,9	88	92,2	85,7	0,000008
Two-keywords	96,3	93,9	95,1	90,8	0,003
Three-keywords	99,7	95,1	97,3	93,3	0,014
Full	**99,2**	**98,1**	**98,7**	**96,3**	N/A

In Figure 4.1, it is observed that the highest accuracy is obtained with Zemberek and Stem-6, followed by Stem-4 and lastly, Stem-5. Stem-N works by removing characters from the right-end until N characters are left.

Figure 4.1: Effect of different stemming techniques on system's accuracy.



This stemming approach has also some benefits. By using Zemberek, the system does classification in 229 seconds, where Stem-6 does the same job in 135 seconds, which is a 69,63 percent increase.

Table 4.3: Effect of weighting techniques on system's accuracy

Weighting Type	Precision (%)	Recall (%)	F-Score(%)	Accuracy(%)	P-Value
TF-IDF	100	83,64	87,64	85,88	N/A
Present	99,2	98,1	98,7	96,3	0,00002

In Table 4.3, it can be observed that using our weighting scheme evaluates better percentages for recall, F-score and accuracy. The main reason for this decrease in percentages is the application of TF-IDF. This weighting scheme discards the keywords whose document frequency is 1, and these keywords are often very distinctive words such as meme, omuz, ayak. P-value is smaller than 0.05, and it proves that the difference between these results are statistically significant.

In the fourth experiment, the system computed TF-IDF scores for every word and automatically selected the top 5 words by the TF-IDF score as dictionary keywords. Those dictionaries are then used to evaluate the system performance, however the accuracy is measured to be very low (13.50 percent). This experiment shows that constructing dictionaries automatically using TF-IDF scores for the presented scenario is not effective. The reason for this low performance is most probably due to the IDF approach that favors keywords that are less frequent in a document set, whereas in our scenario we opt for keywords that are most representative of (more frequent in) a document set.

In the fifth experiment, the system computed TF-DF scores for every word and automatically selected the top 5 words. Those dictionaries are then used to evaluate the system performance, and accuracy is measured to be 42.95 percent.

Automatic selection of dictionary keywords proved uneffective with both weighting schemes. The main reason for this low performance is that term frequency, document frequency and inverse document frequencies are not able to recognize distinctive terms effectively by themselves. A human-engineered dictionary construction proved much more effective as it can be seen in Table 4.2.

4.3 TIME OF COMPUTATION

In order to evaluate the system's computational expense, three trials are done for each test with two different sizes of documents (100 and 1000) and the average process times are given in Table 4.4. These trials are done on a computer with an Intel Core i7 CPU which works at 1.73 GHz, and 4.00 GB RAM. Frequency Analysis took 11,50 seconds for 100 documents, and 80,79 seconds for 1000 documents. The classification process took 61,09 seconds for 100 documents, and 500,39 seconds for 1000 documents.

Table 4.4: Computation time for 100 and 1000 documents

	Frequency Analysis (seconds)		**Classifier (seconds)**	
# of Documents	100	1000	100	1000
Trial 1	42,8	576,9	88,4	1428,9
Trial 2	42,6	576,4	89,5	1430,9
Trial 3	42,7	576,9	88,9	1428,1
Average	**42,7**	**576,7**	**89**	**1429,3**

4.4 ERROR CORRECTION

Our system identified and corrected seven documents into their true respective regions. These documents were pre-labeled documents by the radiology staff. The system proved effective in checking whether this manual labeling is correct or not.

Table 4.5 shows the corrections done by the system after manually checking the mismatches. It is observed that two *shoulder* and two *elbow* documents actually belonged to *spine* class, one *foot* document belonged to *elbow* class, and two *elbow* documents belonged to *foot* class. So, the proposed system corrected 7 previously mis labeled reports.

