TURKISH MEDICAL TEXT PARSING AND CLASSIFICATION

(TÜRKÇE MEDİKAL METİN AYRIŞTIRMA VE SINIFLANDIRMA)

by

AHMET BARDIZ, B.S.

Thesis

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

in

COMPUTER ENGINEERING

in the

GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

of

GALATASARAY UNIVERSITY

Jan 2019

This is to certify that the thesis entitled

TURKISH MEDICAL TEXT PARSING AND CLASSIFICATION

prepared by **Ahmet BARDIZ** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering** at the **Galatasaray University** is approved by the

Examining Committee:

Assist. Prof. İ. Burak PARLAK (Supervisor)
Department of Computer Engineering
Galatasaray University -------------------------

Assist. Prof. Murat AKIN
Department of Computer Engineering
Galatasaray University -------------------------

Assist. Prof. C. Okan ŞAKAR
Department of Computer Engineering
Bahçeşehir University -------------------------

Date: ------------------------

ACKNOWLEDGEMENTS

Firstly, I would like to thank my thesis advisor Assist. Prof. İ. Burak PARLAK of the Computer Engineering Department at Galatasaray University. Assist Prof. Parlak has devoted a wide range of time to me whenever I had a question about my research or writing even during the most intense periods of him.

In addition, I must express my gratitude to my wife and to my company partners for providing me with endless support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Dec 2018
Ahmet BARDIZ

TABLE OF CONTENTS

LIST OF SYMBOLS

NLP **:** Natural Language Processing
SVM **:** Support Vector Machines
NN **:** Neural Network
CNN **:** Convolutional Neural Networks
NB : Naïve Bayes
MNB **:** Multinomial Naïve Bayes
API **:** Application Programming Interface
HIMS : Hospital Information Management Systems
ROC : Receiver Operating Characteristic

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

This study includes approaches to identify the hospital departments with the methods of machine learning by referencing narratives of patients typed by expert physicians in the hospital information management systems. The main challenge is the preservation of semantic integrity while resolving the complex Turkish grammar rules on medical context. A Turkish medical parser was built and was characterized by specific abilities to extract medical entities and to act according to Turkish grammar. This parser identifies the syntactic, morphological, lexical and semantic features on the medical text. This parser basically discovers the misspellings, disassembles the possible morphemes of a term, assigns the optimal part of speech tags, gets rid of the morphemes that do not affect the semantic intensely and then decides to the best parse regarding the language and word2vec models. A medical corpus which includes patients' medical narrative records and hospital departments that corresponds to those narratives has become the basis for these studies. Additionally, a Turkish medical lexicon has been generated to extract medical entities. It consists of clinically favored terms that specify and differentiate diagnoses. These overall parsing steps provide to perform an exhaustive analysis of Turkish texts through the medical and agglutinative linguistic perspective. Secondly, the features extracted by this medical parser were used to predict hospital departments from patients' narrative records. The accuracies of the Multinomial Naïve Bayes, Support Vector Machines and Convolutional Neural Networks are evaluated to classify the medical content. The Support Vector Machines approach has acquired 98.16 % accuracy rate to classify hospital departments from medically parsed patients' narrative text.

ÖZET

Bu çalışma hastane yönetim sistemlerinde yer alan, uzman hekimler tarafından muayene edilen hastalara ait kayıtları referans alarak, hastaların şikayetleriyle ilgilenebilecek en muhtemel hastane branşını makine öğrenmesi yöntemleriyle tespit etmeyi kapsamaktadır. Fakat sınıflandırma aşamasından önce Türkçe dilinin yapısal ve anlamsal çözümlemesini yapmak gerekmektedir. Ayrıca bu çözümlemeyi medikal uzayda özelleştirmek de gerekir. Bu kapsamda Türkçe' nin sondan eklemeli bir dil olma özelliğini ve kendine özgü dil bilgisi kurallarını göz önünde bulundurmakla beraber, medikal sözlüklerle destekli bir şekilde metin ayrıştırma yöntemleri uygulanır. Bu çalışma kapsamında sınıflandırma performansını artırmak için Türkçe medikal kelime veya kelime gruplarını ayrıştıran ve semantik olarak anlamladıran bir medikal Türkçe doğal dil işleme servisi geliştirilmiştir. Bu servis üzerinde işlenip anlamlandırılan medikal metin verileri karar destek makineleri, çok terimli Naïve Bayes ve evrişimsel sinir ağları yöntemleri kullanılarak sınıflandırılmış ve yapılan testler sonucu en yüksek doğruluk oranı %98.16 olarak hesaplanmıştır. Ayrıca medikal Türkçe doğal dil işleme ve sınıflandırma katmanları paketlenerek, herhangi bir sistem veya web ortamına entegre olabilecek şekilde bir API olarak yayınlanmıştır. Sonuç olarak hastanenelerin online randevu sistemlerinde bu hizmetin kullanılmasıyla, hastaların şikayetleriyle örtüşen daha doğru ve özelleşmiş bir hastane branşına yönlendirilmesi hedeflenmiştir.

1. INTRODUCTION

Clinics usually store patient data on their information management systems as narrative text or medical reports. If these medical data are interpreted in a smart way, they can be used to utilize the health care services (Friedman, 2005). We are going to work on Turkish medical text to build a medical parser that morphologically and syntactically analyzes the text to extract medically valuable entities. However, Turkish has some difficulties to apply to natural language processing principles because of the complicated grammar rules and its being an agglutinative language that phrases mostly appear in inflected forms. The morphemes have more importance than other languages like English since they intensely carry semantic and syntactic information. The morphological disambiguation of free typed text in this kind of languages requires the application of specific language processing tasks. The medical terms or terminologies that are tried to extract on text can be disease or medicine names, medical procedures, medical devices or treatments, symptoms and period or status of related symptoms (Lekha, 2015). This medical parser implements natural language processing techniques tokenization, lemmatization, semantic tagging, syntactic parsing and morphological disambiguation (Zhou et al., 2011). It is developed in a modular structure and it has atomic NLP components to be expanded for new features (Eryiğit, 2014). This process is accomplished by applying Turkish grammar rules and supplying some medical vocabulary and other lexicons that contain labeled words from the biomedical domain (Szolovits, 2003).

Patients usually set their appointments from polyclinics by calling the hospital call center, using the online appointment applications through the website or mobile applications of the hospital or directly visiting it. The hospital department they would be appointed is defined by themselves, medical advisers or call center agents. However, they are sometimes appointed to hospital departments that cannot appropriately deal with their diseases or they individually choose the wrong department because of the lack of knowledge about the responsibilities of polyclinics. For instance, most of the patients who have problems with their stomachs get an appointment from the department of internal medicine. However, if they had visited the department of gastroenterology instead of internal medicine, they would have the opportunity to be examined by specialized physicians that have more knowledge and expertise on their problems. In addition, there are also lots of hospital departments especially in comprehensive hospitals or university hospitals. Most of the people have no idea about more specific departments like algology, andrology etc. It is not realistic to expect a non-medical person to distinguish the responsibilities of neurology and neurosurgery or physiotherapy and orthopedics. Patients or patients' relatives mostly make web search on medical forums to discover the abilities of departments to get an appointment. Nevertheless, the information on the web can be manipulative and unreliable so it can mislead them and cause to lose time and money. As a result, patients' getting an appointment from irrelevant hospital department causes a waste of time for both doctors and patients or reduces the quality of the received treatment. In order to improve this appointment process, it is planned to use patient narratives stored on hospital information management systems (HIMS) to create a predictive model that classifies new patients' symptoms and recommend to patients the most appropriate hospital department.

This study aims to develop a machine learning based medical assistant service for online hospital appointment systems to guide patients to the right hospital departments. In the development process of this automated service, related works have been examined in detail as first step. Then the technical approach which includes the steps of parsing the medical texts and training a machine learning model has been explained. Results of applied methods have been discussed. In conclusion part, future works to improve the available service and other application fields of developed framework have been expressed.

2. RELATED WORK

A wide range of studies is carried out in the medical field in terms of machine learning applications. It seems that the recent studies and researches to improve the medical service quality are especially covers the methods like medical image processing, pattern recognition on medical devices or sensors, virtual medical predictions from genomic background, Internet of Medical Things. These research and application fields basically require the processing of multimodal data by applying some statistical or mathematical approaches. However, medical text-based studies are not as widespread as the other modalities such as signal, image or sound based medical acquisitions. Text processing requires more localization than other multimodal data processing tools. For instance, the forms of cancer cells on a radiology image are unlikely to change in different parts of the world. Therefore, a satisfactory project related to an image use case is able to rapidly spread globally.

Although text-based medical applications or researches are not as popular as the image there are various valuable works in terms of medical text parsing or medical text classification. The MIDAS (Medical Diagnosis Assistant) project supported by natural language processing is an expert system that recommends diagnostic codes by referencing the radiology reports (Sotelsek-Margalef & Villena-Román, 2008). Naïve Bayes (NB) and Support Vector Machines (SVM) are experienced to diagnose heart disease for diabetic patients by Parthiban & Srivatsa (2012). The blood test values and personal attributes like weight, age and gender are the reference data to create a predictive model. Semantic analysis of free text in English natural language is also applied to assign clinical ICD-9-CM codes which is a standard of diagnosis representation in medical use. A dependency parsing based semantic analyzer is implemented through an automated diagnosis tool by Chen et al. (2010). Goldstein et al. (2007) similarly emphasize the significance of lexical elements as well as semantic information. It predicts the code of radiology reports and asserts the rule-based systems fed by lexical and semantic features which perform better than algorithmically more complex systems.

Due to the sensitivity and complexity of the healthcare services the practices are concentrated on some specific polyclinics or procedures. This approach minimizes the risk and increases the accuracy rate of medical classification use cases. Consequently, our approach requires more customization and comprehension that covers all polyclinics. We need to clearly identify the discriminative features among hospital departments. In this context, Ge (2015) states that clinical narrative texts are difficult to properly process because of their being ungrammatical, containing lots of acronyms, abbreviations, misspellings and local dialectal phrases. Moreover, a generic parser that handles the basics of natural language processing may not be successful enough in the medical domain since it may need to be fed by medical terminology. To resolve this deficiency Johnson (1999) has prepared a semantic lexicon that associates medical lexemes or phrases with semantic types which are organized ontologically. In a similar approach, to extract information from clinical reports Lekha (2015) identifies that in addition to the fundamental NLP processes like tokenization, part-of-speech-tagging and parsing, relationships and ontologies of medical domain-specific terms and entities must be well defined. Although, Latin medical terms are global, they generally have territorial discourses. Rubio-López et al. (2017) handle one of these custom problems in Spanish. While processing unstructured clinical narratives to perform the context-based search on them, they use machine learning techniques to disambiguate medical acronyms and to assign the most appropriate meaning.

