

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE AND SCHOOL OF PUBLIC HEALTH

MSc. in Health Informatics Program

Applying Text Mining Techniques to Extract Knowledge from Cancer Patients' Medcial Records-The case of Tikur Anbessa Specialized Hospital

By
BETHELHEM ABEBE

January 2018

ADDIS ABABA UNIVERSITY SCHOOL OF INFORMATION SCIENCE AND SCHOOL OF PUBLIC HEALTH

Applying Text Mining Techniques to Extract Knowledge from Cancer Patients' Medcial Records-The case of Tikur Anbessa Specialized Hospital

A Project Submitted to the School of Information Science and Public Health of Addis Ababa University in Partial Fulfillment of the Requirements for Degree of Master of Science in Health Informatics

By:
BETHELHEM ABEBE

January 2018

ADDIS ABABA, ETHIOPIA

ADDIS ABABA UNIVERSITY SCHOOL OF INFORMATION SCIENCE AND SCHOOL OF PUBLIC HEALTH

Applying Text Mining Techniques to Extract Knowledge from Oncology Patients' Medcial Records-The case of Tikur Anbessa Specialized Hospital

Names and Signature of Members of the Examining Board

1.

Examiner Signature Date

2.

Examiner Signature Date

3.

Advisor Signature Date

4.

Advisor Signature Date

DEDICATION

To my beloved Children Nebyu, Arsema and Abenezer Daniel

ACKNOWLEDGMENT

I would like to express my gratitude and heartfelt thanks to my advisors **Dr. Martha Yifiru(phd)** and **Ato Mengistu Yilma** for their guidance, valuable comments and kindness throughout the difficulties I faced while conducting my project. I would also like to extend my deepest gratitude to Tikur Anbessa specialized hospital oncology staffs for sharing information and facilitating things in all aspects of my project.

Finally, I would like to express my gratitude to my family especially my elder brother and beloved husband for all the support they have given me during the master's program.

TABLE OF CONTENTS

LIST OF TABLES **8**
LIST OF FIGURES **9**
ACRONYMS **10**
ABSTRACT **12**
1. INTRODUCTION **14**
1.1. Background **14**
1.2. Statement of the problem **17**
1.3. Objectives of the Project **19**
1.3.1. General Objective 19
1.3.2. Specific Objectives 19
1.4. Significance of the project **20**
1.5. Scope of the project **21**
1.6. Definition of Key Terms **22**
1.7. Ethical Clearance **22**
1.8. Organization of the Project **22**
2. LITERATURE REVIEW **23**
2.1. General Literature **23**
2.1.1. Cancer 23
2.1.2. Cancer in Ethiopia 25
2.1.3. Text Mining 26
2.1.4. Text mining techniques 28
2.1.5. Text Mining Methodologies 29
2.1.6. Text mining tools 33
2.1.7. Machine Learning 34
2.1.8. Clustering Algorithm 35
2.1.9. Distance measures in clustering 40
2.1.10. Approaches to Evaluation 42
2.2. Related Literature **44**
3. METHODOLOGY **48**
3.1. Determine the Purpose of the Project/Business Understanding **48**
3.2. Explore the Availability and the Nature of the text/text understanding **49**
3.3.Preparing the text **52**
3.4.Extraction of Patterns **52**

3.5.Evaluation of the model **52**
3.6.Inclusion criteria of the project **53**
3.7.Exclusion criteria **53**
4. DISCUSSION OF RESULTS **54**
4.1. Prepare the Text **54**
4.2. Extract the Knowledge/Modeling **55**
4.2.1. K means Clustering 60
4.2.2.Hierarchical clustering 63
4.3.Evaluation of the extracted knowledge **67**
4.3.1.Subjective Evaluation 67
4.3.2. Objective Evaluation 68
4.5. Parameters affecting the system performance **71**
4.6. Deployment **71**
4.6.Strength and Limitations of the Project **72**
4.6.1. Strength of the Project 72
4.6.2.Limitations of the Project 72
4.7. Lessons Learnt and Challenges **73**
4.7.1. Lessons Learnt 73
4.7.2. Challenges 73
5. CONCLUSIONS AND RECOMMENDATIONS **74**
5.1. Conclusions 74
5.2. Recommendations 76
APPENDIX **79**
DECLARATION **87**

LIST OF TABLES

Table 1: Term-document matrix……...50

Table 2: The most frequent terms……51

Table 3: Manual cluster done by the principal investigator……..62

Table 4: Cluster done by an oncologist……63

Table 5: Objective evaluation results……64

LIST OF FIGURES

Figure 1: Example of text mining …….24

Figure 2: CRISP methodology …….…….....25

Figure 3: Types of hierarchical clustering……34

Figure 4: Snapshot of the oncology medical record…….47

Figure 5: Word frequency……..52

Figure 6: Word cloud……..53

Figure 7: Optimal number of clusters……...55

Figure 8: K means clustering…….56

Figure 9: Hierarchical clustering……...58

Figure 10: Fan shaped hierarchical clustering…….59

Figure 11: Precision/Recall measures……...63

ACRONYMS

AIDS- Acquired immune deficiency syndrome

Bm- Biological marker

CDSS- Clinical decision support system

Clara- clustering large applications

CRAN- Comprehensive R Archive Network

CRISP-DM- Cross industry standard process for Data mining

DM- Data mining

DTM- Document term matrix

EHR- Electronic health record

EM- Expectation maximization

EM-Enterprise miner

FMOH- Federal ministry of health

GATE- General Architecture for Text Engineering

HAC- Hierarchical agglomerative clustering

HIV- Human immune virus

HMIS- Health management information system

ICT- Information and communication technology

IDE- Integrated development environment

IE- Information extraction

IR-Information retrieval

KDD- Knowledge discovery from Database

MEDLINE-Medical literature analysis and retrieval system online

MeSH-Medical subject headings

NLP- Natural language processing

NMF- non negative matrix factorization

OPD- Out patient department

PAM-partitioning around mediods

SEMMA- sample, explore, modify, model and assess

SIB-Sequential information bottleneck

SKM- Simple k means

SOM- Self organizing maps

SSO_AB- Society of surgical oncology annotated

St- Saint

TASH- Tikur anbessa specialized hospital

TDM- Term document matrix

Tm-Thought marker

TM- Text mining

UMLS- Unified medical language system

VC-DM- Virtuous Cycle of Data Mining

ABSTRACT

Background: Currently, there are a lot of medical texts accumulated than interpreted. These texts have to be organized and analyzed effectively in order to be useful. Nowadays text mining has become very important in analyzing these medical texts and finding patterns effectively. The medical records of chronic ill patients (especially cancer) contain a lot of information both image and textual formats which are vital to immediate patient care which could also help in different researches and finding difficult cases.

Objective: The aim of this project is applying text mining techniques to extract knowledge from oncology patients' medical records.

Method: In order to conduct this project, data was collected from oncology patients' medical records of Tikur Anbessa specialized hospital. The CRISP methodology was applied for describing the pattern in these medical records. For extracting the patterns from 137 medical records, R software was used. After creating a corpus and having pre processed the medical records, a pattern was extracted using hierarchical and k means clustering algorithm. The patterns extracted from these two algorithms were compared and evaluated. To evaluate the pattern extracted both the subjective and objective evaluation approaches were used. The subjective evaluation was done with the help of ten physcians (both residents and oncologists). For the objective evaluation Rand index, accuracy, precison, recall and F measures were performed.

Result: According to the assessment of the medical records indicated, searching the necessary medical records from the record room was difficult almost impossible, their follow up formats are disorganized and the physicians' handwriting is illegible. These make knowledge discovery difficult, time taking and tiresome. As the objective evaluation methods showed the hierarchical algorithm performed better than the k means (Rand index=66.2%, Accuracy=48.8% and Precision and recall=65.6%) and 50% of the physicians chose also the hierarchical algorithm. During the subjective evaluation, out of the ten physicians, three of them (30%) did not have any idea as which one is better because the idea was new to them and difficult to understand the patterns. Two of them (20%) preferred the k means to the hierarchical because the hierarchical seemed complicated to them. The rest (50%) chose the hierarchical algorithm since it tried to show almost the necessary pattern found in the patients' medical records.

Conclusion: In general text mining eases access of the necessary knowledge rather than going through patients' medical records and any one can get the necessary knowledge from the pattern extracted in the oncology patients' medical records. From the project it is easy to see

that the hierarchical algorithm performed better.

Recommendation: Text mining has different applications and each application has different benefits in the medical health care. And different kinds of knowledge can be discovered and predicted not only from cancer medical records but also from other chronic illnesses that need further researches and experiments. For researchers, there is a great need of text mining applications in the medical domain specially using clustering algorithms in order to extract new knowledge. Also the efficiency of the k means and hierarchical clustering needs to be improved. For the health practitioners and Tikur anbessa specialized hospital, this application will give them a great benefit, so handling the patients' medical record in a proper and organized way will give opportunity to give quality of care for the patient. Also for the physicians and different researchers, it will give ease acccess of the necessary data for knowledge extraction and patient management system. For software development business organizations, there is a great opportunity to work on the text mining area especially in the medical domain which needs more structuring and handling medical data.

1. INTRODUCTION

1.1. Background

Currently, there are a lot of medical data accumulated than interpreted. These data have to be organized and analyzed effectively in order to be useful. New information technologies are being used to manage this vast amount of medical data in order to discover useful patterns and new knowledge. And this specific technology is text mining which aims to extract useful knowledge from textual data or documents (1). Knowledge can be gained by drawing conclusion from what is already known before (2).

Whenever there is a large amount of data (both textual and image) at hand any physician will have a difficulty in making decisions. Nowadays, Ethiopia is in the verge of automating the health care system by using the health management information system (HMIS) with the experts at hand. These experts are trying to capture and store the necessary data, which are especially important for researchers and policy makers but this is not enough. A lot of work is waiting ahead of us. The medical data needs a great deal of organization so as to make it more structured in which any one in need of it can have an easy access, be it for clinical decision, research even planning for policy makers who are at the ministry level. In order to deal with the issue of organizing and structuring medical data, health informatics plays a great role. Health informatics is a multidisciplinary field that uses health information technology to improve health care via any combination of higher quality, higher efficiency and new opportunities (3). One of the new opportunities for the organizing and structuring of medical data can be done by using text mining. Text mining is the discovery of interesting knowledge in text documents. It is a very difficult task in finding accurate knowledge (or features) in text documents to help users for the intended use (4).

In (5) text mining is described as the discovery of previously unknown information or concepts from text files by automatically extracting information from several written resources using computer software. In text mining, the files mined are text files which can be in one of the forms: unstructured text is usually in the form of summaries and user reviews whereas structured text consists of text that is organized usually within spreadsheets or tables.

As text data become more bulky and bulky, the challenge is to find new strategies to extract relevant and important information from these data effectively without the need for anyone to read all these texts. All applications for text mining solve this challenge by any means. In

general, text mining is becoming very important to all activities and domains where formal analytical approaches add value (6).

As defined in (7), text mining is the technology which discovers patterns and trends semi-automatically from a vast amount of collections of unstructured text.

From the experience seen in the developed countries, medical data analysis can lead to an enhancement of health care by improving the performance of patient management tasks (8). The application of text mining techniques to the domain of cancer is one of the newest and promising areas of research for the analysis of the data which will discover new knowledge (9).

Having new generation of computational theories and tools to assist people in extracting useful information and knowledge from the rapidly growing volumes of different types of cancer data, and apply decision support and intelligent systems is very important (10).

According to the Federal Ministry of Health, Cancer is an emerging public health issue in Africa. The ministry is starting to include cancer as part of its efforts to control non communicable diseases (11).

There are several directions about text mining applied in cancer research some of them are: hypothesis generation, which helps in drug discovery to explore the possible solution to the existed problems, cancer risk assessment (also famous known as CRA), evaluating the environment influence such as chemical and exposure also knowledge gathering and discovery, which is important in finding the hidden knowledge and rules behind huge amount of information, which will accelerate the research process (7).

Having accurate, comprehensive and timely relevant cancer data is very crucial for studying the causes of cancer, detecting cancer earlier, preventing or determining the effectiveness of treatment, specifying the reasons for the treatment ineffectiveness and cancer control programs (10).

The Federal Ministry of Health (FMoH) in Ethiopia is responsible for controlling the overall operation of the country's healthcare system. One of the transformation agenda of Federal ministry of health is information revolution. The term information revolution refers to the *"phenomenal advancement on the methods and practice of collecting, analyzing, presenting and disseminating information that can influence decisions in the process of transforming*

economic and social sectors". This will change the manual use of data to a more modernized (computerized) system with the necessary information technologies (12).

Federal ministry of health in (12) stated, having the latest health system and ease access of the necessary health information for the public is the key to the quality of health care. With accurate and reliable information, decisions at different levels of the health sector can be very effective.

In Ethiopia in order to meet the needs of physicians in their daily decision making, information technologies are being implemented in health care system**.** The use of information technology in healthcare will help in the comprehensive management of medical knowledge and its secure exchange between healthcare providers and beneficiaries. Having the necessary information in a computerized way can increase the quality of decision making, avoid human error and increase the quality of health care. Whenever a large volume of data is processed manually, the quality of data will be poor and decision making will also be poor. Since new knowledge is being accumulated daily in a growing rate in the health care system, there is a great need to have computerized system. However, changing this information to knowledge is a difficult task. In order to perform this task all health institutions need to have an expert to analyse the medical data which is time taking and expensive (13).