Table 4.5: Document class corrections accomplished by the proposed system

Document's Pre-Defined Class	Class Assigned by the System
Shoulder	Spine
Shoulder	Spine
Elbow	Spine
Elbow	Spine
Foot	Elbow
Elbow	Foot
Elbow	Foot

5. CONCLUSION AND FUTURE WORK

This thesis proposes a new dictionary-based classification system to categorize Turkish radiology documents based on their anatomical regions, and reports a correct categorization rate of 94% over a database of 230 reports acquired from routine clinical practice. It consists of two parts: training and application. The system learns which words are discriminative for pre-defined anatomical regions in the former, and it categorizes new reports in the latter. Different scoring schemes are used in order to see their contribution to the system's performance.

Quantitative evaluation of the system's performance on a dataset of 230 radiology reports revealed an average recall rate of 98,14%, precision rate of 99,21%, and F-score of 98,6%. Furthermore, the system identified and corrected seven mislabelings (errors) done by the radiology staff.

The accuracy of the proposed system could be improved by increasing the training set size, benefiting from natural language processing solutions (e.g. word correction), and exploiting anatomical/pathological information encoded in ontologies. In the future, the proposed system can be (1) combined with a speech recognition solution to automatically convert dictations of radiologists into written reports, (2) extended to take diagnostic decision by combining information from radiology reports, ontologies and image data, and (3) implemented a feedback mechanism where user identifies incorrect classifications done by the system and recommended new keywords to be added in the dictionary.

REFERENCES

Periodicals

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

Robertson, S.E., 1977. The probability ranking principle. *Journal of Documentation*. **33**, pp.294-304.

Rijsbergen, C.J., 1986. A non-classical logic for information retrieval. *The Computer Journal*. **29** (6), pp.481-485.

Aronow, D.B., Fangfang, F., & Croft, W.B., 1999. Ad hoc classification of radiology reports. *Journal of the American Medical Informatics Association*. **6**(5), pp.393-411

Friedman, C., Alderson, P.O., 1994. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association.* **1**(2), pp.161-74.

Thomas, B.J., Ouellette, H., Halpern, E.F., & Rosenthal, D. I., 2005, Automated computer-assisted categorization of radiology reports. *American Journal of Roentgenology*. **184**(2), pp.687-90

Lakhani, P., Kim, W., & Langlotz, C.P., 2012. *Automated detection of critical results in radiology reports. Journal of Digital Imaging*. **25**(1), pp.30-36

Prasad, A.K., Ramakrishna, S., Kumar, D.S., & Rani, B.P., 2010. Extraction of radiology reports using text mining. *International Journal on Computer Science and Engineering.* **2**(05), pp.1558-62

Lacson, R., Khorasani, R., 2011. Natural language processing: the basics (part 1). *Journal of the American College of Radiology*. **8**(6), pp. 436–7

Goldstein, I., Arzrumtsyan, A. & Uzuner, O., 2007. Three approaches to automatic

assignment of ICD-9-CM codes to radiology reports. *AMIA Annual Symposium Proceedings,* **11,** pp.279–83

Hripcsak, G., Austin, J. H., Alderson, P. O., & Friedman, C., 2002. Use of natural language processing to translate clinical information from a database of 889.921 chest radiographic reports. *Radiology*. **224**(1), pp.157-163

Thrall, J.H., 2005. Reinventing radiology in the digital age: Part I. The all-digital department, *Radiology*, **236**(2), pp.382-85

Reiner, B., 2010. Uncovering and improving upon the inherent deficiencies of radiology reporting through data mining. *Journal of Digital Imaging*. **23**(2), pp.109-18

Dunnick, N.R., & Langlotz, C.P., 2008. The radiology report of the future: a summary of the 2007 intersociety conference. *Journal of the American College of Radiology*. **5**(5), pp.626-29

Nadeau, D., Sekine, S., 2007. A survey of named entity recognition and classification. *Lingvisticae Investigatioes*. **30**(1), pp.3-26

Arisoy, E., Dutagaci, H., & Arslan, L.M., 2006. A unified language model for large vocabulary continous speech recognition of Turkish. *Signal Processing*. **86**(10), pp.2844-62