Text classification process primarily requires applying natural language processing techniques to analyze the text and then acquire discriminative features. Kleverwal (2015) has designed a text classification-based system to help medical triage officers to make the right appointment according to the complaints of patients at the emergency call centers. Before the classification process, they have also applied the basic NLP techniques of morphological, lexical, syntactic, semantic and pragmatic analysis to retrieve the information from triage reports. In order to classify the medical reports, they prefer the algorithms of NB, SVM and K-Nearest Neighbors.

Since healthcare domain has low tolerance against technical errors, irrecoverable serious problems would require high costs to fix them. Therefore, the health industry needs to be strengthened by support decision systems for especially human-oriented processes. Liu (2007) emphasizes the significance of text classification in healthcare support systems in terms of their disease-related information like remedy, cause, side effect, symptom etc. In addition to domain knowledge and complicated processing the outcomes were obtained from healthcare support systems to prevent cancer and to classify it into suitable categories.

Our study covers the main titles of Turkish medical text parsing and classification. Parsing step applies the natural language processing techniques like tokenization, morphological, lexical, syntactical and semantical analysis. It customizes these methods by considering the agglutinative form of the Turkish language. Secondly, the features extracted by parsing the medical text are delivered to a classification module to be modeled. In the classification module, a supervised text classification approach which has the best test accuracy score is used. The whole process is published as a restful API to be integrated on the endpoint of hospital appointment systems. Finally, it recommends the most associated hospital departments to patients when they type their medical narratives to the online appointment system's search box.

3. PROPOSED APPROACH

Text parsing is a major step before the classification process to feed the predictive model with extracted features. The applying qualified medical text parsing makes the model more delicate and convenient in terms of specific and complicated cases. After developing a proper text parsing module, the next step is the application of machine learning with extracted features from medical content. The common text classification approaches multinomial naïve Bayes and support vector machines are utilized to predict the hospital department of a medical narrative and a convolutional neural network is also designed as an unconventional alternative.

3.1 Medical Corpus Overview

In this study, we have worked on the real-life data. The dataset has been acquired from Turkish university hospital. The medical narrative records have been collected through the clinics which usually store patient data on their information management systems. Moreover, these records have been extracted through the narrative features of patients who have been examined and then diagnosed by expert physicians. This dataset has the following properties below:

- 10 distinct hospital departments
    - Physical Therapy and Rehabilitation
    - Dermatology
    - Gastroenterology
    - Gynecology
    - Cardiology
    - Otorhinolaryngology

  - Orthopedics and Traumatology
  - Plastic Surgery
  - Psychiatry
  - Urology
- 10.000 documents (Patient Narratives) for each hospital department (100K in total)
- Each document has two properties. First one is the symptoms or complaints of patients and the second one is the medical departments assigned for them by the specialists.
- Number of total tokens: 1.504.859
- Number of distinct tokens: 55.969
- Average token length by characters: 7
- Average sentence length by tokens: 19

This medical corpus covers the data belonging to the most frequently visited hospital departments in terms of patient number. The general information about these departments in terms of main duties and diseases they investigate is presented in Table 3.1.

Table 3.1 Roles and responsibilities of hospital departments that exist on medical corpus

Hospital departments	Principle responsibilities of departments
Dermatology	Dermatologists treat disorders related to the skin and they also treat problems of the hair and nails. They basically treat diseases such as skin cancer, eczema, acne, severe drug rashes.
Physical Therapy and Rehabilitation	Physiotherapists assess, treat and advise patients with a wide range of medical conditions. Especially after surgery, physiotherapists raise body healing through therapies like exercise and manipulation.

	They also provide health education to staff and patients on how to do things more easily. Physiotherapy specialists work closely with orthopedists.
Gastroenterology	This department delivers medical treatment to patients who have disorders with their bowel, stomach and other parts of digestive systems. It covers nutritional services and endoscopy.
Gynecology	The department of gynecology investigates and treats problems of the reproductive organs and female urinary system. Main problems that gynecologists encounter are genital infections and disorders, ovarian cysts, pelvic pain etc.
Cardiology	It deals with the heart and circulatory system problems of patients. Cardiologists basically examine patients with conditions like chest pain, high blood pressure, high cholesterol, hole in the heart, diseases of arteries etc.
Otorhinolaryngology	It is also known as the department of ear, nose and throat. This department provides care for patients with a variety of problems in their ear, nose and throat. Some diseases in the responsibility of this department are hearing loss, ear infections or noise, balance or smell disorders, nose injuries, allergy, cough, sore throat.
Plastic Surgery	A plastic surgeon's primary responsibility is repair, reconstruction, or replacement of physical defects of form or function involving the skin, musculoskeletal system, extremities, breast and external genitalia or cosmetic enhancement to visible areas of the body. Cosmetic plastic surgeons commonly perform breast augmentations and reductions, liposuction, and facelifts.

Orthopedics and Traumatology	It treats problems that affect patients' muscular and skeletal system. These components can be muscles, joints, bones, ligaments, tendons and nerves. Orthopedic trauma includes dislocations and fractures as well as musculoskeletal injuries to soft tissues. Tennis elbow, carpal tunnel syndrome, ligament injuries to the knee and torn meniscus are the most known diseases of this department.
Psychiatry	It is the branch of medicine which deals with the origin, diagnosis, prevention and management of mental disorders, emotional and behavioral disturbances.
Urology	It is the department that investigates all kinds of conditions related to kidney and bladder-based problems. The basic diseases of urology are prostate, anomalies in urine, overactive bladder etc.

There are some sample narratives in Table 3.2. These narratives on corpus are typed by physicians. It is mandatory to be kept a record of every patient by physicians. When they question and examine the patients, they also type complaints of patients and generally perform it very quickly. Therefore, the records are written in the language of physicians, informal, do not appear in sentence form, contain abbreviations, misspellings, ambiguities and Latin terms for Turkish synonyms widely. A machine learning model trained with this raw data can be satisfactory if it just recommends to physicians not to patients. Our approach suggests the most likely hospital departments to a patient when they type their complaints.

Table 3.2 Sample of patient narratives

Patient narratives	Hospital departments
çarpma sonucunda sol el bileğinde ağrı	Orthopedics and Traumatology
yüzde sutür daha önce yüzünde kesisi bulunan hasta	Plastic Surgery
saçlarda dökülme	Dermatology
sık idrara çıkma	Urology
20 ağustosta leke şeklinde kanaması olmuş	Gynecology
her iki dizde ağrı eklem ağrısı	Orthopedics and Traumatology

The development of a text-based recommendation service requires to match physicians' knowledge with patients' statements about their troubles. For example, if a patient defines his/her problem as "bacağımda çekme var", it means "siyatik sinir sıkışması" for a physician. To match these different expressions of the same symptoms we have created a medical lexicon to enrich the medical corpus with various discourses of both patients and physicians. Some different types of samples are listed in Table 3.3.

Table 3.3 The examples of non-standard forms

Statement	Type
apandist / apandisit	Misspelling
tibia / kaval kemiği	Latin/Turkish Variation
kostovertebral açı hassasiyeti / kvah	Physician Abbreviation
Kurdeşen / Ürtiker	Spoken Language

It is realized from the corpus the commonly used medical statements can include multiple words. The different numbers of word occurrence in a phrase are represented with the term n-gram. If n equals to 1 is called as unigram, 2 is bigram, 3 is trigram. It can be extended to any number. Consequently, the evaluation of text data by considering the n-gram forms increases the quality of analysis.

Some samples of the most frequent words or phrases for each hospital department are listed in Table 3.4. Their number of occurrences are also denoted as adjacent to them. Some terms can appear in multiple departments. For instance, the word "ağrı" which means ache repeats too many times in almost every department. This case leads to classification ambiguity while classifying the departments that narratives refer to.

Therefore, it is not a rational approach to decide just by matching the terms and departments. The relationships among the medical terms in each hospital department, the sequence of them in sentences and n-gram forms of them are taken into consideration to resolve this ambiguity.

Table 3.4 Most frequent tokens compared to hospital departments

Most frequent words or n-grams	Hospital departments
kaşıntı (1424), kızarıklık (734), kriyoterapi (558), yüzde sivilce (677), saç dökülmesi (524)	Dermatology
ağrı (17701), lomber (2334), bel ağrısı (2082), eklem (1299), boyun ağrısı (665)	Physical Therapy and Rehabilitation
epigastrium (1901), karın üst kısımda ağrı (1425), karın ağrısı (610), şişkinlik (478), kabızlık (434)	Gastroenterology
uterus (5939), gebelik (5030), vulva vajen (2551), cesamette (2359), plasenta (1875)	Gynecology

ekg (4869), göğüs ağrısı (2282), üfürüm (1318), çarpıntı (1265), nefes darlığı (1051)	Cardiology
timpanik membran (4285), dış kulak yolu (4271), preauriküler (4261), nazal mukoza (2393), orofarenks (2230)	Otorhinolaryngology
pansuman (2648), elde kesi (1871), ayakta yara (1346), nevüs (1324), yüzde kesi (705)	Plastic Surgery
ağrı (7980), düşme (1279), diz ağrısı (1037), bilek ağrısı (848), kalça (461)	Orthopedics and Traumatology
efexor (1457), buspon (1418), sinirlilik (1214), uykusuzluk (471), unutkanlık (456)	Psychiatry
yan ağrısı (2448), sık idrara çıkma (1190), kvah (1454), idrar kaçırma (495), taş düşürme (379)	Urology

3.2 Methodology

Two main processes have been followed to implement a hospital department recommendation service. The first step is designing a medical text parser that applies also Turkish grammar rules using the following order; morphology, syntax, lexicology and semantics. The second step is discovering the best approach to fit a model with a parsed medical dataset and to make a prediction with high accuracy. Finally, a framework has been established to develop medical text-based recommendation systems.