When we come to the issue of cancer patients' medical records, it is difficult, if not impossible, for even the most sophisticated oncologists to locate, read, see patterns emerge and memorize what is important in that bulky medical record. *"The ability to finally find the proverbial haystack is the promise of text and data mining"* (14).

1.2. Statement of the problem

Health informatics has become very important in analyzing the daily accumulating and vastly growing medical data in order to gain new knowledge. This new knowledge in turn can improve the quality of health care offered to patients. But there are a lot of issues that need to be dealt with when manipulating the complex medical data in a fruitful way (15).

These days a patient record management system is highly desired in clinical settings. The major reasons include physicians' significant information needs and clinical information overload. Basically textual health information is classified into two main categories: patient-specific clinical information and knowledge-based information, which includes research reported in journals, books, technical reports, and other sources. Both types of information are growing at an overwhelming pace (16).

Medical records contain a large amount of information, both image and textual formats. And in this project, the main focus is on textual data which have unstructured format. Most of the time, the textual data are patient notes written by clinicians. These notes include the long-term course of patient's illness, procedures, different types of investigations and treatments (especially patients with chronic illness). Also they contain important information which are directly important to the patient care and help in different epidemiological studies. The patient medical records also contain personal and social information in a single patient's complete record (17).

Even if the medical texts of all hospitals and cases are ungrammatical, composed of short and telegraphic phrases, the medical record of cancer cases are the most bulky of all with lots of medical notes, investigation papers and consultation papers. This medical record has a lot of its own format, this makes going through the medical record time taking and tiresome. And with the system being manual and the hospital having a lot of patients and new cases seen every day, it is becoming very challenging for doctors to investigate all these documents and discover significant new knowledge. For example if there has been a difficult case seen and discussed with senior physicians, after solving the problem and treating the patient the medical record will be returned to the record room where it is kept until needed. Whenever this particular medical record is needed for teaching and learning process, it will be very difficult to look for it.

The other problem is difficulty of having the necessary data needed for different researches especially in the area of cancer. The hospital record management system is manual and the oncology unit has only two staffs to handle more than five thousand records and also accepting a lot of new and follow up patients every day. All these factors make things disorganized and messy which could be a problem for teaching/ learning process and for different researches. As Tikur Anbessa specialized hospital is one of the medical teaching hospitals in Ethiopia, the need to have structured information that serves as a base for different researches arises. Structured information is the result of text mining after having been pre processed and transformed into a specific pattern. This pattern in return will help in the knowledge discovery phase.

To this end, this project aimed at the application of text mining techniques to extract knowledge from medical records, specifically from cancer patients' records found at Tikur Anbessa Specialized hospital.

1.3. Objectives of the Project

1.3.1. General Objective

- The general objective of this project is to apply text mining techniques to extract knowledge from cancer patients' medical records.

1.3.2. Specific Objectives

- To assess the nature of texts found in cancer patients' medical records
- To examine text mining methodologies that is needed in the extraction of patterns from the cancer medical records
- To extract patterns from the texts in these medical records
- To examine the patterns extracted from these medical records
- To describe the patterns found using clustering algorithms and
- To evaluate which clustering algorithm best describe the patterns extracted

1.4. Significance of the project

Having information technology in any discipline gives the benefit of storing information in a structured manner (“a manner that the computer can read and understand"), so that it can be served to the end user when needed, in a desired manner (18).

For Health Professionals:

1. With the structured information at hand, the health professionals will be able to make the right decision with the right information to the right patient at the right time.
2. In the future, the result of this project will be a base for clinical decision support system (CDSS) and to be included when the hospital have electronic health record (EHR).
3. Whenever the physician faces difficult cases, this extracted information will be a base for future knowledge discovery and learning.

For Patient:

The patient will get quality of care which is proper treatment with in short period of time. The structured information will enable the physician to make the proper decision as early as possible. Having unnecessary investigations instead of the treatment will cost a patient life. Sometimes a physician may send unnecessary investigation because of a dilemma (without having a clear idea on how to solve the problem). However, if the physician has structured information from previous similar problems, he can solve the problem easily and help the patient in improving his health. So the patient won’t have any ups and downs.

For Policy Makers and FMOH:

1. The structured information will enable the policy makers to make the right decisions in order to allocate the necessary medical resources.
2. For bio surveillance, clinical decision making, text mining and automatic terminology management.

For Researchers:

1. Aggregating the structured information can give us valuable clues for initial medical research, for instance when trying to uncover relations between findings, symptoms and drug use.
2. To have some idea on the nature and conditions of the oncology patients' medical records.
3. To serve as a base for further research both in the area of text mining and cancer

1.5. Scope of the project

The project is about applying text mining techniques to extract knowledge from oncology patients' medical records for the purpose of structuring the information and evaluate the patterns extracted. In order to extract the patterns k means and hierarchical clustering algorithms were used. The study focused on cancer patients' medical records in Tikur Anbessa Hospital because cancer cases are increasing significantly and becoming a challenging issue. Moreover, Tikur Anbessa Hospital is one of the first oncology centers and is a teaching referral hospital in Ethiopia. The data collection is conducted between April and August 2017.

1.6. Definition of Key Terms

Cancer: is non communicable disease that can happen during the abnormal growth of a cell in any of our body organs.

Clinical decision support system (CDSS): is a way of decision making that use different information technologies in order to help oncologists make clinical decisions**.**

Clustering: is a way of grouping similar items or dividing things in to different groups according to their similarity.

Ideal Clustering: is a clustering which is done manually by an expert in order to show the relevant texts that should be included in each cluster.

Medical Record: is a folder that contains oncology patients' medical conditions including history of past, present illneses, diagnosis, treatment or plan and investigations ordered with their results and/or admission and discharge history.

Oncology: is part of medicine that study about cancer.

Oncologist: is a physician who has specialized in oncology.

Text mining(TM): is a technology that enables to extract knowledge from collection of unstructured medical texts.

1.7. Ethical Clearance

The project was carried out after getting permission from the ethical clearance committee of Addis Ababa university school of public health. Official letter of cooperation was written to all respective health centers and permission to conduct the project was secured accordingly.

1.8. Organization of the Project

This project is organized into 5 chapters. Chapter 1 presents an introduction consisting of background, statement of the problem, objective, significance of the study, definition of key terms, scope, and outline of the project. Chapter 2 provides a general literature on cancer and text mining and related works done in the area. Chapter 3 deals with the methods, tools and techniques used to attain the objective of the project. Chapter 4 discusses about the preprocessing of the texts, presents the process of data analysis, findings, strength and limitation of the project, lesson learnt and challenges of the project. Chapter 5 is dedicated to concluding remarks as well as recommendation for future work.

2. LITERATURE REVIEW

2.1. General Literature

2.1.1. Cancer

Cancer in (19) is defined as a term used for diseases in which abnormal cells are divided uncontrollably and invade other tissues and organs. And these cancer cells spread to other parts of the body through the blood and lymph systems. The body is made up of many different types of cells, such as skin cells, muscle cells, and blood cells. There is a continual division of many normal cells in our body inorder to create new cells. Our body has an internal system that can detect when a cell gets old and should make a space for the new cell. During the division of a cell if an error occurs, the new cell become cancerous. And this cancerous cells do not have the mechanism that help the cell to die instead the abnormal cells will be accumulated. When this occurs, it can form into a mass of tissue, called a tumor, or it can crowd out the good, healthy cells, like with leukemia or other cancers that affect the blood. If cancer cells leave their original place and move to other parts of the body, this is called metastasizing.

There are different types of cancers which in turn are grouped into categories, these categorizations describe where the cancer orginally started from. The main categories of cancer are:

- **Carcinoma:** the most common kind of cancer which is generally named by the place in the body where the cancer begins, such as the lung, breast, or colon.
- **Sarcoma:** this type of cancer involves the skin and musculoskeletal system such as bone, muscle, or fat.
- **Leukemia:** Cancer of blood cells or bone marrow
- **Lymphoma:** this type of cancer involves the immune system cells within the lymphatic system.
- **Central nervous system cancers:** Cancer of the brain or spinal cord (19).

In general , if any of these spread of cancers are not controlled on time, it can result in death (20).

As different researchers and studies pointed out cancer has become the second largest contributor to the non-communicable disease burden and its impact continues to rise. It may start developing in any of over 60 body organs and is usually named after the affected organ. There are more than 200 types of cancer that have different causes, symptoms and treatments (21).

In the study done by the Harvard School of Public Health and the World Economic Forum new cancer cases in 2010 were approximately 13.3 million and the number will increase to 21.5 million in 2030. This rises in the number of cases will have a negative impact because treating cancer is becoming more expensive than any other diseases. And this in turn will demand increased share of health system budgets (22).

According to disease prevention and control directorate of Ethiopia cancer kills more than 7.9 million people globally every year constituting close to 13% of total deaths worldwide. Even if the communicable diseases are still the leading killers in many developing countries, the incidence and mortality from non-communicale diseases is raising rapidly. This has created a 'double burden' of diseases, which is imposing strain on existing health system. Because of an increasing prevalence of risk factors such as smoking, overweight, physical inactivity and changing reproductive patterns associated with the economic development, the occurrence of cancer is increasing tremendously. The most frequently diagnosed cancer cases are lung and breast cancer and being the leading causes of cancer death in men and women, respectively, in all over the world (23).

The histological type of cancer, the stage (degree of spread), and the patient's performance status are the main factors that determine the treatment and prognosis of cancer. The Possible treatments include surgery, chemotherapy, and radiotherapy. Survival of the patient depends on the stage of the cancer, health status and other factors like taking other continous medications. The constant and day to day unavoidable exposure to environmental carcinogens complicates the investigation of cancer causes in human beings. The causes of cancer is very complex especially challenging for cancers with long latency, which are associated with exposure to ubiquitous environmental carcinogens. Since cancer care is becoming increasingly sophisticated and complex, the information, image data and steps that comprise a single patient's healthcare journey has grown dramatically (24). Here comes the need for text mining which helps in the new knowledge discovery and pattern finding. In order to find the information we need, we don't necessarily have to read all these documents.

2.1.2. Cancer in Ethiopia

"*Ethiopia is home to a growing population of more than 84 million people and is expected to become the ninth most populous country in the world by 2050, with an estimated parallel rise in cancer burden*" (11).

As in (23), cancer in Ethiopia accounts for about 5.8% of total national mortality. It is estimated that the annual incidence of cancer is around 60,960 cases and the annual mortality is over 44,000. For people under the age of 75 years, the risk of being diagnosed with cancer is 11.3% and the risk of dying from the disease is 9.4%. The most prevalent cancers in Ethiopia among the adult population are breast cancer (30.2%), cancer of the cervix (13.4%) and colorectal cancer (5.7%). About two-thirds of reported annual deaths occur among women.

According to disease prevention and control directorate of Ethiopia, in Tikur Anbessa specialized hospital about 80% of reported cases of cancer are diagnosed at advanced stages when very little can be done to treat the disease. This is largely due to (25):

- The people having low awareness about cancer signs and symptoms,
- Inadequate screening and early detection and treatment services of health facilities,
- Inadequate diagnostic facilities and poorly structured referral of low levels of health system and
- The country has very few cancer specialties.

All these factors create difficulty for the great majority of the population to access cancer treatment services, which results in long waiting times and cause many potentially curable tumors to progress to incurable stages and even death (25).

As presented in (26) most of the patients presented in TASH oncology center are either locally advanced, metastatic or unknown stage (64 %) when they compare with early stage (10 %) and the rest 25% was level as unspecified post operative. In Ethiopia where oncology practice is so young, awareness' even among medical professionals about oncology is much inferior than expected (23).

2.1.3. Text Mining

"The word mining in the term text mining originated from an analogy to coal mining. As with the coal mining, a vast amount of unstructured raw material first be dug up, exposed and processed in text mining, yielding valuable precious metals or, with the latter, profitable knowledge (27)*."*

Text mining is a broad umbrella for describing different technologies that is used to analyze and process semistructured and unstructured text data in order to discover new knowledges and patterns (6).

Text mining is also known as knowledge discovery from text, is considered the method of interesting patterns from a large text collections for discoverying knowledge (28).

Text mining helps in finding or discoverying patterns among words whch was difficult or tiresome to find. The process of text mining includes the following steps:

- Information collection from unstructured data
- Conversion of collected data to structured form
- Pattern identification from structured data
- Pattern analysis
- Extraction and storage (29)

Text mining has become very important in giving rise to an age where vast amounts of textual information can be accessed, analyzed, and processed within a short period of time. The benefits of text mining is not only for search but also have yielded innovations that help people better understand and make use of the information in document repositories. Domain-specific extraction patterns (or something similar) are used to identify relevant information (30).

Text mining aims to extract useful knowledge from textual data or documents. Although text mining is often considered as a subfield of data mining, some text mining techniques have originated from other disciplines such as information retrieval. Most knowledge management, data mining and text mining techniques involve learning patterns from existing data or information and are therefore built upon the foundation of machine learning and artificial intelligence (31).

"Knowledge management, data mining, and text mining techniques have been applied to different areas of biomedicine, ranging from patient record management to clinical diagnosis, from hypothesis generation to gene clustering, and from spike signal detection to protein structure prediction" (16).