Soysal, E., Cicekli, I. & Baykal, N., 2010. Design and evaluation of an ontology based information extraction system for radiological reports. *Computers in Biology and Medicine*. **40**(11), pp.900-11

Mamlin, B.W., Heinze, D.T. & McDonald, C.J., 2003, Automated extraction and normalization of findings from cancer-related free-text radiology reports. *AMIA Annual Symposium Proceedings.* **2003,** pp.420–4

Regev, Y., Finkelstein-Landau, M. & Feldman, R., 2002. Rule-based extraction of experimental evidence in the biomedical domain: The KDD Cup 2002 (task 1)*, ACM SIGKDD Explorations Newsletter.* **4**(2), pp.90–92.

Porter, M.F., 1980. An algorithm for suffix stripping. *Program: electronic library and information systems*. **14**(3), pp.130-7

Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., & Lioma, C., 2006. Terrier: A high performance and scalable information retrieval platform. *Proceedings of*

the OSIR Workshop, pp.18-25

Apte, C., Damerau, F. & Weiss, S.M., 1994. Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems (TOIS)*, **12**(3), pp. 233-251.

Other

NLM, Fact Sheet, 1999, http://www.nlm.nih.gov/pubs/factsheets/mesh.html [retrieval date 20 May 2013]

Alvarez, S.A., 2002. An exact analytical relation among recall, precision and classification accuracy in information retrieval. *Technical Report BCCS-02-01*. Boston College, Boston: USA

Gong, T., Tan, C.L., Leong, T.Y., Lee, C.K., Pang, B.C., Lim, C.C.T., Tian, Q., Tang, S. & Zhang, Z., 2008. Text mining in radiology reports. *Data Mining, 2008, ICDM'08. Eighth IEEE International Conference*, 15-19 December Pisa, Italy: pp. 815-20

Taira, R.K., Soderland, S.G., 1999. A statistical natural language processor for medical reports. *Proceedings of AMIA Symposium,* American Medical Informatics Association, p.970

Hull, D., 1993. Using statistical testing in the evaluation of retrieval experiments. *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval.* July 1993 New York, NY, USA, pp.329-38

Sanderson, M. & Zobel, J., 2005. Information retrieval system evaluation: effort, sensitivity, and reliability. *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.162–9.

Cormack, G. V. & Lynam, T. R., 2007. Validity and power of t-test for comparing map and gmap. *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.*

Ozgur, A., Ozgur, L. & Gungor, T., 2005. Text categorization with class-based and corpus-based keyword selection. 26-28 October 2005 Istanbul, Turkey*: Computer and Information Sciences-ISCIS 2005.* Springer Berlin Heidelberg, pp.606-615

Yildiz, H.K., Genctav, M., Usta, N., Diri, B., and Amasyah, M.F., 2007. A new feature extraction method for text classification. 11-13 June 2007 Eskisehir, Turkey: *IEEE 15th Signal Processing and Communications Applications, SIU 2007*, pp.1-4.

Cataltepe, Z., Turan, Y. and Kesgin, F., 2007. Turkish document classification using shorter roots. 11-13 June 2007 Eskisehir, Turkey: *IEEE 15th Signal Processing and Communications Applications, SIU 2007,* pp.1-4.

Pala, N., Cicekli, I., 2007. Turkish keyphrase extraction using KEA. 7-9 November 2007 Ankara, Turkey: *22nd international symposium on Computer and information sciences*, *ISCIS 2007*, pp.1-5

Oguz, B., Bilge, U., & Saka, O., 2007. Text mining in medicine. 15-18 November Antalya, Turkey: *Medical Informatics '07, 4th TurkMIA Congress.*

Ceylan, N.M., Alpkocak, A., Esatoglu, A.E., 2012. An intelligent system to help on assignment of ICD-10 codes to medical records. 15-17 November Antalya, Turkey: *Turkish Medical Informatics Conference, TurkMIC'12*.