3.3 Medical Text Parsing

3.3.1 Preprocessing

First of all, the raw clinical text is split into sentences by taking into consideration the punctuations like “.”, “!”, “?” as the last character of a sentence. The dot mark for the abbreviations is excluded. After that, each sentence is tokenized separately. When the whitespace, double quote or comma is treated as a delimiter pipe, single quote or dot are not interpreted as a delimiter because of their semantic significance for Turkish morphology and medical context. The sequence between tokenized words is always preserved by establishing a linked list of tokens.

3.3.2 Morphological Analysis

The agglutinative form of the Turkish language makes the morphological analysis of an inflected word or word groups challenging. To explore possible word structures morphological analyzer of the Zemberek[1] open source NLP framework for Turkic languages is used (Akın & Akın, 2007). It is satisfactory in terms of extracting all variants of all morphological forms of an inflected term. It presents all possible analyzed forms of a word as roots, lemmas and morphemes. However, the syntax of the analyzed words is required to be taken into account. Token analysis is performed to find the most appropriate form within the context of a sentence. When some morphemes like possessive suffix, past tense suffix or plural suffix are omitted to lemmatize some others medically matter like privative suffix (-sız,-siz) or negative suffix (-ma, -me) are preserved not to lose the semantic sensitivity.

[1] https://github.com/ahmetaa/zemberek-nlp

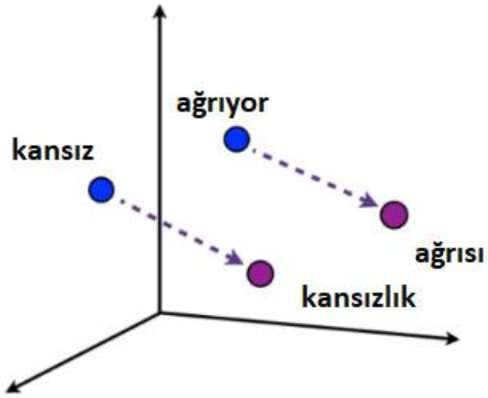

Figure 3.1 Relationships of inflected forms in word2vec space

In addition, word sequences also affect the lemmatization process to decide which lemma must be kept. To perform this approach a word2vec model is generated to associate the word embeddings and word morphology (Cao & Rei, 2016). It simply creates a medical lexical database that let to query the familiarity score of two separate words. We have also built a word2vec model by modeling the whole medical corpus. The agglutinative forms that drive from the same root are ensembled closely as shown in Figure 3.1. After detecting the proper POS-tagging of a sentence's words by applying the POS-tagger of Zemberek, word2vec layer provides to determine the most possible lemmas of those morphological forms of words. For instance, if a sentence "kansızlık yaşıyorum" which means "I am having anemia" is resolved, "kansızlık" is tagged as Noun and "yaşıyorum" is tagged as Verb by Zemberek. However, the inflected word "kansızlık" have distinct lemma forms that each one has also different POS-tagging like "kan": Noun, "kansız": Adjective and "kansızlık": Noun. To identify the most appropriate lemma in Noun form, word2vec model is activated to evaluate the relationship of noun lemmas of "kansızlık" with the verb "yaşıyorum". Finally, it selects the most pervasive couples and it repeats this process for each next token sequentially. Thus, the contextual proximity among possible inflected forms of two words is revealed rationally. Some samples are listed in Table 3.5.

Table 3.5 Customized lemmatization for medical context

Inflected form	Stem	Lemma
kansızlığımdan	kan	kansızlık
halsizliktendir	hal	halsizlik
kanaması	kan	kanama
kanımın	kan	kan
kanamamasından	kan	kanamamak

Moreover, misspelled words are corrected before applying the morphological analysis. Misspellings are detected by using a unigram language model created on Zemberek framework. It just tolerates one manipulated character of the correct form of a word. Additionally, a Turkish deasciifier is used to normalize the Turkish characters that do not exist on the Latin alphabet. It analyzes the characters of a single word and converts the characters typed in ASCII form to the appropriate Turkish letters like ş, ı, ö, ç, ğ, ü. It simply prefers the most occurred sequences of the characters in a word by referencing a Turkish pattern table which stores the occurrences of possible character pairs.

3.3.3 Syntactical Analysis

Assigning part of speech (POS) tags are simpler in Indo-European languages rather than Turkic languages that words can have excessively agglutinative forms. Therefore, it is not appropriate for Turkish to express with the limited POS tags exist in the Penn

Treebank (Marcus et al., 1993). Applying the finite tagset approach to the languages like Turkish may cause to lose information (Hakkani-Tür et al., 2002). If the word “iyileştirmek” which means to cure is morphologically analyzed each suffix manipulates POS tagging category as shown below.

iyi + leş + tir + mek

[(iyi) (Adjective)

(Verb; Become: leş)

(Verb; Causative: tir)

(Noun; Infinitive1: mek)]

In the morphosyntactic approach, all morphological features of a word are considered. Therefore, the compound form is analyzed instead of just considering POS tag of the root form. To eliminate the disambiguation and determine the appropriate POS tag the Zemberek framework applies the perceptron algorithm of Sak et. al (2007). In this approach a statistical trigram-based model to enumerate the top best list of candidate morphological parse sequences for each sentence is referenced. Then, the perceptron algorithm is applied to rank the top best list by caring about the possible morphological characteristics of related trigram sequences.

3.3.4 Lexical Analysis

Medical lexicons are generated by specialists or they are extracted from expert systems. After lemmatization of the words with the appropriate form normalized tokens or n-grams are tagged by referencing a medical lexicon that covers medically valuable entities like symptoms, medicines, devices, procedures, examinations, treatments etc. On the other hand, the lexicon is also useful for detection of synonyms, variants or medical statements in folk speech. They are detected, normalized and converted to the most common form.

Consequently, hidden medical terms in the text are uncovered to be interpreted as semantically afterward. The medical lexicon can be elaborated on synonyms, genders, diagnosis or more specific medical concepts. Our lexicon currently includes 10 the entity types which are BODY_PART, MEDICAL_ABBREVIATION, MEDICAL_EXAMINATION, MEDICINE, ORGAN, PERIOD, PROCEDURE, SPECIFIC_SYMPTOM, STATUS and SYMPTOM. Sample values for these entity types are listed in Table 3.6.

Table 3.6 Referenced Turkish medical lexicon

Entity	Variant	Type
dış kulak yolu		BODY_PART
timpanik membran	kulak zarı	BODY_PART
batın	karın	BODY_PART
g2p1	2 gebelik 1 doğum	MEDICAL_ABBREVIATION
kan testi		MEDICAL_EXAMINATION
paxil 20 mg		MEDICINE
paxera 30 mg		MEDICINE
haftada bir gün		PERIOD
burun tıkanıklığı		SPECIFIC_SYMPTOM
plasenta posterior		SYMPTOM
kolonoskopi		PROCEDURE

3.3.5 Semantic Analysis

Medical text parser can gain more specific abilities by evaluating the attributes acquired from morphological and lexical analysis semantically. In this regard, some rule-based logics are deployed on the parser. For instance, if an organ is followed by a symptom, a new tag type SPECIFIC_SYMPTOM is automatically generated and associated tokens are grouped under this tag. If the specified rules are valid the parser can automatically create compound words recursively. Figure 3.2 indicates an example of the auto-created hierarchical token group.

Figure 3.2 Automatic tag type generation

Besides, a medical corpus that used for optimum lemmatization is also referenced for the association of medical entities with related hospital departments. However, some ambiguous situation emerges when some entities are related with more than one hospital department. Some ambiguous cases are illustrated in Table 3.7 and the percentage rates of entities that causes ambiguity are also denoted by considering their distribution in hospital departments. To overcome this situation a relevancy score is assigned for each related department by frequency analysis.

Table 3.7 Ambiguous cases

Entity	Type	Hospital departments
sık idrara çıkma	SYMPTOM	Urology (96%) Gynecology (3%)
nefes darlığı	SYMPTOM	Cardiology (83%) Psychiatry (13%) Physical Therapy and Rehabilitation (1%) Gastroenterology (1%)
sırt ağrısı	SYMPTOM	Cardiology (22%) Physical Therapy and Rehabilitation (69%) Gastroenterology (3%) Orthopedics and Traumatology (5%)
kolonoskopi	PROCEDURE	Gastroenterology (95%) Urology (3%)

Furthermore, lexical collocation analysis also semantically strengthens the medical parser. McKeown & Radev (2000) states that “A collocation is a group of words that occur together more often than by chance”. This medical parser especially focuses on noun + verb constructions. For example, “tahlil + yaptırmak”, “kas + yırtılmak”, “reçete + yazmak” etc. Collocations are detected on the medical corpus by applying frequency analysis of nouns that are neighbor to verbs. Lexical collocation analysis is beneficial to decrease the morphological disambiguation in terms of determining the possible lemma commonly used with a related verb.

3.3.6 Publishing the Text Parsing API

The main parsing steps; how to treat a sample sentence and to analyze a sentence with the related functions; tokenizer, deasciifier, spell-checker, lemmatizer, POS-tagger and entity extractor are illustrated in Figure 3.3.

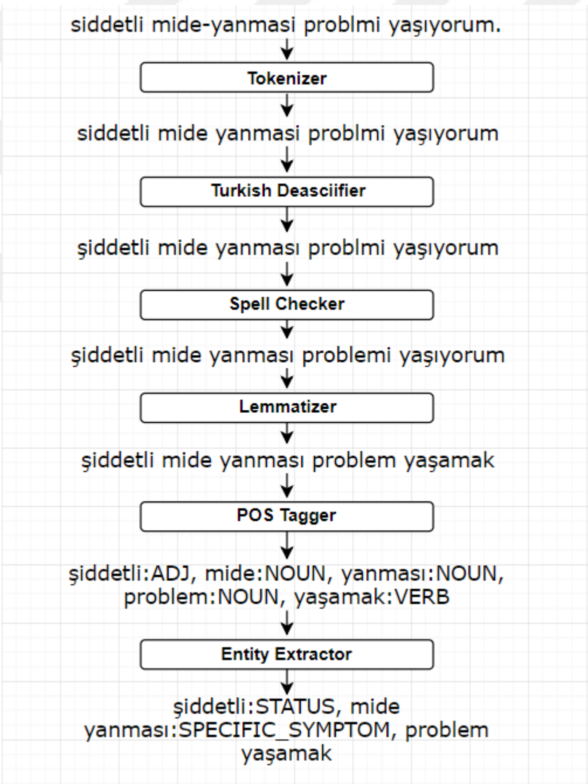

Figure 3.3 Main parsing steps

The Turkish medical text parsing API is developed on Java 8+. The followed methods that the API applies to the user request are displayed in Figure 3.4. The basic reason to choose Java for the development of our service is the benefits of its architecture for a backend project. Since it can scale to distributed systems and multi-processors, it is suitable and reliable for cross-platform usage, it is successful on memory management, it well handles the multi-threading, it introduces security advantages and also it has a wide ecosystem.