The primary goal of text mining is to extract the knowledge that is hidden in text and to present it in a clear and correct form to medical professionals or researchers. Text mining applications integrate a wide range of different data resources, providing tools for the analysis, extraction and visualization of information, with the purpose of helping biologists to transform available data into usable information and knowledge (1).

Text mining involves three major activities. These are:

1. The information retrieval, to gather relevant text;
2. The information extraction, to identify and extract specific information from the text of interest; and
3. The knowledge discovery, to find associations among pieces of information extracted from various text sources (1)**.**

In the perspective of healthcare, text mining technologies can give several benefits such as:

- Provide appropriate access to the key information recorded in free-text such as patient's diagnoses, lab tests preformed, medications prescribed, and their outcome that would facilitate the sound clinical decision making in a timely manner.
- Provide quick access to new and past results (such as patient's response to a therapy) that would increase patient safety and effectiveness of care.
- Enhance legibility and reduce the redundant experiments or tests performed over patients that can effectively reduce the cost of treatment.
- Generate timely alerts and computerized decision support systems that would give best clinical practices and accelerate services to the patients.
- Identify suitable individuals for clinical trials or comparative effectiveness studies.
- Facilitate the enrichment of databases and literature-based knowledge bases.
- Perform knowledge discovery and association mining in order to find the association or linkage between different biomedical events (32).

2.1.4. Text mining techniques

According to (32) text mining involves the application of techniques from areas such as information retrieval, natural language processing, information extraction and data mining. These various stages can be combined together into a single workflow:

Information Retrieval **(IR)** systems identify the documents in a collection which match a user's query. For example, if a researcher is interested in mining information only about protein interactions, he/she might restrict their analysis to documents that contain the name of a protein, or some form of the verb 'to interact', or one of its synonyms. Already, through application of information retrieval, the vast accumulation of scientific research information can be reduced to a smaller subset of relevant items.

Natural Language Processing **(NLP)** is the analysis of human language so that computers can understand research terms in the same way as humans do. Although this goal is still some way off, natural language processing can perform some types of analysis with a high degree of success. The role of natural language processing is to provide the systems in the information extraction phase with linguistic data that the computer needs to perform its 'mining' task.

Information Extraction **(IE)** is the process of automatically obtaining structured data from an unstructured natural language document. Often this involves defining the general form of the information that the researcher is interested in as one or more templates, which are then used to guide the extraction process. Information extraction systems rely heavily on the data generated by natural language processing systems.

Data Mining **(DM)** (often known as knowledge discovery) is the process of identifying patterns in large sets of data. When used in text mining, data mining is applied to the facts generated by the information extraction phase. The results of the data mining process are put into another database that can be queried by the end-user via a suitable graphical interface. Even though most of the experiment in biomedicine has been done using text/data mining techniques, its concepts go beyond biomedicine to many other disciplines.

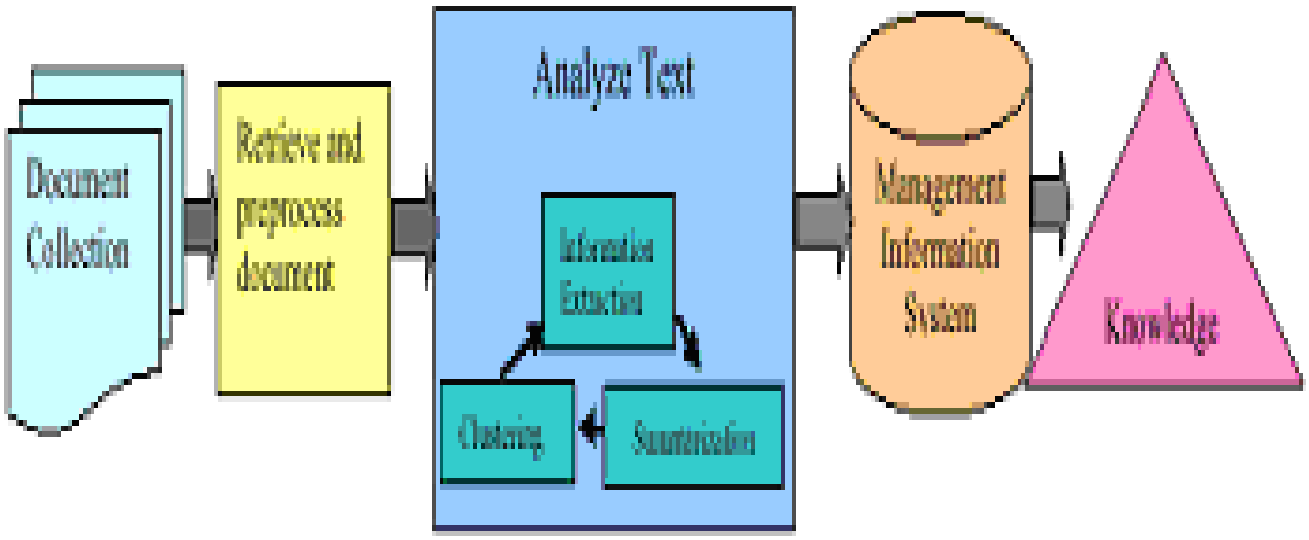


Fig 1. An example of text mining[(6)**: p.2]**

2.1.5. Text Mining Methodologies

A methodology is a documented and somewhat standardized process for executing and managing complex projects that include many interrelated tasks (i.e., extracting knowledge from textual data sources) by the use of a variety of methods, tools, and techniques. A well-designed and properly followed/ implemented methodology can help to ensure consistent and successful result (6).

Text mining is often considered as a subfield of data mining (33). Since I couldn't able to find a text mining methodology and in different applications of text mining many text miners used data mining methodologies, I have disscussed some of the data mining methodologies. In different literatures, various Data Mining methodologies are proposed, in form of scenarios of gathering and preparing data for further analysis, as well as dissemination of results for implementationof certain solutions. The most frequently used ones are CRISP (CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING), KDD, SEMMA, VC-DM.

2.1.5.1. CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING (CRISP-DM)

CRISP-DM is a general purpose methodolgy which is industry independent technology and it is said to be the de facto standard for data mining (DM). The first version of the CRISP-DM specification was developed by a consortium of European and American private companies in

1996, aiming to create a non- proprietary and freely available standardized process model and tool set for DM (data mining) application engineering. As shown in figure 2 the CRISP-DM proposes an iterative process flow with non strictly defined loops between phases and an overall iterative cyclical nature of the data mining project itself. CRISP-DM breaks down the life cycle of a data mining project into six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment (6).

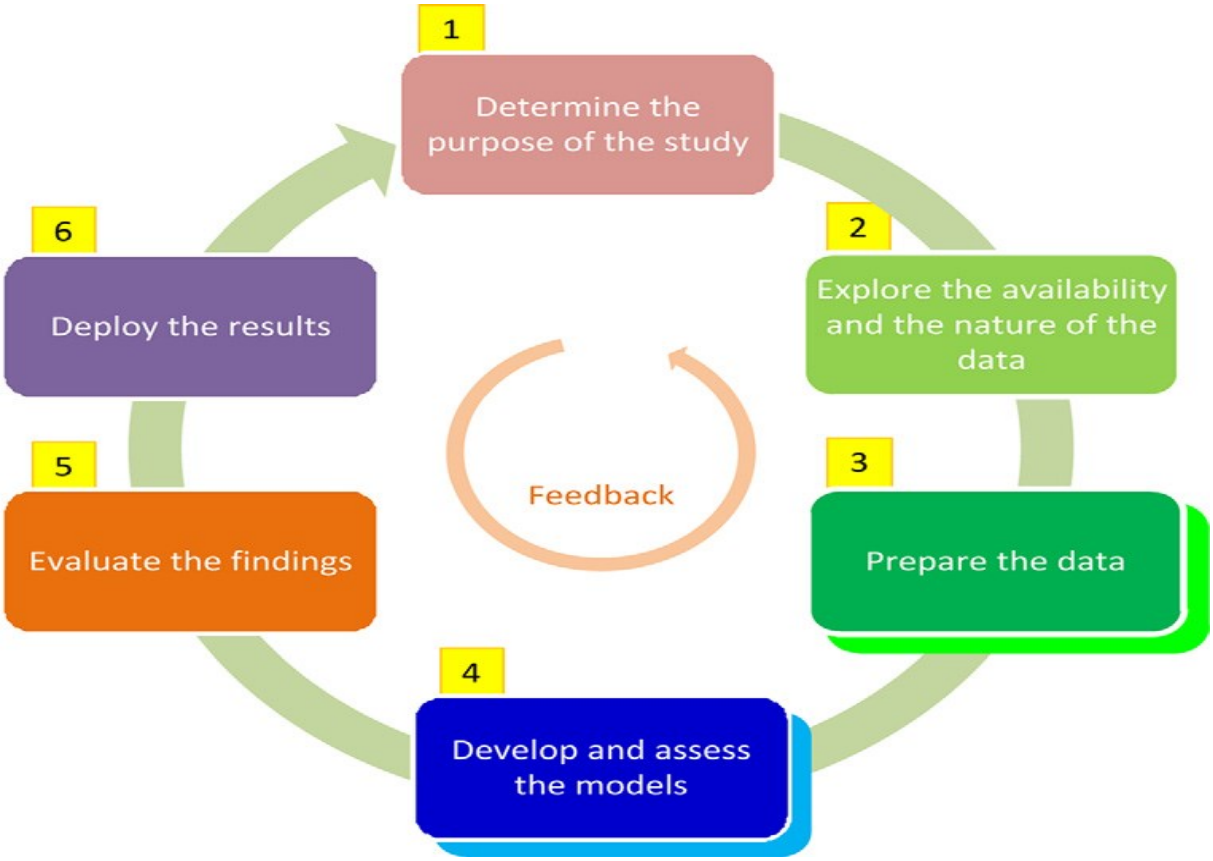


Figure 2. CRISP-DM Process flow [(6) **p.75]**

Phase one: Determine the Purpose of the Project/Business Understanding

Like any other project activity, text mining study starts with the determination of the purpose of the study which requires a thorough understanding of the business case and what the study aims to accomplish. Before defining the aims of the study, it is necessary to assess the nature of the problem (or opportunity) that initiated the study (6).

Phase two: Explore the Availability and the Nature of the text/text understanding

The data understanding phase starts with an initial data collection. The analyst then proceeds to increase familiarity with the data, to identify data quality problems, to discover initial insights into the data, or to detect interesting subsets to form hypotheses about hidden information. The data understanding phase involves four steps, including the collection of initial data, the description of data, the exploration of data, and the verification of data quality (6).

Phase Three: Data Preparation

The data preparation phase covers all activities to construct the final data set or the data that will be fed into the modeling tool(s) from the initial raw data. Tasks include table, record, and attribute selection, as well as transformation and cleaning of data for modeling tools. The five steps in data preparation are the selection of data, the cleansing of data, the construction of data, the integration of data, and the formatting of data (34).

Phase Four: Modeling

In this phase, various modeling techniques are to be selected and applied and their parameters are calibrated to optimal values. Typically, several techniques exist for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase may be necessary. Modeling steps include the selection of the modeling technique, the generation of test design, the creation of models, and the assessment of models (34).

Phase Five: Evaluation

Before proceeding to final deployment of the model built by the data analyst, it is important to more thoroughly evaluate the model and review the model's construction to be certain it properly achieves the business objectives. Here it is critical to determine if some important business issue has not been sufficiently considered. At the end of this phase, the project leader then should decide exactly how to use the data mining results. The key steps here are the evaluation of results, the process review, and the determination of next steps (34).

Phase Six: Deployment

Model creation is generally not the end of the project. The knowledge gained must be organized and presented in a way that the customer can use it, which often involves applying "live" models within an organization's decision-making processes. It also includes a summary of the project and its experiences and all of the previous deliverables, summarising and organising the results (34).

2.1.5.2. Knowledge Discovery in Database (KDD)

The term Knowledge Discovery in Database (KDD) was used in 1991 for the first time. Subsequent work led to establishing the KDD process as a methodology. It contained results of cooperation of many researchers and business analysts. The KDD process comprises 5 stages (35):

• Selection (of a data set).
• Pre-Processing (data cleaning and preparation to modeling).
• Transformation (converting data for application of a specific method).
• Data Mining (use of DM tools to search for hidden patterns).
• Interpretation/Evaluation (interpretation and assessment of "unearthed" knowledge).

2.1.5.3. SEMMA

This is another methodology, used just as frequently, assumed to be competitive to CRISP-DM, is known as SEMMA. The name was proposed by Bulkley in 1991, but it was not commercially implemented until 2008. It is used in the Enterprise Miner (EM) software and obviously is most effective with this software. It mainly focus on the modeling tasks, leaving the business aspect out. There are 5 distinct stages of knowledge discovery, for which there are tools available in the EM (35):

• Sampling (Input Data Source, Sampling, Data Partition).
• Exploration (Distribution Explorer, Multiplot, Insight, Association, Variable Selection, Link Analysis).
• Modification (Data Set Attributes, Transform Variables, Filter Outliers, Replacement, Clustering, SOM/Kohonen, Time Series).
• Model (Regression, Tree, Neural Network, Princomp/ Dmneural, User Defined Model, Ensemble, Memory-Based Reasoning, Two Stage Model).
• Verification (Assessment and Reporter).