Maghsoodi, A., Sevenster, M., Scholtes, J., & Nalbantov, G., 2012, Sentence-based classification of free-text breast cancer radiology reports. *Computer-Based Medical Systems (CBMS), 2012 25th International Symposium,* 20-22 June Rome, Italy: pp.1-4

Cilden, E.K., 2006. Stemming turkish words using snowball. *Snowball ile Çalışmalar Raporu,* Cybersoft Information Technologies: Turkey.

Makhoul, J., Kubala, F., Schwartz, R. & Weischedel, R., 1999, February. Performance measures for information extraction. *Proceedings of DARPA. Broadcast News* Workshop, pp.249-52

APPENDICES

APPENDIX-1: FREQUENCY ANALYSIS CODE: DESTROY

This code performs frequency analysis.

```
#!/bin/bash
STEMTOOL="../stembro/run"
for FOLDER in learningDocuments/* ;
do
   if [ -d $FOLDER ]; then
      echo "Processing $FOLDER ..."
         for REPORT in $FOLDER/*.doc; do antiword $REPORT | sed -e "/^|/d" -e
"/Sayg/{Q}" | tr "\n" " " | sed -e "s/[^a-zA-Z0-9'-]/ /g" -e "s/[ \t\r]\+/ /g" | tr "[:upper:]"
"[:lower:]" | tr " " "\n" | sed -e "s/-//g" | $STEMTOOL | sed -e
"y/ğüşıöçĞÜŞİÖÇ/gusiocGUSIOC/"; done | sort | uniq -c | sort -r -g >
$FOLDER.summary
   fi
done
```

APPENDIX-2: WEIGHTING CODE: LOKI

This code performs weighting.

```
#!/bin/bash
cd words/frequency
FREQ_FILES=*
for DOG in $FREQ_FILES
do
   echo $DOG
   REPORTCOUNT=$(ls -1 ../../learningDocuments/$DOG/*.doc | wc -l)
   WORDS=$(cut -d" " -f2 $DOG)
   rm ../weight/$DOG 2> /dev/null
   for WORD in $WORDS
   do
#        echo "  $WORD"
         WORDCOUNT=0
         for REPORT in ../../learningDocuments/$DOG/*.doc
         do
            antiword $REPORT | sed -e "/^|/d" -e "/Sayg/{Q}" | tr "\n" " " | sed -e "s/[^a-
zA-Z0-9'-]/ /g" -e "s/[ \t\r]\+/ /g" -e "s/ş/s/g" -e "y/ğüşıöçĞÜŞİÖÇ/gusiocGUSIOC/" | tr
"[:upper:]" "[:lower:]" | tr " " "\n" | grep -i "$WORD" > /dev/null
            if [ "$?" == "0" ]; then
                  WORDCOUNT=$(( $WORDCOUNT + 1 ))
            fi
         done
                  if [ $WORDCOUNT -gt 0 ]; then
            WEIGHT=$(echo "scale=4;$WORDCOUNT / $REPORTCOUNT" | bc -l)
            echo "         $DOG:$WORD:$WORDCOUNT/$REPORTCOUNT =
$WEIGHT"
            echo "$WEIGHT:$WORD" >> ../weight/$DOG
         fi
   done
```