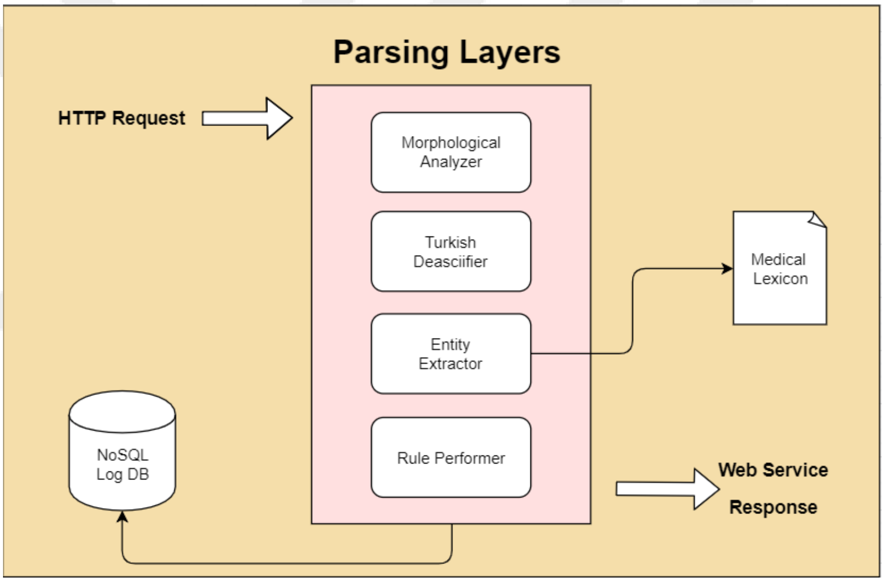

Figure 3.4 Rest-API Architecture

After applying all these approaches to the free text input it returns a response in the JSON format. A sample JSON output of the parser service is shown in Figure 3.5 below. In this output, for the supplied free text input misspelling is corrected, one of the tokens "kafa" is replaced with its synonym "baş" which is used more commonly, the entity type SPECIFIC_SYMPTOM is generated automatically because of getting together of entities ORGAN and SYMPTOM. It is lemmatized and normalized by preserving the important morphemes and eliminating the useless suffixes. The Figure 3.5 displays the summary which is just one tagged compound word, but each token is also recursively processed in the background.


```
"tokens":
{
"token": "şiddetli kafa ağrsından",
"tokenType": "SPECIFIC_SYMPTOM",
"normalize": "şiddetli baş ağrısı",
"meta": {
        "synonym": true,
        "fuzzy": true,
 }
```


Figure 3.5 Turkish medical text parser API sample output

3.4 Text Classification

Assigning free-text documents into a fixed number of predefined categories is characterized as text classification. Classifiers are the machine learning or deep learning models that can assign labels to the content automatically. Classifiers need to be trained with labeled data to be able to make specific predictions. A classifier decides more accurate when it is trained to serve for a specific use case. Some use cases that text classifiers benefit are spam filtering, sentiment analysis, language detection and categorization of various kind of text-based sources like news, customer emails, social media posts etc.

When the medical parsing service is called on the sample data, cleaned and valuable tokens that tagged medically would be acquired. Thus, a classification technique which runs reliably with text data would be sufficient to make polyclinic prediction of some specific symptoms. For this purpose, the multiclass architecture is modeled by training it with an extracted part of the sampled HIMS data and by testing it with already trained model through the remnant part. In addition to the SVM and MNB which Rennie & Rifkin (2001) defines them as very strong supervised learning methods for text classification, the multichannel Convolutional Neural Networks (CNN) would be used to train the multiclass predictive models. Finally, test performance of them has been evaluated to decide which one has better results.

The medical dataset has been split into two parts as training and test. 80% of the 100K documents were put aside for training and rest of them 20% of the whole data were stored for testing. Test dataset is absolutely unseen during training of a classifier.

3.4.1 Technical Requirements

We have preferred to use Python while applying the text processing approaches. Python is very powerful in terms of specific open source packages for NLP and easy to implement while preprocessing of text data, executing classification techniques and measuring the performance of designed classification models. To import, analyze and preprocess the data, the packages of Numpy and Pandas are very advantageous. Scikit-learn is one of the most popular machine learning packages. Python also supports the Tensorflow framework which is designed to process numerical computations with high performance. Therefore, it is very successful to handle complex computations of deep learning architecture. Finally, Matplotlib is used to visualize the data distributions and classification performance metrics.

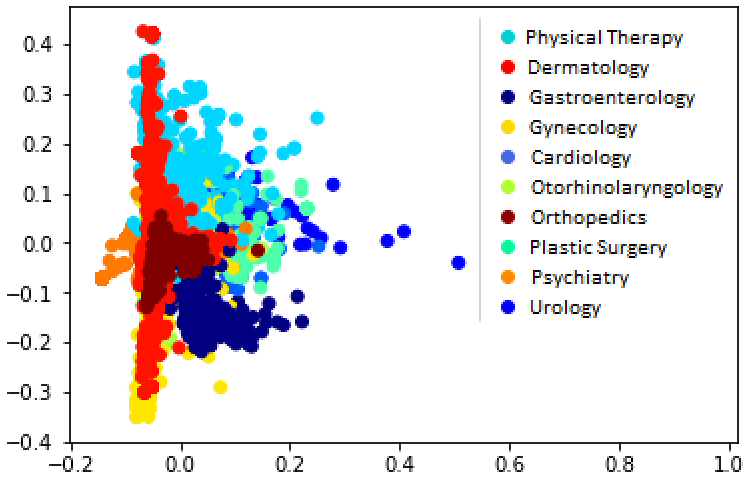

Figure 3.6 Projections of the vectorized text documents on a 2D plane

To operate the classification algorithms, we need to vectorize raw text documents. Figure 3.6 indicates the document vectors that belong to distinct classes. They are converted from textual content to mathematical representatives by using tf-idf vectorizer that converts text documents into matrix of tf-idf features. Tf-idf points to term frequency - inverse document frequency. It is basically the measurement of how important a word in a text document. Tf-idf of a term t can be simply calculated by taking the product of tf and idf. They are also calculated with the following formulas.

tf(t) = (# of times t appears in a document) / (Total # of terms in the document) (3.1)

idf(t) = log_e(total # of documents / # of documents contains t) (3.2)

3.4.2 Support Vector Machines Classifier

SVM is one of the strongest methodologies that commonly used for text categorization (Joachims, 1998). In this study as a first approach, we are going to use the SVM to classify medical text data and then to predict the class of test data through the trained model.

At the beginning step of creating a predictive model on SVM, the documents need to be vectorized to be represented on a hyperspace. Then the set of vectors which deliminates the same vector sets as shown in Figure 3.7 are grouped by the hyperplanes. Thus, each class possesses an area stored with related documents. While predicting the class of a new document it is firstly vectorized and then the area is discovered on the plane it belongs to.

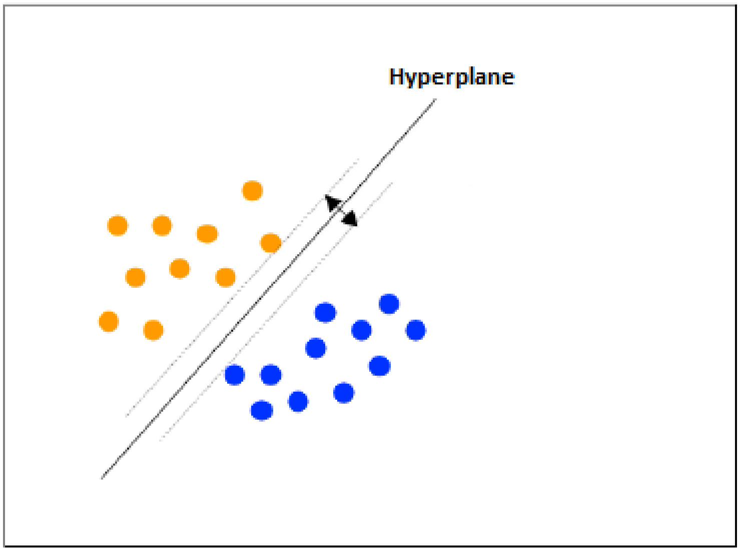

Figure 3.7 SVM hyperplane (Caragea et al., 2005)

Before the application of SVM classification, the training dataset is processed thorough custom medical parser to tag medical and valuable tokens or token groups. Then these tagged text data are classified by SVM and a model is created to predict the classes of documents exist in the test dataset. This SVM classification study is implemented with SVM linear kernel in the Scikit-learn machine learning library (Pedregosa et al., 2011). It basically applies the following functions.

Given training vectors $x_i \in \mathbb{R}^p$, i=1,..,n, in two classes, and a vector $y \in \{1, -1\}^n$, the following dual problem is solved by SVC:

$$\min_{\alpha} \frac{1}{2}\alpha^T Q \alpha - e^T \alpha$$
$$\text{subject to } y^T \alpha = 0$$
$$0 \leq \alpha_i \leq C, i = 1, ..., n \qquad (3.3)$$

where $C > 0$ is the upper bound, e is the vector of all ones, Q is an n by n positive semidefinite matrix, $Q_{ij} \equiv y_i y_j K(x_i, x_j)$, where $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ is the kernel. The function ϕ implicitly maps training vectors into a higher dimensional space.

The decision function is:

$$\operatorname{sgn}(\sum_{i=1}^{n} y_i \alpha_i K(x_i, x) + \rho)$$

(Cortes & Vapnik, 1995) (3.4)

Except the linear one SVM can be applied with sigmoid, polynomial and radial basis function kernels.