In the above presented approach, data analysis is started by identification of the research problem, while further exploration is conducted on a data sample obtained from a larger data set. Then, relations between data are mostly looked for using data visualization tools (Explore) and the data set is prepared for modeling (Modification). Subsequently, the DM techniques are used to discover hidden knowledge (Model – the main processing phase). The last stage (Assess) is the evaluation of obtained results and attempt at their translation into real conditions of company functioning (post-processing phase). Sometimes, SEMMA is not considered as a methodology, but rather as a process model, focused on the core tasks of DM. Different studies indicate that the three methodologies described above were the most frequently used ones between 2007 and 2014.

2.1.5.4. Virtuous Cycle of Data Mining (VCDM)

The last methodology considered significant despite being rarely used, is known as the Virtuous Cycle of Data Mining (VCDM), proposed by Berry and Linoff as early as in 1997. It consists of four basic stages (35):

• Identify the business problem.
• Transform data into actionable result.
• Act on the information.
• Measure the results.

2.1.6. Text mining tools

There are more than 27 text mining tools but the commonly used ones are (36) :

1. **Rapid Miner** (formerly known as YALE): written in the java programming language, this tool provides advanced analytics through template-based frameworks. Users hardly have to write any code. Offered as a service, rather than a piece of local software. It also provides functionality like data preprocessing and visualization, predictive analytics and statistical modeling, evaluation and deployment.
2. **Weka**- usually used for data mining. It is a java based. This tool is very sophisticated and used in different applications including visualization and algorithms for data analysis and predictive modeling. Users can customize it however they like please. It supports data preprocessing, clustering, classification, regression, visualization and feature selection.

3. **R-programming**: it is primarily written in C and FORTRAN. And a lot of its modules are written in R itself. It is a free software environment for statistical computing and graphics. It provides linear, non-linear modeling, classical statistical tests, time-series analysis, classification, clustering and others.
4. **Orange**: a python-based powerful and open source tool for both novices and experts. It also has components for machine learning, add-ons for bioinformatics and text mining. It's packed with features for data analytics.
5. **GATE:** the General Architecture for Text Engineering. It is used for all sorts of language processing tasks and applications, including cancer research, drug research, decision support, recruitment, web mining, information extraction and semantic annotations.

2.1.7. Machine Learning

Machine learning is basically an intersection of elements from the fields of computer science, statistics, and mathematics, which uses concepts from artificial intelligence, pattern detection, optimization, and learning theory to develop algorithms and techniques which can learn from and make predictions on data without being explicitly programmed. The learning is about the ability to make computers or machines intelligent based on the data and algorithms which are provide to them so that they start detecting patterns and insights from the provided data. This learning ensures that machines can detect patterns on the data fed to it without explicitly programming them every time. The initial data or observations are fed to the machine and the machine learning algorithm works on that data to generate some output which can be a prediction, a hypothesis, or even some numerical result. Based on this output, there can be feedback mechanisms to our machine learning algorithm to improve our results. This whole system forms a machine learning model which can be used directly on completely new data or observations to get results from it without needing to write any separate algorithm again to work on that data.

There are many applications of machine learning in the real world such as

- Retail and e-commerce
- Advertising
- Filtering of spam emails and messages
- Fraud detection and prediction

- **Health care**: Machine learning algorithms are used widely in the healthcare vertical for more effective treatment of patients, disease detection and prediction and studying complex structures such as the human brain .

Each machine learning algorithm depends on what type of data it can work on and what type of problem are we trying to solve. Machine learning is categorized in to two groups: Supervised and unsupervised machine learning algorithms (37).

2.1.7.1. Supervised machine learning algorithms

The supervised learning algorithms are a subset of the family of machine learning algorithms which are mainly used in predictive modeling. The main types of supervised algorithms are classfication and regression (37).

2.1.7.2. Unsupervised machine learning algorithms

The unsupervised learning algorithms are the family of machine learning algorithms which are mainly used in pattern detection and descriptive modeling. These algorithms try to use techniques on the input data to mine for rules, detect patterns, and summarize and group the data points which help in deriving meaningful insights and describe the data better to the users. There is no specific concept of training or testing data here since we do not have any specific relationship mapping and we are just trying to get useful insights and descriptions from the data we are trying to analyze. The main types of unsupervised learning algorithms include clustering and association rule (37).

2.1.8. Clustering Algorithm

Clustering is an unsupervised machine learning task that automatically divides the data into **clusters,** or groups of similar items. It does this without having been told how the groups should look ahead of time that is, no prior knolwedge is needed. As we may not even know what we're looking for, clustering is used for knowledge discovery rather than prediction. It provides an insight into the natural groupings found within data. Clustering is guided by the principle that items inside a cluster should be very similar to each other, but very different from those outside. The definition of similarity might vary across applications, but the basic idea is always the same, group the data so that the related elements are placed together. The resulting clusters can then be used for further action. Overall, clustering is useful whenever diverse and varied data can be exemplified by a much smaller number of groups. It results in meaningful

and actionable data structures that reduce complexity and provide insight into patterns of relationships (38).

The choice of a suitable clustering algorithm and depends on the clustering objects and the clustering task. There are different clusteing stratgies (39):

1. Partitioning clustering
2. Hierarchical clustering
3. Advanced clustering which are further classified into:
   3.1. Fuzzy clustering
   3.2. Model-based clustering
   3.3. Density- based clustering
   3.4. Grid-based clustering

But the two standard clustering strategies are partitioning (K means) and hierarchical clustering.

2.1.8.1. Partitioning clustering

This clustering method is used to classify observations, within a data set, into multiple groups based on their similarity. Each partition represents a cluster. The algorithms require the analyst to specify the number of clusters to be generated. The commonly used partitioning clustering algorithms (39) are:

• **K-means clustering:** in which, each cluster is represented by the center or means of the data points belonging to the cluster. The K-means method is sensitive to anomalous data points and outliers.
• **K-medoids clustering** or **PAM** (Partitioning Around Medoids): in which, each cluster is represented by one of the objects in the cluster. PAM is less sensitive to outliers compared to k-means.
• **CLARA algorithm** (Clustering Large Applications): which is an extension to PAM adapted for large data sets.

The **k-means clustering** is perhaps the most commonly used clustering method. Having been studied for several decades, it serves as the foundation for many more sophisticated clustering techniques. It is a type of partitioning clustering algorithm which aims to partition the points into k groups such that the sum of squares from points to the assigned cluster centers is minimized (38).

The algorithm starts by randomly selecting k objects from the data set to serve as the initial centers for the clusters. The selected objects are also known as cluster means or centroids. The algorithm involves assignment and updating of the centroid. The cluster assignment and centroid update steps are iteratively repeated until the cluster assignments stop changing (i.e until *convergence* is achieved) (39) .

$$SSE(C) = \sum_{k=1}^{K} \sum_{Xi \epsilon Ck} |Xi - Ck|2$$

In order to determine the optimal number of clusters there are different strategies:

1. Sometimes the number of clusters is dictated by business requirements or the motivation for the analysis.
2. **One rule of thumb**: without any prior knowledge, one rule of thumb suggests setting *k* equal to the square root of *(n / 2)*, where *n* is the number of examples in the dataset. However, this rule of thumb is likely to result in an unwieldy number of clusters for large datasets and it is not very reliable.
3. Luckily, there are other statistical methods that can assist in finding a suitable k-means cluster set such as elbow method, cross validation, silhouette method and x means clustering;

A. **Elbow method:** attempts to gauge how the homogeneity or heterogeneity within the clusters changes for various values of *k*. The homogeneity within clusters is expected to increase as additional clusters are added; similarly, heterogeneity will also continue to decrease with more clusters. As it could be continued to see improvements until each example is in its own cluster, the goal is not to maximize homogeneity or minimize heterogeneity, but rather to find *k* so that there are diminishing returns beyond that point. This value of *k* is known as the **elbow point** because it looks like an elbow. With the elbow method the solution criterion value (with in groups sum of squares) tended to decrease substantially with each successive increase in the number of clusters.

B. **Cross Validation:** It's a commonly used method for determining k value. It divides the data into X parts. Then, it trains the model on X-1 parts and validates (test) the model on the remaining part. The model is validated by checking the value of the sum of squared distance to the centroid. This final value is calculated by averaging over X clusters. Practically, for different values of k, we perform cross validation and then choose the value which returns the lowest error.

C. **Silhouette Method:** It returns a value between -1 and 1 based on the similarity of an observation with its own cluster. Similarly, the observation is also compared with

other clusters to derive at the similarity score. High value indicates high match, and vice versa. We can use any distance metric (explained above) to calculate the silhouette score.

D. **X means Clustering:** This method is a modification of the k means technique. In simple words, it starts from k = 1 and continues to divide the set of observations into clusters until the best split is found or the stopping criterion is reached. But, how does it find the best split? It uses the Bayesian information criterion to decide the best split.

4. In most clustering applications, it suffices to choose a *k* value based on convenience rather than strict performance requirements (38).

2.1.8.2. Hierarchical Clustering

It is the clustering method by which the data are grouped together in form of trees. The hierarchical clustering is generally classified into two types of approach such as agglomerative approach and divisive approach. Agglomerative approach is the clustering technique in which bottom up strategy is used to cluster the objects. It merges the atomic clusters into larger and larger until all the objects are merged into single cluster. Divisive approach is the clustering technique in which top down strategy is used to cluster the objects. In this method the larger clusters are divided into smaller clusters until each object forms cluster of its own. Figure 3 below shows simple example of hierarchical clustering (40).

The output of hierarchical clustering will be a dendrogram, which is a tree-like diagram that shows the arrangement of the various clusters (38).
Here is the simplified description of how it works:

1. Assign each term to its own (single memeber) cluster
2. Find the pair of clusters that are closest to each other and merge them. So you now have one cluster less than before
3. Compute distances between the new cluster and each of the old clusters
4. Repeat step 2 and 3 until you have a single cluster containing all terms (41).

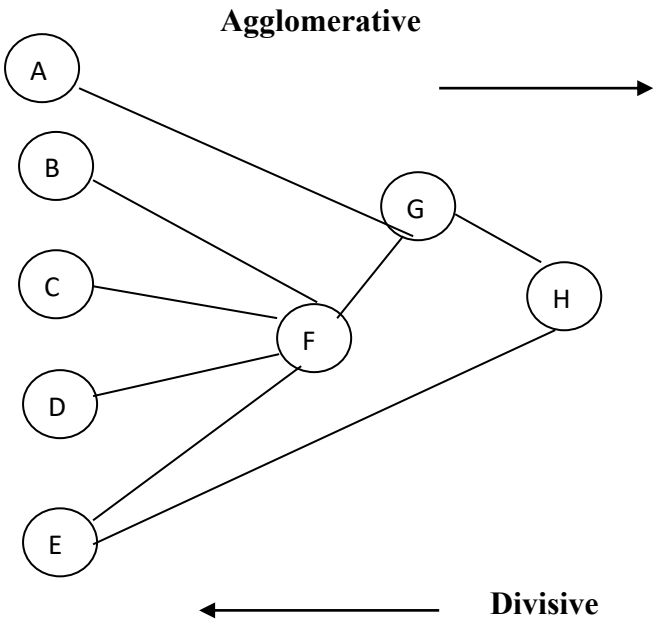


Figure 3. Types of Hierarchical clustering [(40)**: p.2]**

2.1.8.3. Fuzzy clustering

It is also known as soft clustering. Standard clustering approachesproduce partitions in which each observations belong to only one cluster. This is known as hard clustering. In this clustering, items can be a member of more than one cluster. Each item has a membership coefficients corresponding to the degree of being in a given cluster (39).

2.1.8.4. Model-based clustering

The data are viewed as coming from a distribution that is mixture of two or more clusters.It finds best fits of models to data and estimates the number of clusters (39).

2.1.8.5. Density-based clustering

It can find out clusters of different shapes and sizes from data containing noise and outliers. The basic idea behind density based clustering approach is derived from a human intuitive clustering method which is grouping objects into one cluster if they are connected to one another by densely populated area (39).

2.1.8.6. Grid-based clustering

Among the existing clustering algorithms, grid-based algorithms generally have a fast processing time, which first employ a uniform grid to collect the regional statistic data and then, perform the clustering on the grid, instead of the database directly. The performance of grid-based approach normally depends on the size of the grid which is usually much less than

the database. However, for highly irregular data distributions, using a single uniform grid may not be sufficient to obtain a required clustering quality or fulfill the time requirement. There are different types of grid based clustering technique such as Sting, Wave Cluster and Clique (40).