done

APPENDIX-3: APPLICATION CODE: THOR

This code performs categorization.

```
#!/bin/bash
#LC_ALL=en_US.UTF-8
STEMTOOL="../stembro/run"
CATEGORY_FILES=words/weight/*
CATEGORIES=(  )
for CAT in $CATEGORY_FILES
do
   CAT=$(echo $CAT | grep -o "[^/]\+$" )
   CATEGORIES=( ${CATEGORIES[@]} $CAT )
done
TOTAL=$(cat $CATEGORY_FILES | wc -l)
COUNT=$(( ${#CATEGORIES[@]} - 1 ))
for INDEX in $(seq 0 $COUNT)
do
   CAT=${CATEGORIES[$INDEX]}
   PTR=WORDS_$CAT
   KEYWORDS=$(cat words/weight/$CAT)
   eval "$PTR='$KEYWORDS'"
done
DOCUMENTS=$(find testDocuments/ -name "*.doc")
#DOCUMENTS=$(find learningDocuments/ -name "*.doc")
for DOC in $DOCUMENTS
do
   # split the query document into lowercase words and then stem them
   WORDS=$(antiword $DOC | sed -e "s/-//g" -e "/^|/d" -e "/Sayg/{Q}" | tr "\n" " " | sed
-e "s/[^a-zA-Z0-9'-]/ /g" -e "s/[ \t\r]\+/ /g" | tr "[:upper:]" "[:lower:]" | tr " " "\n" |
$STEMTOOL | sed -e "y/ğüşıöçĞÜŞİÖÇ/gusiocGUSIOC/")
   # category index
```

```
  I=0
  # index of the maximum best category
  I_MAX=0
  # score of the maximum best category
  O_MAX=-1
  #echo "Checking $DOC with ${#WORDS} words"
  for CAT in $CATEGORY_FILES
  do
    #echo "We are in $CAT"
    CAT=$(echo $CAT | grep -o "[^/]\+$" )
    PTR=WORDS_$CAT
    KEYWORDS=${!PTR}
      #no normalization below
    KEYWORD_COUNT=1
      #normalization below
      #KEYWORD_COUNT=$(cat words/frequency/$CAT | wc -l)
    SCORE=0
    for KEYWORD_DEFINITION in $KEYWORDS
    do
        KEYWORD_PARTS=(${KEYWORD_DEFINITION//:/ })
        WEIGHT=${KEYWORD_PARTS[0]}
        KEYWORD=${KEYWORD_PARTS[1]}
      # TODO: use weighted score here
      KEYWORD_SCORE=$(echo -e "$WORDS" | grep -c "^$KEYWORD$")
      KEYWORD_WEIGHTED_SCORE=$(echo "$KEYWORD_SCORE *
$WEIGHT" | bc -l)
      SCORE=$( echo "$SCORE + $KEYWORD_WEIGHTED_SCORE" | bc -l)
        IS_GREATER_THAN=$(echo "$KEYWORD_WEIGHTED_SCORE > 0" |
bc -l)
        if [ "$IS_GREATER_THAN" == "1" ]; then
        #echo -e "\t$KEYWORD: $KEYWORD_WEIGHTED_SCORE"
```

```
                echo -e "\t$KEYWORD: $KEYWORD_WEIGHTED_SCORE" >
/dev/null
            fi
        done
            SCORE=$(echo "scale=4;$SCORE / $KEYWORD_COUNT" | bc -l)
        #echo -e "score of $CAT: $SCORE\n"
            IS_GREATER_THAN=$(echo "$SCORE > $O_MAX" |  bc -l)
            if [ "$IS_GREATER_THAN" == "1" ]; then
            O_MAX=$SCORE
            I_MAX=$I
        fi
        I=$(( $I + 1 ))
    done
    echo "$DOC belongs to ${CATEGORIES[$I_MAX]} ($O_MAX)"
done
```

APPENDIX 4: TABLE 1: RECALL, PRECISION, F-SCORE AND ACCURACY PERCENTAGES FOR TEST DOCUMENTS

	TP	FP	FN	Recall (%)	Precision(%)	F-Score(%)	Accuracy (%)
Spine	30	2		100	93,8	96,8	93,8
Thorax	3			100	100	100	100
Shoulder	17			100	100	100	100
Head	41		2	95,4	100	97,6	95,4
Foot	15			100	100	100	100
Elbow	16		1	94,1	100	97	94,1
Pelvis	22		1	95,7	100	97,8	95,7
Breast	13			100	100	98,6	100