3.4.3 Multinomial Naïve Bayes Classifier

Naive Bayes classifiers are used in different fields due to its robustness, its accuracy, its low algorithmic complexity and its speed. Kazmierska & Malicki (2008) studies the applications of NB to optimize the treatment decisions for assessment of progression after radiotherapy and individual cancer risk. The multinomial NB is preferred when the multiple occurrences of the words matter in text classification (McCallum & Nigam, 1998). In this part of the study, we are going to classify the sample test data after creating a Bayesian model with the training data. It basically applies the following function.

The vectors $\theta_y = (\theta_{y1}, \ldots, \theta_{yn})$ parametrizes the distribution for each class y, where n is the number of features (the size of the vocabulary in text classification) and θ_{yi} is the probability $P(x_i \mid y)$ of feature i appearing in a sample belonging to class y.

A smoothed version of maximum likelihood estimates the parameters θ_y , i.e. relative frequency counting:

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n} \tag{3.5}$$

$$N_y = \sum_{i=1}^{|T|} N_{yi} \tag{3.6}$$

where $N_{yi} = \sum_{x \in T} x_i$ is the number of times feature i appears in a sample of class y in the training set T, and is the total count of all features for class y.

The smoothing priors $\alpha \geq 0$ justify features not present in the learning samples and prevent zero probabilities in further computations (Pedregosa & Varoquaux, 2010). The other Naïve Bayes methods Gaussian and Bernoulli can also be experimented to compare with the Multinomial one.

3.4.4 Convolutional Neural Networks

A high-level neural networks API Keras were used to build a neural networks classifier. Keras runs on top of popular deep learning frameworks Tensorflow and Theano. We have preferred CNN of Tensorflow in our experiment. Unlike the standard neural network architecture, instead of parameter optimization in the fully connected layers, the convolutional layer provides to process on parameters belong some fractions (Goodfellow et al., 2016). Common CNN architecture is visualized in Figure 3.8.

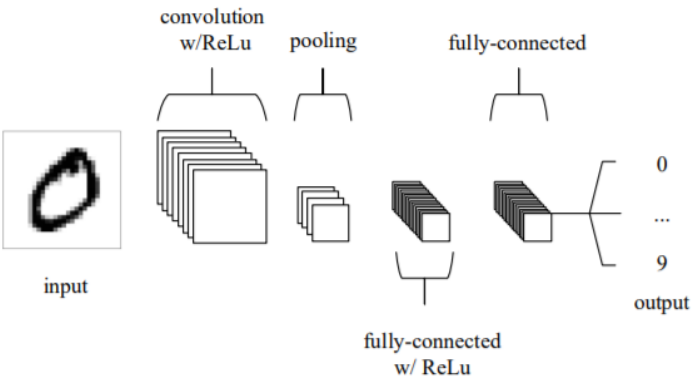

Figure 3.8 CNN architecture basics (O'Shea & Nash, 2015)

CNN is usually applied for image classification scenarios. In CNN architecture, convolutions basically treat as filters that are specialized to extract features on each applied layer. Therefore, the key features like corners, shadows, edges etc. that characterize an image can be decomposed by convolutions. Recently, CNN is preferred for NLP tasks like sentiment analysis, question answering and machine translation. We have used CNN for text classification from a different perspective. Unlike the bag of words approach in SVM and MNB the sequence of the words is considered by applying the approach of n-gram multichannel convolutional neural network (Kim, 2014). This method basically creates parallel convolutional neural networks that each one processes words in different window intervals. Figure 3.9 summarizes the channel network structure.

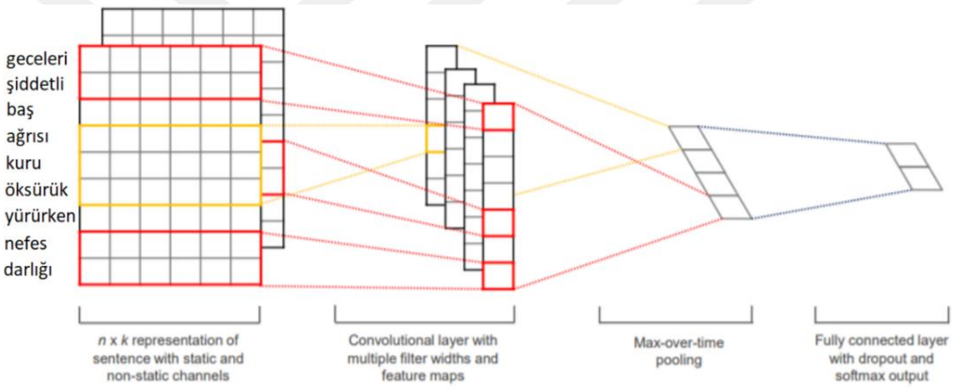

Figure 3.9 Two channels CNN model architecture (Kim, 2014)

We have used four channels respectively denote 4-grams, 6-grams, 8-grams and 10-grams. Each kernel acts as an independent network and the prediction probabilities of each channel is argmaxed to obtain the most probable output. Each channel has the same network architecture but except the kernel sizes. Different kernel sizes provide to investigate the relationship of tokens or phrases in different window sizes. Figure 3.10 displays the general architecture of our network.

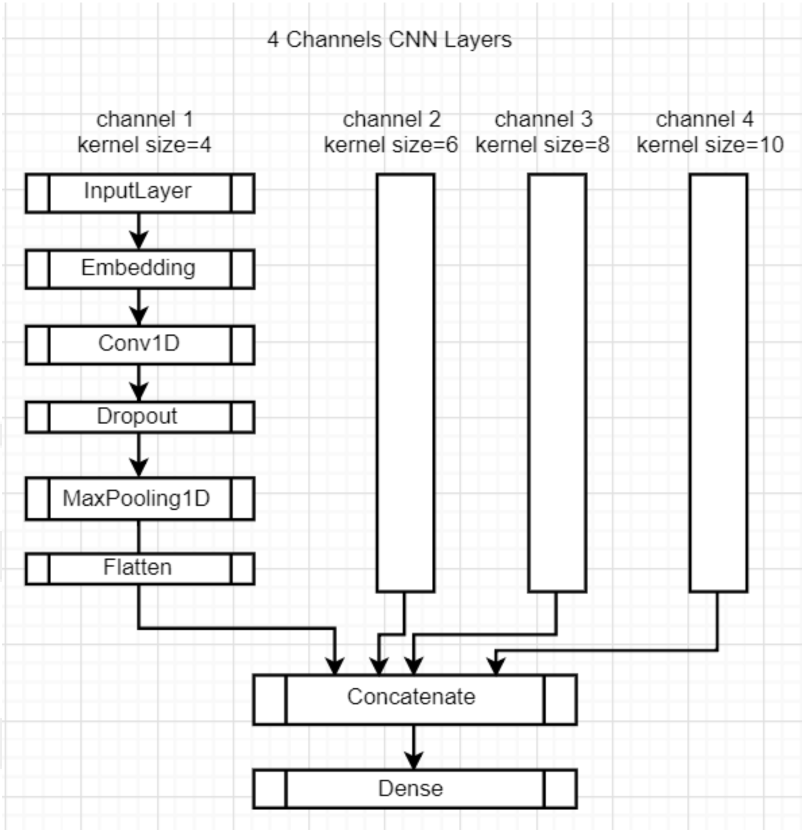

Figure 3.10 Four channels CNN architecture

While building our network we have first used an input layer. It is used to instantiate a tensor which is a term for the unit of data. Secondly, an embedding layer is appended to both encode the words in efficient way and to extract the feature of relations among words. Then, Conv1D layer performs the magic of convolutional approach that creates filters to emphasizes the distinct features of input data. Specific weights of neurons can make the model too specialized to the training data. Dropout layer makes the model more generalized and prevents the overfitting. Dropout is a regularization technique where randomly selected neurons are dismissed during the training process. Then, the model is simplified, and the sharpest features are extracted with the maxPooling1D layer. The Flatten layer reduces the tree-dimensional output to two dimensions before the concatenation process. The concatenate layer merges output of each channel into one vector. The dense layer where the matrix vector multiplication is performed. The values in the matrix are trainable parameters and are updated during backpropagation step. These parameters are updated according to the activation function.

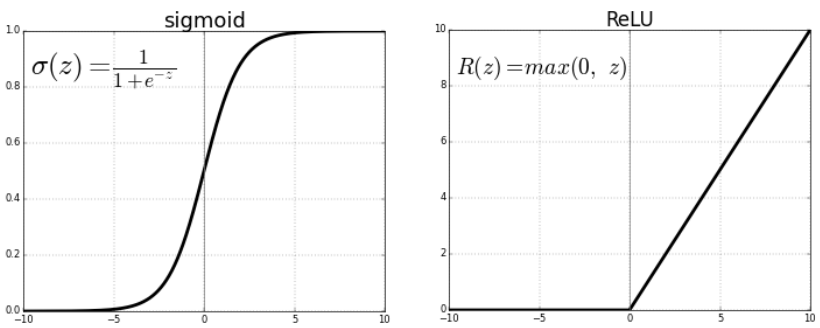

Figure 3.11 The comparison of sigmoid and RELU

There are different kinds of activation functions which have different sensitivity like Tanh, Sigmoid, Relu etc. These functions give the strength of non-linear computation to neural networks. We have preferred to use Relu because of its becoming much simpler computationally and its having wider input range to be activated unlike sigmoid. These facts are clearly visible in Figure 3.11. RELU just picks the value greater than 0 but sigmoid performs expensive exponential operations.

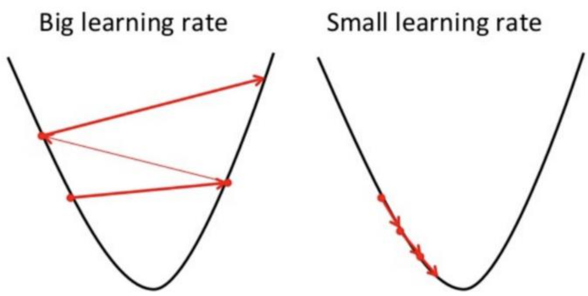

Figure 3.12 The impact of learning rate to reach the local minima of a function

Optimizing the parameters is the key process to run designed neural networks efficiently. To measure the performance of designed network batch size and epoch are the most focused parameters. To optimize these parameters, it is important to know how a neural network run. Therefore, we need to focus on the most popular optimization algorithm which gradient descent is. The neural networks models learn by minimizing the cost function iteratively. At this stage gradient descent takes a role. It can be considered as climbing down to the bottom of a valley. Figure 3.12 displays the difference between big and small learning rate which is influenced by batch size. If it becomes too small, it iterates slowly, and it can be stuck in local minimums, so it cannot move to the deeper valleys of the mountain. However, if it becomes too big it is going to have bigger steps, so it is going to be faster and more likely to find the deeper valley but less sensitive to find the deepest point.