2.1.9. Distance measures in clustering

In order to perform a clustering algorithm, an important point should be disscussed first, that is distance. The notions of 'distance' and 'similarity' are related, since the smaller the distance between two objects, the more similar they are to each other. All measures refer to the feature values in some way, but they consider different properties of the feature vector. The distance measure reflects the degree of similarity or dissimilarity of the target documents and should correspond to the characteristics that are believed to distinguish the clusters embedded in the data. The nature of similarity measure plays a very important role in the success or failure of a clustering method; it is a very critical point in clustering (42). Similarity or distance measures are core components used by distance-based clustering algorithms to cluster similar data points into the same clusters, while dissimilar or distant data points are placed into different clusters (43). There are different methods for distance measures but the common ones are:

1. **Minkowski**: a classical distance measure which performs well when the dataset clusters are isolated or compacted Euclidean and Manhattan.
   Manhattan Distance is calculated as the absolute value of the sum of differences in the given coordinates. While Euclidean distance is the length of the line connecting the two points which is simply the sum of the squares of the differences between the two coordinates of the two points representing the terms. The distance between two terms (let's call them x and y) have coordinates (word frequencies) x1, x2, x3,…..xn and y1,y2,y3….,yn then one can say define the straight line distance also called Euclidean distance
   D(X, Y) =$\sqrt{(X1 - Y1)^2 + (X2 - Y2)^2 + \ldots.. + (Xn - Yn)^2}$ where n is the number of terms (41).
2. **Co relational:** widely used in clustering gene expression data. The distance between two objects is 0 when they are perfectly correlated. The correlational based distances are Pearson, Spearman, Kendall. Pearson measures the degree of a linear relationship between two profiles. Pearson's correlation is quite sensitive to outliers while the spearman correlation method computes the correlation between the rank of x and the

rank of y variables. Kendall and Spearman correlations are non-parametric and they are used to perform rank-based correlatinal analysis (39).

3. **Cosine:** mostly used in document similarity by calculating the *cosine of the angle* between their feature vectors.

There are ways of measuring the dissimilarity distance between two clusters which are called linkages. These are (44)

1. **Complete or maximum linkage:** computes all pairwise dissimilarities between the elements in cluster 1 and the elements in cluster 2 and considers the largest value (that is maximum value) of these dissimilarities as the distance between the two clusters. It tends to produce more compact clusters.

2. **Single or minimum linkage:** computes all pairwise dissimilarities between the elements in cluster 1 and the elements in cluster 2 and considers the smallest value (that is minimum value) of these dissimilarities as the distance between the two clusters. It tends to produce more compact clusters. It tends to produce long, “loose” clusters.

3. **Average linkage:** computes all pairwise dissimilarities between the elements in cluster 1 and the elements in cluster 2 and considers the average value of these dissimilarities as the distance between the two clusters. It tends to produce more compact clusters..

4. **Ward’s minimum variance method:** minimizes the total with in cluster variance. At each step the pair of clusters with the minimum between cluster distance are merged.

2.1.10. Approaches to Evaluation

There are two broad methodologies for evaluating clusterings: internal quality and external quality. Internal quality evaluates a clustering only in terms of a function of the clusters themselves. External quality evaluates a clustering using external information, such as an ideal clustering. When external information is available, external quality is more appropriate because it allows the evaluation to reflect performance relative to the desired output. There are three main approaches to evaluation using the external methodology: gold-standard, task-oriented, and user evaluation.

1. Gold-standard approaches manually construct an ideal clustering, which is then compared against the actual clustering.
2. Task-oriented approaches evaluate how well some predefined task is solved. Task-oriented methods have a bias towards the selected task.
3. User evaluation approaches involve directly studying the usefulness for users and often involve observation, log file analysis, and user studies similar to those carried out in the user evaluation of Grouper. User evaluation methods are very difficult to reproduce as they are dependent on the users. The large cost, and time involved in conducting good user evaluations is also a significant drawback. The lack of reproducibility, large cost, and time involved in conducting user evaluations makes them poor candidates for a standardized clustering evaluation method (45).

External Quality Measures

There are different external quality meausres. The most frequently used are:

1. **Rand Index:** compares the two clusters and tries to find the ratio of matching and unmatched observations among two clustering structures (C1 andC2).

Think of C1 as your predicted cluster output and C2 as the actual cluster output. The higher the value, the better the score. Its formula is given by

Rand index= (a+d)/ (a+b+c+d)

Where a= observations which were available in the same cluster in both structures (C1 and C2)

b=observations which were available in a cluster in C1 and not in the same cluster C2

c= observations which were available in a cluster in C2 and not in the same cluster C1

d= observations which were available in different clusters in C1 andC2 (46)

2. **Precision/Recall Measures:** this metric is derived from the confusion matrix.
   Precision is a measure of correctly extracted items while recall is a measure of matching items from all the correctly extracted items (45).
   Precision= TP/ (FP+TP) where TP is relevant texts extracted and FP is irrelevant extracted (47).
   Recall= TP/ (TP+FN) where FN is relevant texts omitted (47).
3. **F measure:** is more oriented toward measuring the effectiveness of a hierarchical clustering. It is a measure that combines the precision and recall ideas.
   F measure= (2*precision*recall) / (precision + recall) (46)
4. **Accuracy**: the term in which it measures the degree of proximity of a quantity to the quantity's actual true label values. In other words, it is defined as the number of exactly determined data objects of cluster results in contrast to the known true labels divided by the total number of instances in the dataset (46).

2.2. Related Literature

In (42), the researchers applied text mining process for pattern finding and clustering similar information from text so they mined the frequent terms from documents and generating the plot diagram for frequent terms using R Studio integrated development environment(IDE). The aim of the paper was clustering the documents using k-means, partitoning around mediods (PAM) and hierarchical agglomerative clustering (HAC) methods in R as statistical analysis tool and calculating the precision, recall and F-measure values for clusters and they compared the three clustering algorithms. They used Euclidean and cosine similarity methods to find the distance/similarity between clusters. For comparing they took the class labels documents from the 20 news group data set. After preprocessing, approximately 2% words of the corpus was used for clustering and pattern analysis. As their results indicate the partitoning around mediods(PAM) technique is better than the standard K-means approach and as good as or better than the hierarchical approaches that they have tested. And partitioning around mediods approach produced significantly better clustering solutions quite consistently according to F measure and overall similarity measures of cluster quality. And they were trying to implement the various probablistic and statistical models for selection of feature vectors from text corpus before applying text mining using R studio IDE.

In the study titled text mining in health care application and opporutnities, it was explained that domain knowledge plays a vital role in extraction of knowledge from text. The purpose of this project was evaluating the applicability of text mining in the domain of clinical records. In order to achieve their objective, they made a collaboration between University of Alabama at Bringngham and University of Alabama to explore the applications of text mining in electronic clinical records. SAS provided the software support needed for the project. University of Alabama at Bringngham data set contained patient diagnosis and prescription information. These record contained vast amounts of unstructured data that was typically overlooked due to the immense size of the medcial record collection at the hospital. Each medical record had numerous different documents making a single complete record very large and cumbersome to work with. So they narrowed their analysis to pathology reports and discahrge summaries. The project began with the pathology report form of set of 1500 records. Clustering of the documents specifically the term weight entropy was used to discover the patterns that existed in the textual contents after removing any noise or skewness. So the documents were clustered as cytologies, bone marrow biopsies, kidney pathologies, tumors and thin prep cytologies. And from the 15,000 discharge summaries they extracted

medication information, they clustered the medications as antibiotics, muscle relaxants and heart medications. As their results showed they were able to group patients with similar conditions together. At last they concluded that text mining can be an effective tool in health care data sets. With the shift to electronic clincal records and availabilty of standardized vocabulary the future of text mining in the health care domain is bright. And it is a great deal to create predictive models using patient records that will significantly reduce medical errors and patient dying from these medical errors (48) .

Another study focused on using unsupervised machine learning to distinguish between actual suicide notes and newsgroups. The purpose of the paper was to describe one experiment in a series of experiments to develop a tool that combines Biological Markers (Bm) with Thought Markers (Tm), and use machine learning to compute a real-time index for assessing the likelihood repeated suicide attempt in the next six-months. Their main focus was distinguishing between actual suicide notes and newsgroups using the unsupervised machine learning. They particularly wanted to determine if modern machine learning methods could be applied to free-text from those who committed suicide.The findings showed that mental health professionals accurately selected genuine suicide notes 50% of the time and the supervised machine learning methods were accurate 78% . They shifted from supervised to unsupervised machine learning methods. The rationale for the study, then, was that since the ultimate goal was to create a Suicide Risk Index that incorporates biological and thought markers it was important to determine if unsupervised methods can distinguish between suicidal and nonsuicidal writings. They used 866 suicide notes and 4000 newsgroups to develop the corpus. Basic statistics were calculated using variables extracted by Linguistic Inquiry and Word Count version 2007 software (LIWC2007). Calculations were done using open source software called *R*. Clustering was done with the following algorithms: expectation maximization (EM), simple k-means with euclidean distance (SKM), and sequential information bottleneck algorithm (sIB). The last approach has been shown to work well when clustering documents. Specificity, sensitivity and F1 measure were used as performance measures. Multidimensional scaling with euclidean distance measures was used for visualization purposes. To extract features that represent each cluster, Pearson correlation coefficient was used. In general sequential information bottleneck (sIB) clustering algorithm performed best for all data sets with respect to F1 measure. The average score also did not change when the number of clusters varied from two to six. Performance of k-means and expectation maximization algorithm was much worse. Their findings suggested that unsupervised methods can distinguish between suicide notes and newsgroups (49).

The study in (40) showed the vast amount of hidden data in huge databases has created tremendous interests in the field of data mining. Their paper discussed the data analytical tools and data mining techniques to analyze the medical data as well as spatial data. Spatial data mining includes discovery of interesting and useful patterns from spatial databases by grouping the objects into clusters. Recently many commercial data mining clustering techniques have been developed and their usage is increasing tremendously to achieve desired goal. Researchers are putting their best efforts to achieve the fast and efficient algorithm for the abstraction of spatial medical data sets. The proposed work focused on challenges related to clustering on medical spatial datasets. Clustering is the unsupervised classification of patterns into clusters. They prepared series of experiments to determine relevant pattern detection for medical diagnosis. The dataset consisted of number of cancer patients. The database analysis was done using TANAGRA tool kit that has several data mining software for data analysis, statistical tools in data base This paper focused on clustering algorithms such as HAC (hierarchial agglomerative clustering) and Kmeans in which, hierarchial agglomerative clustering was applied on K-means to determine the number of clusters. The quality of cluster was improved, if hierarchial agglomerative clustering was applied on K-means. The application can be used to demonstrate how data mining technique can be combined with medical data sets and can be effectively demonstrated in modifying the clinical research. The experimental results showed that there are certain facts that are evolved and can not be superficially retrieved from raw data.

In (50), different data clustering algorithms were compared. The algorithms under investigation were: K-means algorithm, Hierarchical Clustering algorithm, Self-Organizing Maps (SOMs) algorithm, and Expectation Maximization (EM) Clustering algorithm. All these algorithms were compared according to the following factors: size of dataset, number of clusters, type of dataset and type of software used. The dataset that was used to test the clustering algorithms and compare among them was obtained from a web site The comparison between the four algorithms were done using both large and small datasets and found the following conclusions: as the number of clusters, k, became greater, the performance of SOM algorithm became lower; the performance of K-means and EM algorithms were better than hierarchical clustering algorithm; SOM algorithm showed more accuracy in classifying most the objects into their suitable clusters than other algorithms; as the value of k became greater, the accuracy of hierarchical clustering becam better until it reached the accuracy of SOM algorithm; K-means and EM algorithms had less quality (accuracy) than the others; all the algorithms had some ambiguity in some (noisy) data when clustered; the quality of EM and K-means algorithms became very good when using huge dataset while hierarchical clustering

and SOM algorithms showed good results when using small dataset; hierarchical clustering and SOM algorithms gave better results compared to K-means and EM algorithms when using random dataset and the vice versa; K-means and EM algorithms were very sensitive for noise in dataset. This noise made it difficult for the algorithm to cluster an object into its suitable cluster; hierarchical clustering algorithm was more sensitive for noisy dataset than SOM algorithm; running the clustering algorithms using any software gives almost the same results even when changing any of the factors because most software use the same procedures and ideas in any algorithm implemented by them. The recommendation made was to compare between these four algorithms (or may other algorithms) can be attempted according to different factors other than those considered in this study. One important factor is normalization. Comparing between the results of algorithms using normalized data or non-normalized data would give different results. Of course normalization would affect the performance of the algorithm and the quality of the results.

Most of the literatures reviewed tried to show that lots of text mining projects in the cancer medical domain have been performed internationally using classification methods especially in the biogenetic discipline and document clustering which is different from text clustering. Little has been done on the area of cancer medical records using clustering algorithms which help in knowledge discovery. As can be seen from these literatures hierarchical clustering performed better if the data set was small otherwise k means was better when the two algorithms were compared. Since most of the medical records were electronic, preprocessing of the texts and extraction of the pattern would be a little bit easier. And the texts in electronic medical records were semi structured whereas the texts in Tikur Anbessa specialized hospital are unstructured that are difficult and tiresome to handle and preprocess.

3. METHODOLOGY

Text mining applications are so broad in their scope and have different goals that it is difficult to express the accomplishment of it in general terms. Compared to other well-established statistical methods, text mining is a relatively new and unstandardized analytical technique for knowledge discovery. Therefore, it is challenging to create a road map of operations to perform its methodology (6).

As discussed earlier in section 2.1.5 CRISP-DM is the most referenced and used in practice of data mining methodologies. It also provides the most complete tool set to date for data mining practitioners so I used CRISP-DM. According to (6), the primary distinction between data mining and text mining is simply the type of data involved in the knowledge discovery process. That is, text mining is extraction of data from unstructured formats which is text whereas data mining is extracted from databases. Thus, the use of a data mining methodology for the purpose of undertaking a text mining project is still appropriate. CRISP-DM has an iterative cycle with six phases: business understanding, exploration of data availability, data preparation, model development and assessement and result evaluation and deployment.