APPENDIX-5: TABLE 2: RECALL, PRECISION, F-SCORE, ACCURACY PERCENTAGES FOR SINGLE-KEYWORD EXPERIMENT

	TP	FP	FN	Recall (%)	Precision(%)	F-Score(%)	Accuracy (%)
Spine	30			100	100	100	100
Thorax	3	1		100	75	85,7	75
Shoulder	16		1	94,1	100	97	94,1
Head	38		5	88,4	100	93,8	88,4
Foot	15			100	100	100	100
Elbow	14		3	82,4	100	90,3	82,4
Pelvis	9		14	39,1	100	56,3	39,1
Breast	13			100	100	100	100

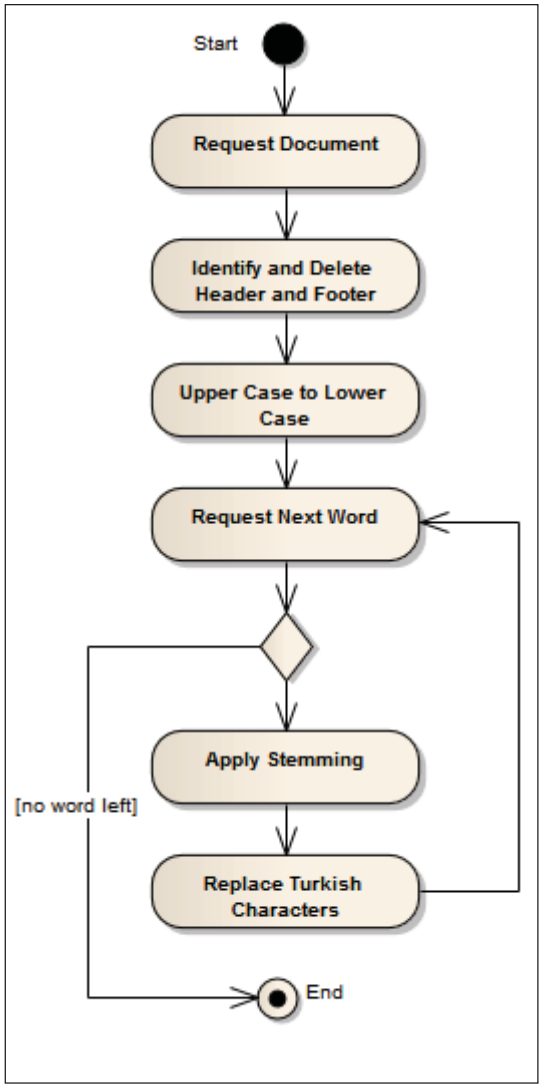
APPENDIX-6: TABLE 3: RECALL, PRECISION, F-SCORE, ACCURACY PERCENTAGES FOR TWO-KEYWORD EXPERIMENT

	TP	FP	FN	Recall (%)	Precision(%)	F-Score(%)	Accuracy (%)
Spine	29		1	96,7	100	98,3	96,7
Thorax	3	1		100	75	85,7	75
Shoulder	17			100	100	100	100
Head	38	2	5	88,4	95	91,6	84,4
Foot	15			100	100	100	100
Elbow	15		2	88,2	100	93,8	88,2
Pelvis	18		5	78,3	100	87,8	78,3
Breast	13			100	100	94,6	100

APPENDIX-7: TABLE 4: RECALL, PRECISION, F-SCORE, ACCURACY PERCENTAGES FOR THREE-KEYWORD EXPERIMENT

	TP	FP	FN	Recall(%)	Precision(%)	F-Score(%)	Accuracy(%)
Spine	30			100	100	100	100
Thorax	3	1		100	100	100	100
Shoulder	17			100	100	100	100
Head	40		3	93,0	97,6	95,2	90,9
Foot	15			100	100	100	100
Elbow	15		2	88,2	100	93,8	88,2
Pelvis	19		5	79,2	100	88,4	79,2
Breast	13			100	100	100	100

APPENDIX-8: FIGURE 1: PREPROCESSING FLOWCHART



APPENDIX-9: FIGURE 2: WEIGHTING FLOWCHART