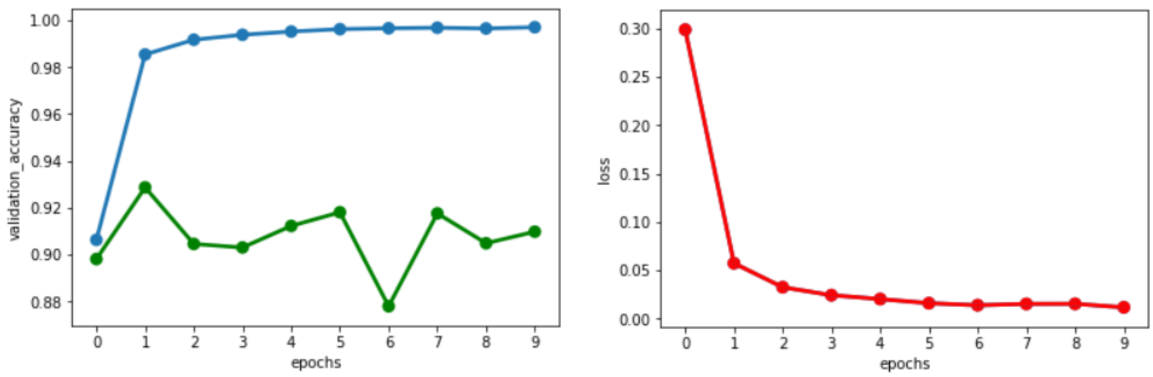

Figure 3.13 Train, validation and loss accuracy changes against each epoch

While running our CNN model we have chosen the batch size as 32. The entire dataset is passed forward and backward through the network by considering the batch size. Each pass of entire dataset is named as the epoch. We have assigned 10 to epoch parameter. Accuracy and loss rates of our optimized CNN architecture are displayed in Figure 3.13. The blue line indicates the training accuracy rate by each corresponding epoch. The green line is the validation accuracy rate which is calculated by random allocation of a validation dataset for each epoch. Finally, the red line represents not the rate, just the value of loss function which calculates the error rates on training and validation dataset. The gradual reduction of the loss function confirms the successful design of our model.

3.4.5 Hospital Department Recommendation API

After approving the success of SVM the model has been created is published as a service in order to direct the patients to the right medical department. It can serve patients directly through integration with the online appointment process of hospitals or it can be integrated with call center processes to guide the agents to recommend the medical departments for the supplied symptoms. To properly match the typed user's symptoms with the model, typed symptoms are also needed to be parsed through the medical parser. The data flow of the classification service is shown in Figure 3.14

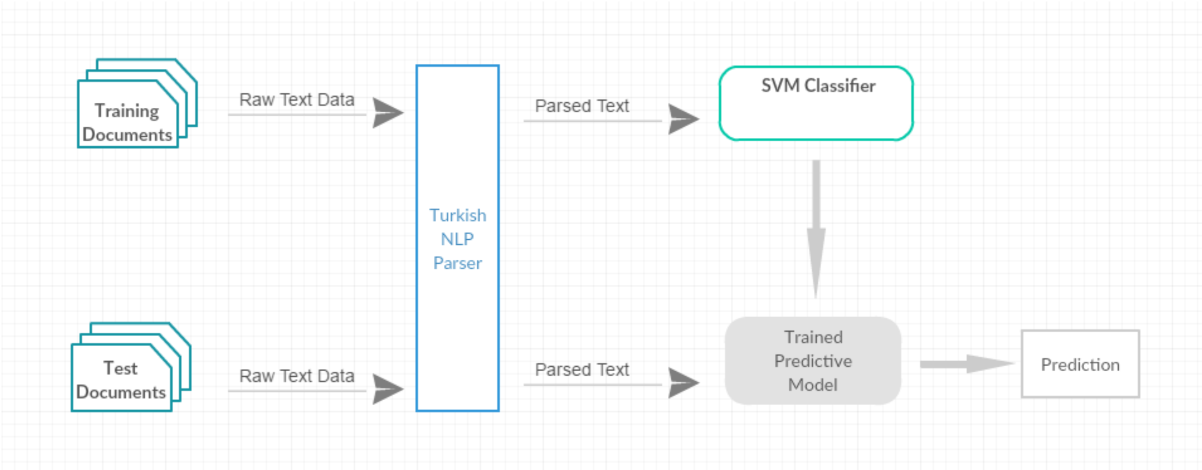

Figure 3.14 The predictive API service architecture

4. RESULTS

4.1 Support Vector Machines Test Results

During the training and test process, the polyclinics assigned for the symptoms on the sample data are accepted as true. When the polyclinics exist on the raw data are compared with predictions of SVM classifier for the related symptoms, the success rate of the SVM is measured as 98.16%. This seems a very high rate. It is probably because of typing both training and testing data by medical specialists. If the test data is typed by patients the success rate would decrease undoubtedly.

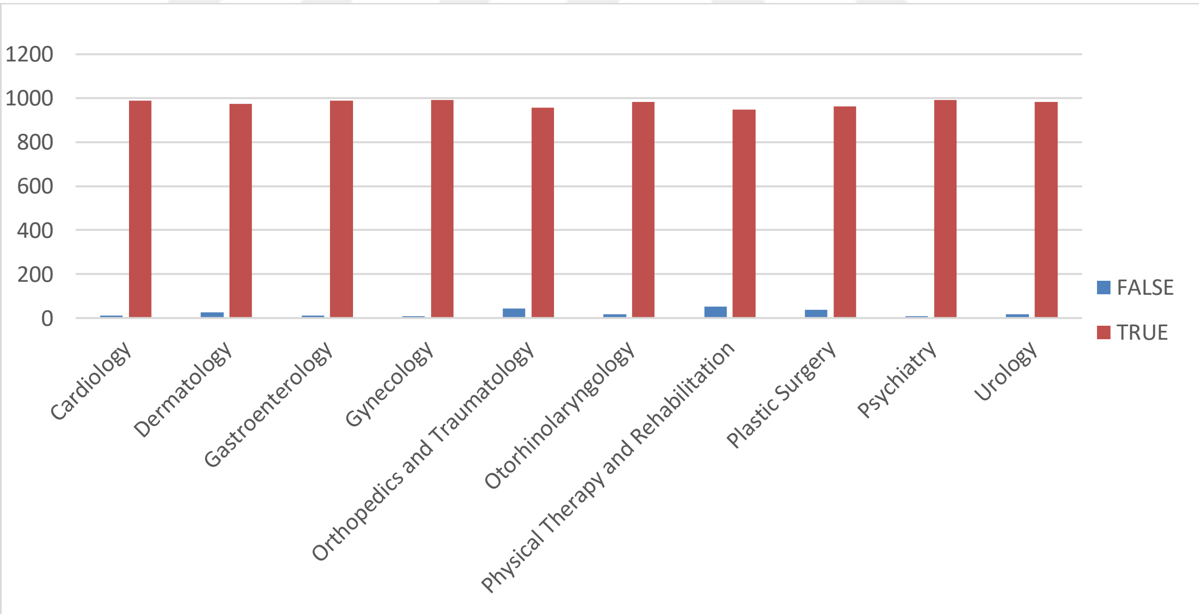

Figure 4.1 Test results of the medical SVM classifier

The distribution counts of the test result as true and false are shown in Figure 4.1 above. If we pay attention to SVM classifier's most mispredicted two classes, they are "Orthopedics and Traumatology" and "Physical Therapy and Rehabilitation". It can be predicted in advance since specialties of both these polyclinics are very close to each other.

Table 4.1 Accuracy rates of different SVM kernels

SVM options	Accuracy rate (%)
Radial basis function (rbf)	91.11
Polynomial	90.10
Linear	97.90
Linear and n-gram	98.16

There are different SVM kernel types. Kernels are the main mathematical functions that SVM uses to separate hyperplanes. Some of them are linear, radial basis function (RBF), sigmoid and polynomial. When we compared the most popular SVM kernels that linear, RBF and polynomial, the best performance in terms of accuracy and speed belonged to the linear one. Tests have been performed with the same training and test datasets for all kernels. The accuracy rate of RBF is measured as 91.11%, the polynomial is 90.10% and linear is 97.90 % as shown in Table 4.1. In addition, when the linear kernel was used with n-grams the highest accuracy was obtained with the bigrams and measured as 98.16 %. If unigram and bigram accuracy rates are compared, bigram increases the accuracy by 0.26 %. N-grams are interpreted as a single token and they expand the corpus. For instance, a phrase like "baş ağrısı" is tokenized as "baş, baş ağrısı, ağrısı". On the other hand, when we increase the n-gram range, the duration of building the predictive model has also lasted longer.

4.2 Multinomial Naïve Bayes Test Results

When the training and test datasets are used the success rate of Multinomial Naïve Bayes (MNB) is measured as 96.7 %. There is just one percent difference between SVM and MNB classifiers. Moreover, false predictions of SVM are distributed more balanced but the false predictions of MNB are mostly concentrated on the class of "Orthopedics and Traumatology". Distribution of all classes is indicated in Figure 4.2.