3.1. Determine the Purpose of the Project/Business Understanding

The “business understanding” phase of CRISP-DM involves establishing the business case and aims of the project.

In order to understand the business, informal interviews were conducted with the head nurse, record room clerk and the medical residents who were working in the oncology department of Tikur Anbessa specialized hospital.

The project was conducted at Tikur Anbessa Specialized hospital oncology unit by collecting the data between April and July 2017. This hospital was chosen because it is the first and teaching hospital with both radiotherapy and chemotherapy for all cancer cases in Ethiopia. Actually St. Paulos hospital and some private hospitals like Girum hospital have started giving only chemotherapy for all cancer cases and Zweditu Memorial hospital has started giving chemotherapy for cervical cancer patients. The oncology unit of Tikur Anbessa specialized hospital has its own record room where patients’ medical records are being kept. Here, there are more than five thousands patients’ medical records and only two staffs to manage these records which are making retrieval of patients’ records very difficult.

For an oncology patient to be seen in this hospital, the patient needs to have a referral paper. With the referral paper at hand, first the patient needs to be seen in the adult outpatient department (OPD). In the outpatient department, medical residents take full patient history and do some necessary investigations. If all the investigation results confirm the cancer, the patient will be directly referred to the oncology unit. Else the patient needs further investigation so the patient will be referred to either to the surgical department or gynecology department. Then after having removed the affected body part the patient will be sent to the oncology unit to start either chemotherapy or radiotherapy or both. Since the oncology unit has a limited number of oncologists, each cancer patient needs to be seen by a medical resident. The medical resident examines the patient and decides for the patient to have either chemotherapy or radiotherapy or both. Since there are a lot of patients enrolled in the oncology unit of the hospital, patients who need the radiotherapy have to be booked first to get the treatment. Also patients who need chemotherapy have to wait for their turn to get the bed.

Whenever there is a difficult case the resident will either consult the oncologists or search for a previous similar case in order to share other physicians' experience on the case so that the patient can start the necessary treatment. But all these processes are tiresome and difficult because the physicians write the patient's progress, the investigation results and the treatment plan wherever they like that is wherever free space is found. Since the case (cancer) needs a thorough investigation to confirm the disease, its stage and also to follow the patient's health status and performance level, the patient needs to have a lot of investigation results which make the medical record bulky. Because of all these some of the necessary patient history may be missed.

Even some times the patient may take his record to other units or departments for other related cases like ART (anti retroviral therapy) follow up and his record may not be returned. Because of all these problems the patient may lose his record and needs to have a new one this will in turn create a big problem in the patient's care.

So as mentioned in chapter one the purpose of the study is to apply text mining techniques to extract knowledge from oncology patients' medical records.

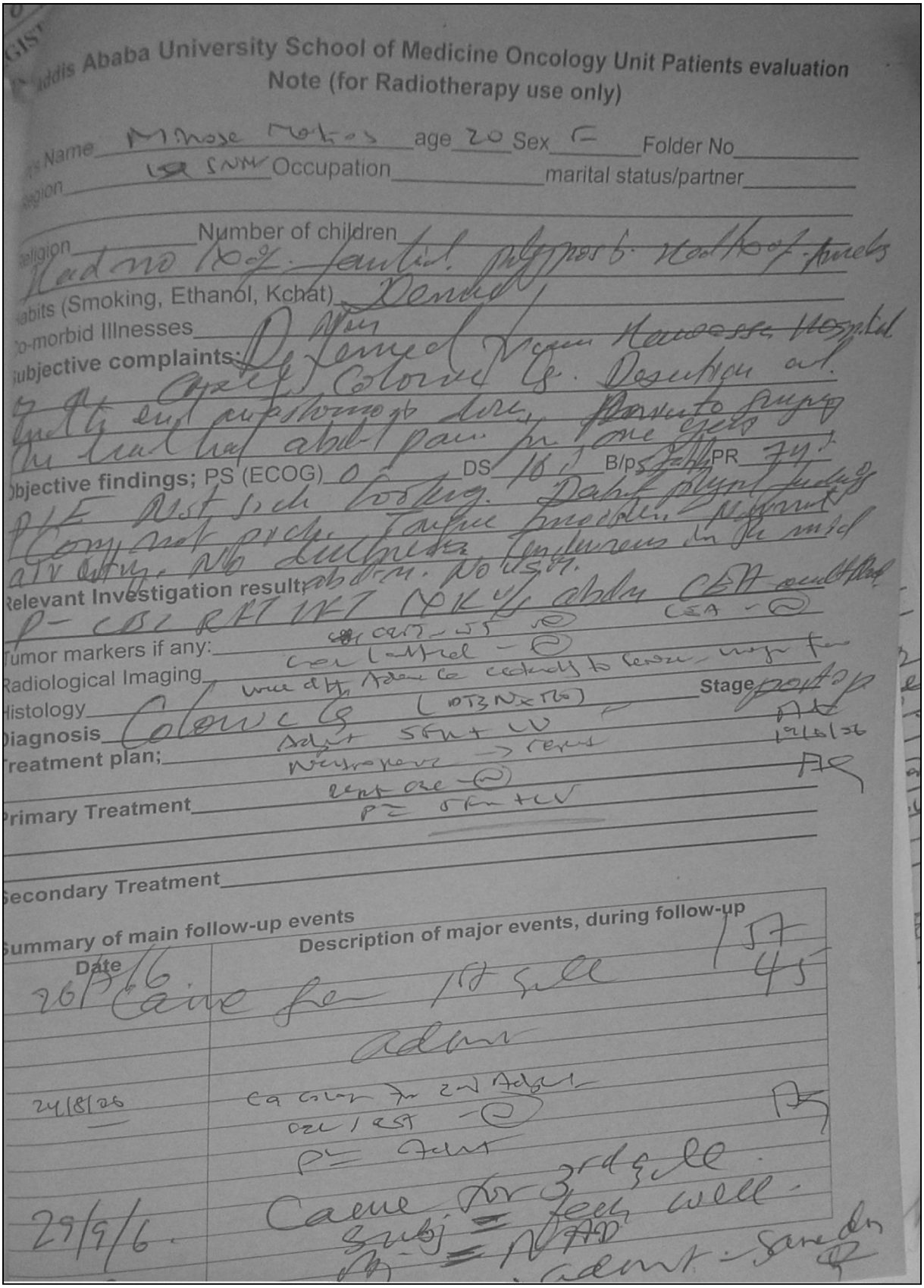
3.2. Explore the Availability and the Nature of the text/text understanding

The text understanding phase starts with an initial data collection. Then it will be proceed to increasing familiarity with the data, identifying data quality problems, discovering initial

insights into the data, or detecting interesting subsets to form hypotheses about hidden information.

Since the record management system of the hospital oncology unit was not yet automated, it was very difficult to access the necessary data because the medical record was very bulky with lots of investigation papers, history sheets and the oncology's own different format papers. Moreover, there were missed papers from the patient's medical records which were important part. As known physician hand writing's are illelligible, so it was very difficult to extract the necessray information from the patient medical records.

The data was collected by the principal investigator using camscanner of a tablet, to take pictures of the necessary information. For this project, it was very difficult to take a sample size because the total number of medical records found in the oncology unit was not known. The sampling technique used is non probabiltiy sampling specifically, purposive sampling. Using the inclusion and exclusion criteria, I checked 600 medical records out of which, only 137 records were collected and used in the corpus (collection of texts) because the rest of the records have missed texts and some have characters which were difficult to recognize and read. As anyone can see in figure 4 most of the medical records have illegible handwriting and texts difficult to interpret.



Addis Ababa University School of Medicine Oncology Unit Patients evaluation Note (for Radiotherapy use only)

Name Minose Matios age 20 Sex F Folder No

Region SNM Occupation marital status/partner

Religion Number of children

Habits (Smoking, Ethanol, Kchat) Denied

Co-morbid Illnesses

Subjective complaints: Deferred from Hawassa Hospital

Objective findings; PS (ECOG) 0 DS 16 B/p PR 74

Relevant Investigation result:

Tumor markers if any:

Radiological Imaging

Histology Stage

Diagnosis Colonic Ca

Treatment plan;

Primary Treatment

Secondary Treatment

Summary of main follow-up events

Date	Description of major events, during follow-up
26/8/6	Came for 1st cycle
24/8/26	Came for 2nd Adjuvant
29/9/6	Came for 3rd cycle

Figure 4. Snapshot of the oncology medical record format: *this figure shows the oncology patient medical record format*

3.3.Preparing the text

The text preparation phase covers all activities to construct the final data set or the text that will be fed into the modeling tool(s) from the initial raw data.

In order to prepare the text, R version 3.4.0 was used, which is both a language and environment oriented towards statistical computing and graphics creation. It is also popular in a lot of text mining projects and researches. *"Due to its extensibility and versatility, R has remained consistently popular for data and text mining applications across many domains"* (51).

All R functions and datasets are stored in packages. Only when a package is loaded are its contents available. There are two types of packages: standard and contributed packages. The standard (or base) packages are considered part of the R source code. They contain the basic functions that allow R to work, and the datasets and standard statistical and graphical. There are thousands of contributed packages for R, written by many different authors. Some of these packages implement specialized statistical methods, others give access to data or hardware, and others are designed to complement textbooks. Some (the recommended packages) are distributed with every binary distribution of R. Most are available for download from comprhensive R archive network(CRAN) and other repositories such as Bioconductor and Omegahat (52). So some of the contributed packages are dowloaded from https://CRAN.R-project,org/package-tm like NLP, snowballC, ggplot2, wordcloud and Rgraphviz. The primary package for text mining in R is tm.

3.4.Extraction of Patterns

In this phase, various modeling techniques are to be selected and applied and their parameters are calibrated to optimal values.

After preprocessing the texts, the clustering algorithm was used in order to extract the pattern. There are different clustering algorithms but most commonly used ones are K means and hierarchical clustering algorithm.

3.5.Evaluation of the model

In the evaluation of the clustering algorithms, both subjective and objective methods were used. For the objective method a gold standard approach which uses the manual construction of an ideal clustering which is then compared against the actual clustering. This approach used

accuracy, rand index, presicion, recall and F measures to compare the pattern extracted using the algorithms mentioned. While the subjective evaluation was done by interviewing ten phyicians from the oncology department.

3.6.Inclusion criteria of the project

Any medical record with eligible hand writing and complete information such as the primary and secondary treatment plan (patients who were booked for radiotherapy, patients who took chemotherapy and were still taking chemotherapy) were included in the project.

3.7.Exclusion criteria

Any medical record with illegible hand writing, missed patient history sheet, incomplete data and newly diagnosed patients who were waiting for the physicians decisions (patients' who were diagnosed after April 2017 because they haven't yet started neither chemotherapy nor booked for radiotherapy during the data collection period) were excluded from this project.

4. DISCUSSION OF RESULTS

In the previous chapter, the methodology used for the performance of this project is discussed and described. As mentioned earlier R software was used for the preprocessing of the text and extraction of the patterns. In this section, pre processing, extraction of the pattern and evaluation of the model will be discussed in detail.

4.1. Prepare the Text

During the collection of data, patient privacy has been kept; neither the patient name nor his medical record number has been mentioned.

"*Corpus is the main structure for managing documents"*. It repesents a collection of text documents and can be considered as a database for texts (53). The quality and quantity of the data are the most important elements of both data mining or text mining projects (18) All the texts collected, which were used in the corpus, were typed into the computer because text fairy, optical character recoginzer, couldn't be able to recognize the characters. And these typed records changed to txt files to facilitate working with R. The last step was changing all the abbreviations in to text, again for this process phyton was tried but some words have more than one interpretation for example

1. m stands for male and also metastasis
2. PR stands for pulse rate and per rectal
3. CT stands for chemotherapy and computed tomography etc.

So inorder to solve this prblem, all the abbreviations are changed manually by going through all the texts accordingly.

The data preparation step in R starts with the corpus establishment. After calling the library tm and library Natural language processing (NLP) which is used for data preparation, the corpus was esatblished.. The corpus, for this project, contains 137 narratives. This corpus was clinical narratives of the cancer patients.

The process of standardizing the many different file types to a common text-based representation is called preprocessing. Preprocessing can also mean adding to or modifying the text to make it more usable by a library. For instance, certain libraries may expect the input to already have sentences identified. Ultimately, preprocessing includes any steps that must be undertaken before inputting data into the library or application for its intended use (54).

The assumptions made here is that the "meaning" of a document can be represented with a list and frequency of the terms used in that document. Some terms, such as articles, auxiliary verbs, and terms used in almost all of the documents in the corpus, have no distinguishing power and therefore should be excluded from the indexing process. With in each corpus ,each text was treated as separate record. To perform text transformation tm_map() function was applied from the tm package. This step involved

I. Removing punctuation and numbers
II. Remove unwanted characters like ",@,#,&, in the text data by applying the **gsub** transformation (55)

    These two steps I and II are called tokenization
III. Removing stop words, these are words which have no importance for the process and called english stop words. They are commonly known stopwords (the , has , any, that, am,in, for etc)
IV. There were some terms which are found on the radiotherapy format which were redundant in every text format which will be a great problem during the data analysis and have no significance. These terms (age, sex, marital status, religion, region, comorbid illnesses, habits, performance status, daignosis specific etc**)** also needs to be removed by calling remove mystopwords(). In this step we need to specify the words to be removed by ourselves. So I specifically picked out these words in order to reduce redundancy, avoid complication and get a true result.