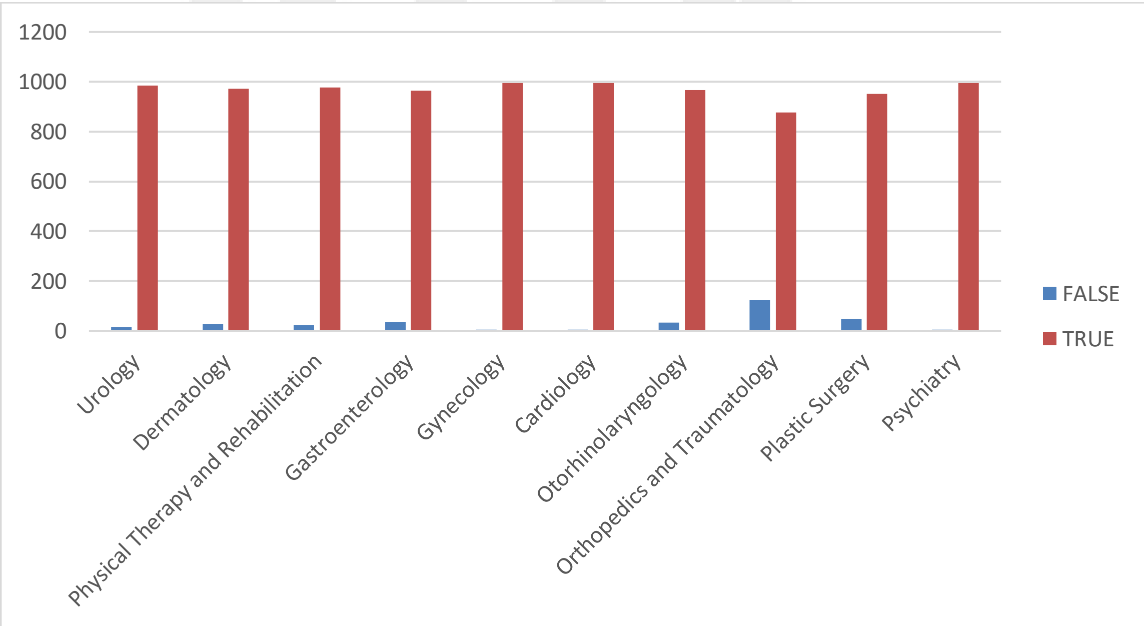

Figure 4.2 Test results of medical MNB classifier

According to the test results, SVM seems a little bit more powerful than MNB. The obtained performance metrics of SVM and MNB support the statement of Tong & Koller (2001) which is "The Naive Bayes classifier provides an interpretable model and principled ways to incorporate prior knowledge and data with missing values. However, it typically does not perform as well as discriminative methods such as SVMs, particularly in the text classification domain"

4.3 Convolutional Neural Networks Test Results

When the training and test datasets are used the accuracy rate of Neural Networks (NN) is measured as 94.6 %. It is the least satisfactory one compared to SVM and MNB classifiers. The accuracies for all classes can be analyzed in Figure 4.3.

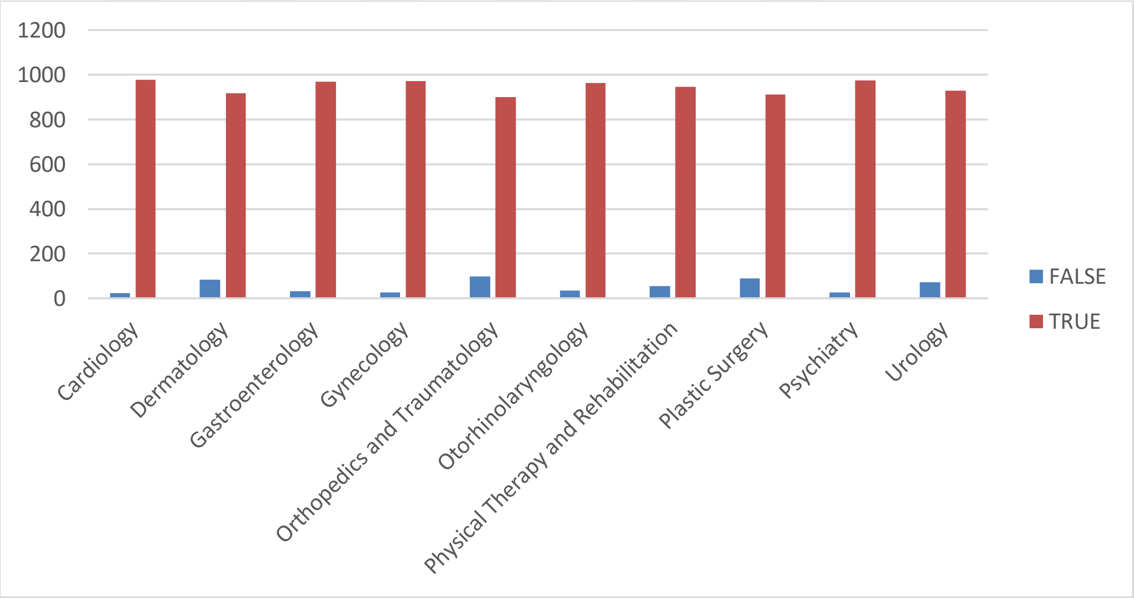

Figure 4.3 The test results of medical CNN classifier

Although CNN is more complicated and harder to implement it has the worst accuracy rate. Its additionally taking into consideration of sequence of words is expected to increase the accuracy rate, however, it appears to be lower than other approaches. Patients' narratives' being typed with just a few words by experts as short descriptions can be the reason for lower accuracy. Therefore, experimenting with more formal typed dataset can make the CNN more suitable.

4.4 Precision, Recall and F-measure

The precision is the rate of retrieved documents that are relevant, recall is the rate of relevant documents successfully retrieved and F-measure is the harmonic mean of them. They are the common metrics to measure the per-class classification accuracies (Flach & Kull, 2015). To calculate the precision, recall and F-measure the confusion matrix is needed to be generated. It is a table such as Figure 4.4 which indicates the prediction results on a classification task.

		Predicted class	
		P	*N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Figure 4.4 The confusion matrix

The formulas which are listed below use the values on confusion matrix for precision, recall and F-measure calculation.

$$\textit{Precision} = \frac{TP}{TP + FP} \tag{4.1}$$

$$\textit{Recall} = \frac{TP}{TP + FN} \tag{4.2}$$

$$F\text{ - }measure = \frac{2*Recall*Precision}{Recall + Precision} \tag{4.3}$$

The rates of the precision, recall and F-measure are listed in tables below for SVM, MNB and NN. They are calculated for each of the ten classes separately.

Table 4.2 SVM Precision, Recall and F-measures

Classes	Precision	Recall	F-measure
Physical Therapy and Rehabilitation	0.977	0.972	0.974
Gastroenterology	0.979	0.96	0.969
Psychiatry	0.969	0.989	0.979
Gynecology	0.996	0.991	0.993
Urology	0.993	0.99	0.991
Plastic Surgery	0.985	0.989	0.987
Otorhinolaryngology	0.949	0.957	0.953
Orthopedics and Traumatology	0.958	0.964	0.961
Cardiology	0.996	0.991	0.993
Dermatology	0.989	0.987	0.988

Table 4.3 MNB Precision, Recall and F-measures

Classes	Precision	Recall	F-measure
Physical Therapy and Rehabilitation	0.97	0.973	0.972
Gastroenterology	0.89	0.977	0.931
Psychiatry	0.991	0.961	0.976
Gynecology	0.98	0.996	0.988
Urology	0.989	0.996	0.993
Plastic Surgery	0.992	0.967	0.979
Otorhinolaryngology	0.95	0.876	0.912
Orthopedics and Traumatology	0.965	0.951	0.958
Cardiology	0.982	0.993	0.988
Dermatology	0.975	0.987	0.981

Table 4.4 CNN Precision, Recall and F-measures

Classes	Precision	Recall	F-measure
Dermatology	0.983	0.978	0.98
Plastic Surgery	0.923	0.916	0.92
Gynecology	0.851	0.968	0.906
Physical Therapy and Rehabilitation	0.988	0.973	0.98

Cardiology	0.919	0.901	0.91
Orthopedics and Traumatology	0.983	0.964	0.973
Psychiatry	0.927	0.946	0.937
Urology	0.926	0.912	0.919
Gastroenterology	0.993	0.974	0.983
Otorhinolaryngology	0.984	0.928	0.955

The precision, recall and F-measure rates of SVM are well balanced and they are similar to the accuracy rate of 98.16 % that has been found above. However, when the MNB and NN rates are analyzed in detail, the precision score for the class of Gastroenterology and recall score for the class of Otorhinolaryngology indicate that MNB has lower accuracy on some specific classes. Moreover, NN has a lower precision score for the Department of Gynecology. Although MNB and NN have a satisfying accuracy rate in general, it can be disappointing to predict some specific hospital departments correctly.

4.5 ROC Analysis

In terms of accuracy and precision-recall scores SVM seems to be the best classifier in our motivation. To verify the SVM in detail the ROC analysis tool which evaluates the performance of classification models in machine learning can be used (Majnik & Bosnić, 2013). When the ROC curves of each hospital department are drawn it is good to see that they are stuck on the top left corner of the diagram in Figure 4.5. It's becoming close to the y-axis and its quickly reaching to the top level support the SVM's satisfactory performance.

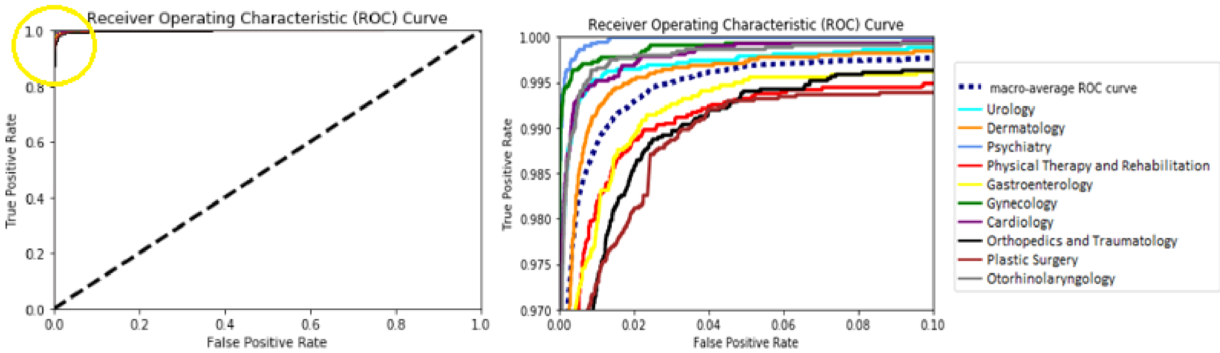

Figure 4.5 ROC curves of each hospital department

4.6 Analysis the Impact of Different Sets of Data

We have observed the change of accuracy rates when the proportion of train and test sets have been changed. It is expected that when the model is trained with less data it would be more delimited and the accuracy rate would probably decrease. In our study, we have put aside the test dataset as 20% of the all corpus and we have measured the accuracy rate as 98.16% on SVM. However, if we choose the test ratio as 40% the accuracy rate decreases as expected and it becomes 93.21%.

In addition, we have also evaluated the test accuracies by changing the number of classes in our corpus. The corpus we work on contains ten departments and each one corresponds to a class. If the number of classes is decreased, SVM would have limited hyperplanes so it would deliminate the classes more precisely. When our corpus is rebuilt with five classes which are dermatology, cardiology, gynecology, gastroenterology, psychiatry and the ratio of allocated test data is 20%, the test accuracy is measured as 99.4 %. As a result, the accuracy increases in parallel with our expectation.

5. CONCLUSION

This Turkish medical parser is developed by taking into consideration the Turkish natural language characteristics and the medical context. It can be made more comprehensive with additional lexicons and corpora, receiving mentorship from medical experts or increase in the ability of NLP techniques. As a next step, the features such as age, gender and location of a patient can be taken into consideration when building the machine learning model. Moreover, a medical treebank that provides the ability of dependency parsing can be adapted to improve the performance of this medical parser. The significance of such a medical parser cannot be ignored when it is evaluated as a tool which facilitates healthcare services for both physicians and patients. Since it provides infrastructure to develop some useful medical applications. When the parser is supported with artificial learning to classify the medical content, medical recommendation services can be implemented and integrated into available medical web or mobile applications.

As in our classification case, we have acquired the highest accuracy rate of 98.16 % by using n-gram approach on SVM. A recommendation system can suggest possible hospital departments for patients through the online appointment systems by repeating our approaches on a more comprehensive dataset that includes enough data for all hospital departments. When the service is integrated on a live system the machine learning model can be feedbacked with the higher volume of data. Therefore, periodically feeding the model with current data improve the service results and makes predictions more accurate. In a parallel perspective, it can also recommend possible diagnosis for physicians in the hospital information management system. On the other hand, after the integration of this parser through the channels stream medical text data, alarm mechanisms can be adjusted for real-time detection of epidemics. This innovation enables early intervention to epidemics before intensely spread. It can also be evolved into a personal medical consultant application by interacting with patients in a mobile interface. Some demands and advice needs of patients can be operated on a mobile application by firstly classifying their intent and then carrying on a context-specific dialogue. In addition, it leads to run analytics on unstructured medical text data according to symptoms, procedures, treatments, medicines etc.

This framework also supports the Health 4.0 concept which is a strategy to encourage the personalization of healthcare. It is achieved by the Internet of Things, cloud computing, mobile communication, personal medical mobile devices, real time data processing and interpretation, automatization of processes. Personal healthcare quality can be increased through text-based recommendation services in order to give the most accurate information to the patients.

In conclusion, the approaches followed on our study can be applied to another language. It just requires customizing the steps on natural language processing by paying attention to the characteristics of the targeted language. Thus, multilingual text-based medical applications can be developed. If we evaluate the possible future works from a different point of view, the covered methods can form a strong basis for another domain except the medical. For instance, a machine learning based decision support system for attorneys can be very beneficial. If it is accepted that we already have architecture and technological details of the recommendation web service, compelling technical requirements to develop a recommendation service to serve judicial system are preparing a judicial corpus that contains precedent cases and a specific lexicon that includes prominent terms or phrases.

REFERENCES

Akın, A. A., Akın, M. D. (2007). Zemberek, an open source nlp framework for turkic languages, Structure, pp. 1-5. URL: http://zemberek.googlecode.com/files/zemberek makale.pdf

Cao, K., Rei, M. (2016). A joint model for word embedding and word morphology, *arXiv* preprint *arXiv:1606.02601*.

Caragea, D., Cook, D., Honavar, V. G. (2005). Visual methods for examining support vector machine results, with applications to gene expression data analysis, *Technical report, Iowa State University*

Chen, P., Barrera, A., & Rhodes, C. (2010). Semantic analysis of free text and its application on automatically assigning ICD-9-CM codes to patient records, *In 2010 9th IEEE International Conference on Cognitive Informatics (ICCI)*, pp. 68-74.

Cortes, C., Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), pp. 273-297.

Eryiğit, G. (2014). ITU Turkish NLP web service, *In Proceedings of the Demonstrations at the 14th Conference of the* European *Chapter of the Association for Computational Linguistics*, pp. 1-4.

Flach, P., Kull, M. (2015). Precision-recall-gain curves: Pr analysis done right, *In Advances in Neural Information Processing Systems*, pp. 838-846.

Friedman, C. (2005). Semantic Text Parsing for Patient Records, *Medical Informatics, Vol 8 of Integrated Series in Information Systems*, Springer, Boston, pp. 423-448.

Ge, Y. (2015). Current Development and Technology in the Information Extraction for Clinical Narrative Text, International *Journal of Computer Science and Application*.

Goldstein, I., Arzumtsyan, A., & Uzuner, Ö. (2007). Three approaches to automatic assignment of ICD-9-CM codes to radiology reports, *In AMIA Annual Symposium Proceedings, Vol. 2007, American Medical Informatics Association*, p. 279.

Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y. (2016). Vol 1 of Deep learning, Cambridge: MIT press.

Hakkani-Tür, D. Z., Oflazer, K., Tür, G. (2002). Statistical morphological disambiguation for agglutinative languages, Computers *and the Humanities*, pp. 381-410.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *In* European *conference on machine learning*, Springer, Berlin, Heidelberg, pp. 137-142.

Johnson, S. B. (1999). A semantic lexicon for medical language processing, *Journal of the American Medical* Informatics *Association*, pp. 205-218.

Kazmierska, J., Malicki, J. (2008). Application of the Naïve Bayesian Classifier to optimize treatment decisions, Radiotherapy *and Oncology*, 86(2), pp. 211-216.

Kim, Y. (2014). Convolutional neural networks for sentence classification, *arXiv preprint arXiv*:1408.5882.

Kleverwal, J. (2015). *Supervised text classification of medical triage reports*, Master's thesis, University of Twente.

Lekha, (2015). Information Extraction in Medical Domain, pp. 1-14.

Liu, R. L. (2007). Text classification for healthcare information support, *In International Conference on Industrial*, Engineering *and Other Applications of Applied Intelligent Systems*, Springer, Berlin, Heidelberg. pp.44-53.

Majnik, M., Bosnić, Z. (2013). ROC analysis of classifiers in machine learning: A survey. *Intelligent data analysis*, 17(3), pp. 531-558.

Marcus, M. P., Marcinkiewicz, M. A., Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank, *Computational linguistics,* pp. 313-330.

McCallum, A., Nigam, K. (1998). A comparison of event models for naive bayes text classification. *In AAAI-98* workshop *on learning for text categorization*, Vol. 752, No. 1, pp. 41-48.

McKeown, K. R., Radev, D. R. (2000). Collocations, *Handbook of Natural Language Processing. Marcel Dekker*.

O'Shea, K., Nash, R. (2015). An introduction to convolutional neural networks, *arXiv preprint arXiv:*1511.08458.

Parthiban, G., Srivatsa, S. K. (2012). Applying machine learning methods in diagnosing heart disease for diabetic patients. *Vol 3 of International Journal of Applied Information Systems (IJAIS)*

Pedregosa, F., Varoquaux, G. (2010). Scikit-Learn User Guide, *Tech. rep.*

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Vanderplas, J. (2011). Scikit-*learn: Machine learning in Python. Journal of machine learning research*, pp. 2825-2830.

Rennie, J. D., Rifkin, R. (2001). Improving multiclass text classification with the support vector machine. *Technical* Report *AIM-2001-026, Massachusetts Institute of Technology*, AI Memo.

Rubio-López, I., Costumero, R., Ambit, H., Gonzalo-Martín, C., Menasalvas, E., & Rodríguez, A. (2017). Acronym disambiguation in spanish electronic health narratives using machine learning techniques, *Studies in health technology and informatics*, pp. 235-251.

Sak, H., Güngör, T., Saraçlar, M. (2007, February). Morphological disambiguation of Turkish text with perceptron algorithm. *In International Conference on Intelligent Text Processing* and *Computational Linguistics*, Springer, Berlin, Heidelberg, pp. 107-118.

Sotelsek-Margalef, A., & Villena-Román, J. (2008). MIDAS: an information-extraction approach to medical text classification. *Procesamiento del lenguaje Natural*, pp. 97-104.

Szolovits, P. (2003). Adding a medical lexicon to an English parser, *In AMIA annual symposium proceedings (Vol. 2003, p. 639).* American *Medical Informatics Association.*

Tong, S., Koller, D. (2001). Support vector machine active learning with applications to text classification, *Journal of machine learning research*, pp. 45-66.

Zhou, L., Plasek, J. M., Mahoney, L. M., Karipineni, N., Chang, F., Yan, X., Chang, F., Dimaggio, D., Goldman, D. S., Rocha, R. A. (2011). Using Medical Text Extraction, Reasoning and Mapping System (MTERMS) to process medication information in outpatient clinical notes, *Annual Symposium proceedings. AMIA Symposium*, pp. 1639–1648.

BIOGRAPHICAL SKECTH

Ahmet BARDIZ was born in 1988 in K.Maraş, Turkey. After graduating from the Sivas Science High School, he has studied Computer Science at Bilkent University. He has completed his undergraduate education in 2011 and then he has begun to work on various data-driven projects of telecommunication and finance companies. In 2017, he has established his own startup company which develops products and solutions based on the concepts of real-time data analysis, natural language processing and artificial intelligence. Recently, he works as a data scientist and studies the Master of Science in Computer Engineering at Galatasaray University.

PUBLICATIONS

Bardız, A. (2017). Medikal metin ayrıştırma ve sınıflandırma: semptomdan hastane branşına, *2. Ulusal Biyomedikal Cihaz Tasarımı ve Üretimi Sempozyumu,* UBICTUS.

Bardız, A., Parlak, İ. B. (2018). A new method for Turkish medical text parsing, *International Congress on* Engineering *and Life Science,* ICELIS.

Bardız, A., Parlak, İ. B. (2019). A new method of classification for clinical text processing, *The Sixth International Symposium on Engineering, Artificial Intelligence and Applications,* ISEAIA.