    Step III and IV are called filtering (56)
V. Changing texts in to lower cases, stripping white spaces and
VI. Stemming words. The stemming algorithm stemDocument included in the R package tm where the function calls the porter stemming algorithm that removes common word endings for English words, such as "es", "ed" and "'s" (38).

4.2. Extract the Knowledge/Modeling

After Completing the transformation of the corpus, the next step was developing a model using the clustering algorithm. In order to use this algorithm, one has to create either a **Document-Term Matrix** (**DTM**) or **Term-Document Matrix** (**TDM**). This step helps in changing the preprocessed texts in to structured formats. Document-Term Matrix (DTM) or Term-Document Matrix (TDM) has the same purpose, it is only a matter of putting the word counts and the individual matrix either in a row or in a column, the result in both way is the same. So for this project, the TDM was used. A TDM would have the documents as columns and the words as

row, it is the transposition of DTM. The relationships between the terms and the documents are characterized by indices, which are relational measures, such as how frequently a given term occurs in a document. In the corpus created, there were 137 documents and 1429 terms. As everything set up in the term-document matrix, then the next step was exploring the most frequent terms (terms with the highest occurrences frequencies) by creating an object with the row sums. I then converted the term document matrix to a simple matrix for writing to a CSV file, for example, for loading the data into other software in this case in to excel. Once converted into a standard matrix the usual write.csv() can be used to write the data to file (55) (38).

	Documents								
Terms	10.txt	100.txt	101.txt	102.txt	103.txt	104.txt	105.txt	106.txt	107.txt
Abdomin	2	1	0	1	2	1	2	2	2
Adjuv	2	0	0	0	0	0	0	0	0
Admit	1	0	0	0	1	0	0	0	0
Antigen	1	0	0	0	1	1	0	2	0
Bluish	1	0	0	0	0	0	0	0	0
Came	3	0	0	0	0	0	0	0	0
Cancer	2	3	0	3	1	0	1	1	0
Carcinoembryon	1	0	0	0	0	1	0	1	0
Caviti	1	0	0	0	0	0	0	0	0
Chemotherapi	1	0	0	0	0	0	0	1	0
Colon	2	0	0	0	0	0	0	0	0
Cycl	3	0	0	0	0	0	0	0	0
Deni	1	0	0	0	0	1	0	0	0
Diagnos	1	0	0	0	0	0	0	0	0
Diagnosi	1	2	2	2	0	2	2	2	2
Examination	1	0	0	0	0	0	0	0	0
Femal	1	0	1	1	1	0	0	0	0
First	1	0	0	0	0	0	0	0	0
Fluorouracil	2	0	0	0	1	0	0	0	0
Follow	2	2	2	2	0	2	2	2	2

Table 1. Term-document matrix: *sample of term document matrix that was saved in the excel using write.csv () function*

Term by document matrices are generally quite sparse – they contain lots of zeros. Terms which occur only once or twice are likely to consume a lot of computational resources without adding anything useful to the analysis. So there is a need to reduce the size of the tdm without losing much useful information. Such sparse terms can be removed from the document term matrix quite easily using removeSparseTerms () function (55).

When I checked the sparsity of the terms, I got 96% (0.96). The sparsity parameter of 0.96 says to remove all terms from the tdm with zeros for the term count in 96% of the documents. When the remove sparse terms function was applied to the data, the number of terms was reduced from 1429 to 246. Again when the remove sparse terms function with the parameter of 0.85 was applied to the data, the number of terms was reduced from 246 to 64.

One thing usually done first is to get an idea of the most frequent terms in the corpus and these terms are rated important. So I used the function findFreqTerms () (55) to get the most frequent terms in my corpus. To show one example, I limited the output to those terms that occur at least 40 times. Then the result showed the table below:

Word	Freq
Cancer	196
Mass	114
Radiotherapi	104
Breast	98
Carcinoma	89
Abdomin	86
Book	84
Pain	71
Cervic	69
Cell	65

Table 2. The most frequent terms that occur at least 40 times in the corpus developed

Using the ggplot2 library the most frequent words mentioned above are depicted using a bar graph as shown below.

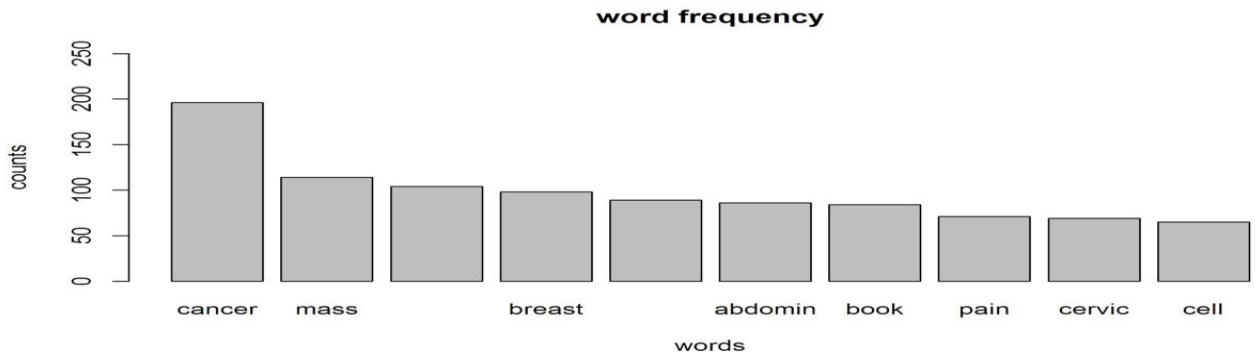


Figure 5. Word frequency

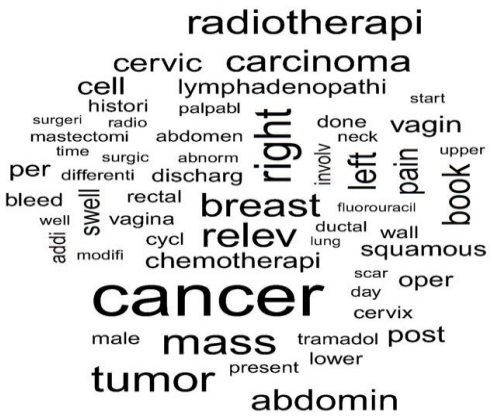


Figure 6. Word Cloud

Then I depicted a word cloud, an effective alternative to providing a quick visual overview of the frequency of words for example at least 40 times in a corpus. The word cloud () package provides the required function (57).

After having the most frequent terms in the corpus, the next step is modeling or extracting the necessary knowledge. In order to extract the knowledge in text mining, there are different technologies: classification, clustering, feature extraction, association, regression and anomaly detection (58). And for this project, clustering algorithm was used. There are five clustering methods namely partitioning, hierarchical, fuzzy, density based and model based. For identifying groups of related terms the key clustering methods are hierarchical and k means

clustering. Also these two algorithms are more applicable and popular in the medical domain of text mining (59). For computing these algorithms library stats was used (55). The clustering was performed on the terms which passed the process of sparsity which were 64 terms.

4.2.1. K means Clustering

The k-means algorithm begins by choosing *k* points in the feature space to serve as the cluster centers. The procedure was used to break out the term-document matrix into subsets based on clusters. With k-means clustering the user must specify "k", the number of clusters, as well as the distance measure (55). The algorithm essentially involves two phases. First, it assigns examples to an initial set of *k* clusters. Then, it updates the assignments by adjusting the cluster boundaries that currently fall into the cluster. The process of updating and assigning occurs several times until changes no longer improve the cluster fit. At this point, the process stops and the clusters are finalized (38).

Determining the optimal number of clusters (k)

Before initiating this algorithm, the first step is choosing k which is the number of clusters which serves as the center of the clusters. There are different ways for determining the number of clusters:

In order to choose an optimal number of clusters, I used one of the statistical method which is the elbow method. Simplistically, an optimal number of clusters was identified once a "kink" in the line plot was observed (42).

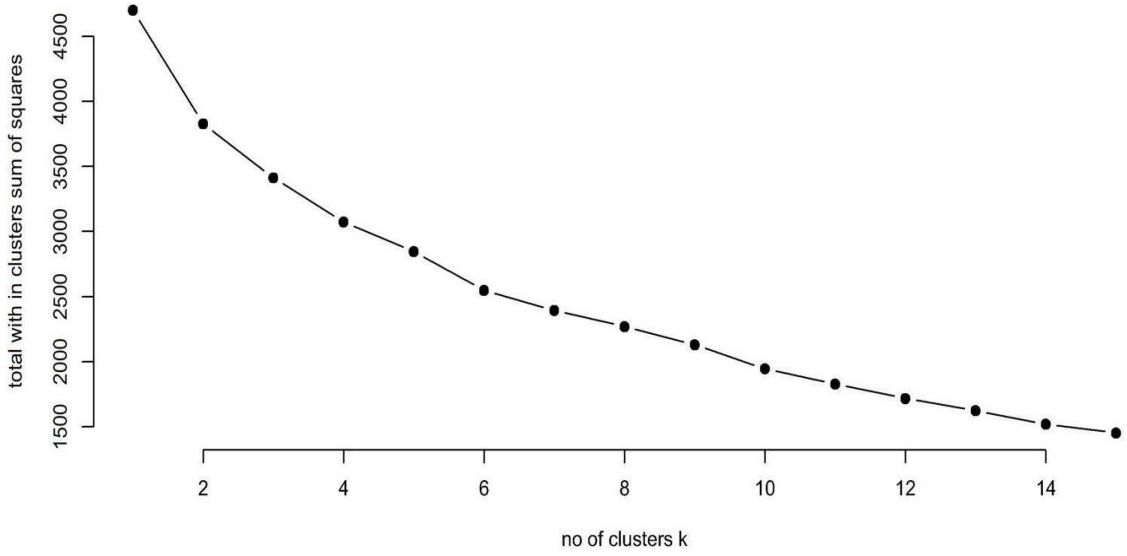


Figure 7. Optimal number of clusters(k)

As can be seen in figure 7, it was easy to say that after 6 clusters the observed difference in the with in cluster dissimilarity is not substantial. Consequently, I could say with some reasonable confidence that the optimal number of clusters to be used was 6.

After determing the optimal number of clusters, the distance should be calculated using the Euclidean distance method which is the default one. This calculated distance will also be used in the hierarchical clustering method. Euclidean distance between an observation and initial cluster centroids 1 and 2 is calculated. Based on Euclidean distance each observation has assigned to one of the clusters based on minimum distance. Euclidean distance is widely used in clustering problems including clustering text (58).

Because the k-means algorithm utilizes random starting points, the set.seed() function was used to ensure that the results match the output (38). In the absence of this statement, the results would vary each time the k-means algorithm ran:

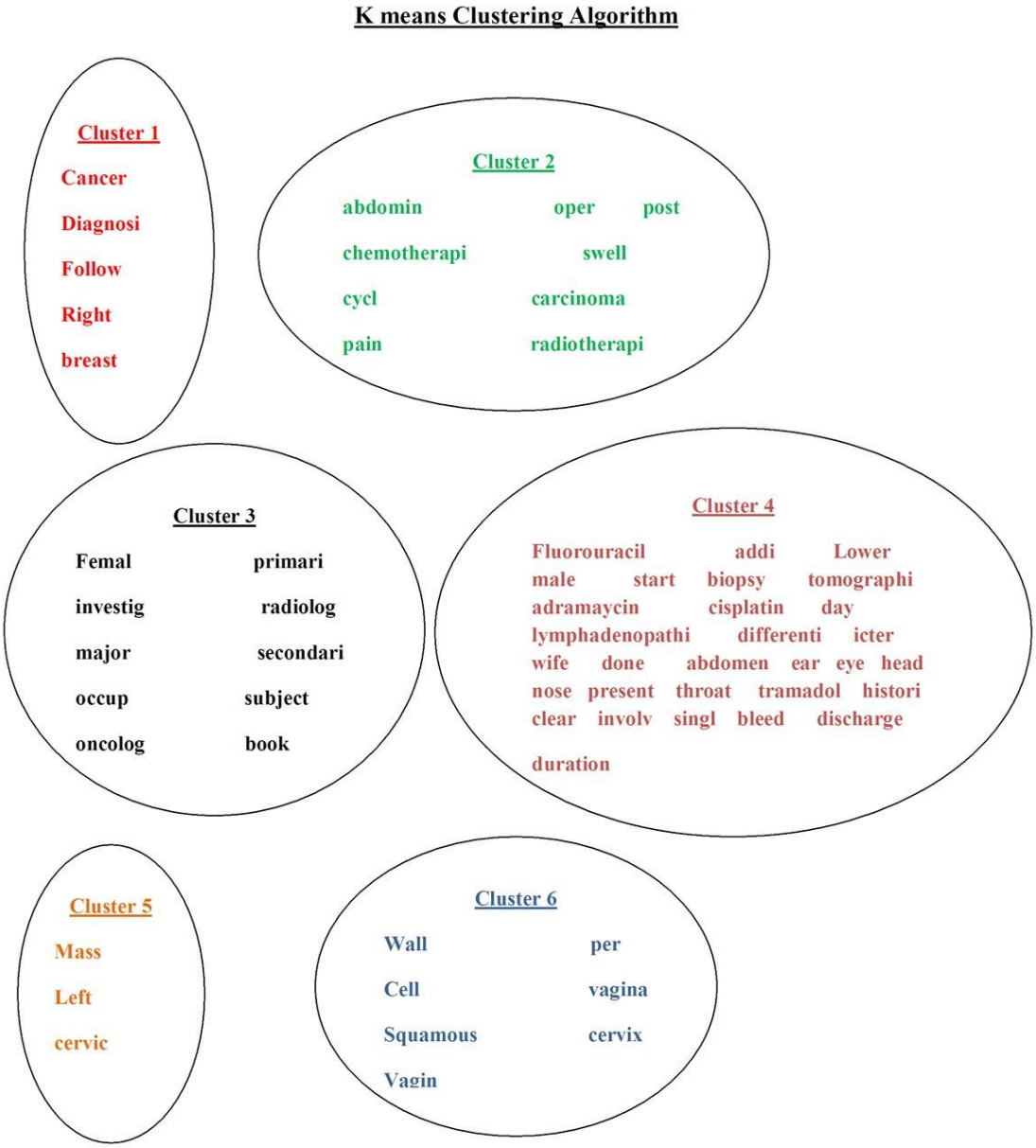


Figure 8. kmeans algorithm: *this figure is drawn to show the clusters found during the computation of the k means algorithm and drawn manually to show the kind of similarities they have.*

4.2.2.Hierarchical clustering

Hierarchical clustering is an agglomerative or bottom-up technique. It means that all observations are their own cluster. From there, the algorithm proceeds iteratively by searching all the pairwise points and finding the two clusters that are the most similar. So, after the first iteration, there are *n-1* clusters and after the second iteration, there are *n-2* clusters, and so forth. As the iterations continue, it is important to understand that in addition to the distance measure, we need to specify the linkage between the groups of observations (38).

Different types of datasets will demand that we use different cluster linkages namely ward, complete, single and average. Ward linkage minimizes the total within-cluster variance as measured by the sum of squared errors from the cluster points to its centroid and is the most commonly used one and it is a default in R, so ward linkage was used in this project. To bulid a hierarchical cluster model in R, I used the hclust () function in the base stats package. The two primary inputs needed for the function are a distance matrix and the clustering method (38).

The distance matrix was computed with the dist () function and distance name Euclidean which is usually considered as the default.

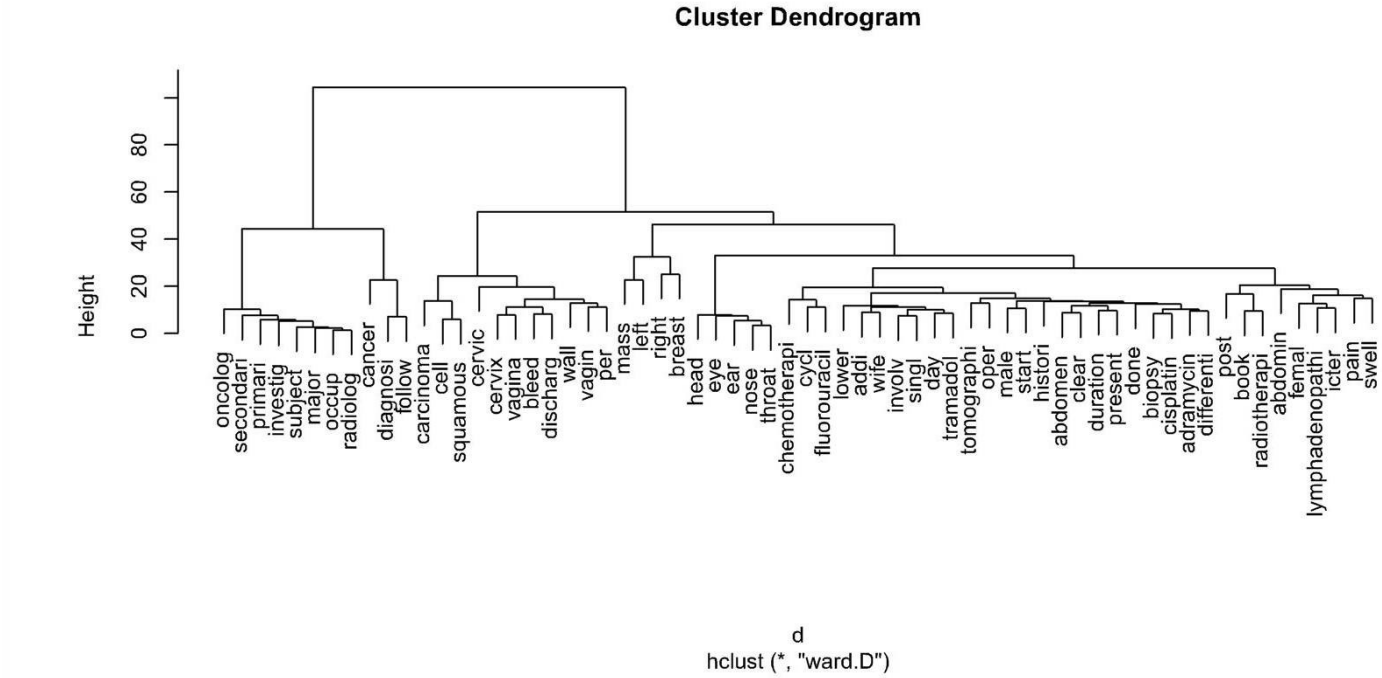


Figure 9. Hierarchical clustering: *this figure shows the result of the project with 64 objects(texts)which are stemmed (texts which seem misspelled are stemmed texts)*

In figure 9, each branch point encountred is the distance point at which a cluster merge occurred. Clearly the most well-defined clusters were those that have the largest separation, many closely spaced branch points indicate a lack of dissmislarity (distance) between clusters (41).

In order to help in the evaluation process of the hierarchical with the k means algorithm cut tree() function which cuts the hierarchical algorithm in to 6 clusters which is the same number of clusters as that of the k means was used. And a very nice tool for displaying more appealing trees is provided by the R package ape so that any one can differentitate the clusters with various colors:

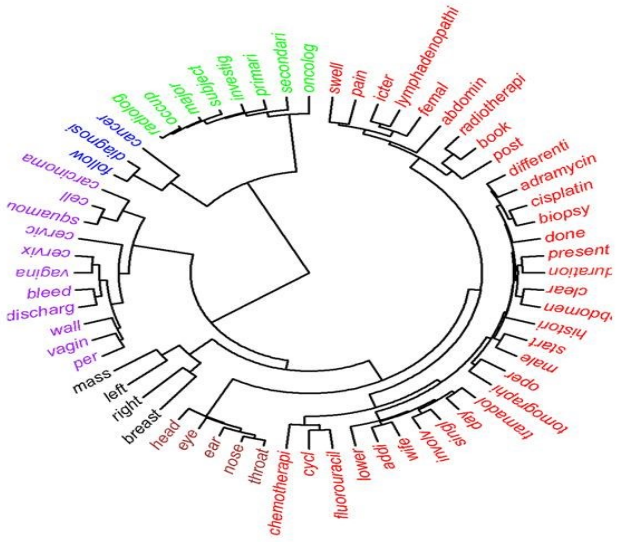


Figure 10. Fan Shaped hierarchical clustering: *this figure clarifies the pattern of each clusters using different colors*

As can be in the figures (8 and 9), the clustering algorithms give the following results:

1. The most cancer cases found in the corpus are breast and cervical cancers which involve both the right and left breast
2. Chemotherapy and cycle are clustered in one cluster, radiotherapy and book are clustered in another
3. Fluorouracil, cisplatin and adramycin are the chemotherapy drugs found in the corpus and tramadol
4. Sign and symptoms of the cervical cancer including vaginal bleeding and vaginal discharge are clustered together
5. The commonly used computed tomography (CT) and biopsy are clustered together which are used as a diagnosis method.

4.3.Evaluation of the extracted knowledge

There are two main approaches to evaluation using the external methodology:

1. **Objective Evaluation:** this method involves the availablity of gold standards.The gold-standard approach manually constructs an ideal clustering which is then compared against the actual clustering (45)**.**
2. **Subjective Evaluation:** involves random selection of a set of results and manual evaluation of the selected samples by the help of domain experts in this case the oncologists and medical residents (60).

4.3.1.Subjective Evaluation

In this phase, the subjective evaluation of the hierarchical and k means clustering algorithm is done by interviewing the oncologists and medical residents to which algorithm shows the best result. That is, which one of the two algorithms extract the best pattern or knowledge from the oncology patients' medical records. For this purpose, I have interviewed two oncologists and eight medical residents. Out of the ten physicians, three of them (30%) did not have any idea as which one is better because the idea was new to them. And it was difficult for them to understand the pattern. Two of the physicians (20%) prefer the k means to the hierarchical algorithm because the hierarchical algorithm seems complicated to them. Because the algorithm tried to show almost the necessary pattern that was found in the patient' medical records, the rest (50%) of the physicians chose the hierarchical algorithm. That is, the commonly used chemotherapy medications (fluorouracil, cisplatin and adramycin) and most commonly used anti pain (tramadol) which is given once a day are clustered in one cluster and the commonly occurring cancers in female are the breast and cervical cancer in different clusters. And the hierarchical algorithm showed that most of the cervical cancer patients have vaginal bleeding and discharge. It also showed that most of the cervical cancers are squamous cell carcinoma and for a patient to get radiotherapy treatment he or she needs to be booked first. All these things are grouped according to their similarity in accordance to the patients' medical records.

Whereas in case of the k means algorithm, for example,

1. Book and radiotherapy are in different clusters
2. Carcinoma and squamous cell are in different clusters
3. Chemotherapy and the medication used are again in different clusters

4. Bleeding, discharge and vaginal are in different clusters which raise a question as to where this bleeding and discharge has come

All these things may create a difficulty of understanding as to what kind of pattern does exist in the medical records because most of the terms mentioned above were supposed to be clustered together.

4.3.2. Objective Evaluation

During this evaluation the first step is developing the cluster manually by the principal investigator, has a medical background, then this cluster was subjectively evaluated by an oncologist. While clustering was being performed manually by the principal investigator, there was some bias because some of the medical terms were complex and ambiguous. After the oncologist evaluated the cluster, he did some rearranging to create the ideal cluster. This kind of evaluation is called gold standard approach which helps in comparing the system clustering to the real scenario.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Carcinoma	Mass	Chemotherapi	Icter	Oncolog	Head
cell	Left	Cycle	Pain	Secondari	Eye
Squamous	Right	Fluorouracil	Swell	Primari	Ear
Cervix	Breast	Cisplatin	Lymphadenopathi	Investing	Nose
Vagina		Adramycin	Abdomen	Subject	Throat
Bleed		Book	Abdomin	Major	
Discharge		Radiotherapi		Occup	
Wall		Singl		Radiolog	
Vagin		Day		Diagnosi	
Per		Tramadol		Follow	
Cancer		Oper		Male	
		Biopsy		Addi	
		Done		Wife	
		Tomographi		Involv	
		Post		Femal	
		Histori		Duration	
				Present	
				Differenti	
				Lower	
				Start	

Table 3 Manual cluster done by the principal investigator: *the texts mentioned in the table are stemmed texts*

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Carcinoma	Mass	Chemotherapi	Icter	Oncolog	Biopsy
cell	Left	Cycl	Lower	Secondari	Done
Squamous	Right	Fluorouracil	Swell	Primari	Tomographi
Cervix	Breast	Cisplatin	Lymphadenopathi	Investig	Abdomin
Vagina	Cancer	Adramycin	Head	Subject	Biopsy
Bleed	Pain	Book	Eye	Major	
Discharg		Radiotherapi	Ear	Occup	
Wall		Singl	Nose	Radiolog	
Vagin		Day	Throat	Diagnosi	
Per		Tramadol	Abdomen	Follow	
Involve		Oper		Male	
Cervic		Post		Addi	
		Start		Wife	
				Histori	
				Femal	
				Duration	
				Present	
				Differenti	
				Clear	

Table 4. The cluster done by an oncologist

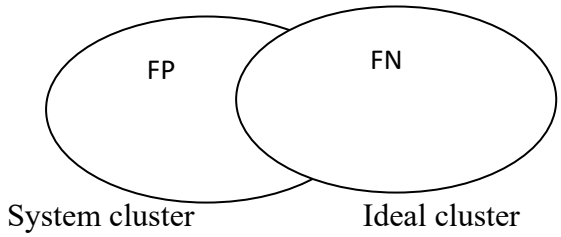


Figure 11. Precision/Recall measures

In order to evaluate the clusters, external measures were used. The external measures are calculated by matching the structure of the clusters with some pre-defined classification of instances in the data. These measures are rand index, accuracy, precision/recall measures and F measures.

1. **Rand index**: tries to find the ratio of matching and unmatched observations among the two clustering structures.
   Rand index= (a+d)/ (a+b+c+d)

   Where a= observations which were available in the same cluster in both structures (C1 andC2)
   b=observations which were available in a cluster in C1 and not in the same cluster C2

c= observations which were available in a cluster in C2 and not in the same cluster C1

d= observations which were available in different clusters in C1 andC2 (46).

For hierarchical: a=42, b=22, c=22, d=44, (42+44) / (42+22+22+44) = 66.2% and

For k means: a=27, b=37, c=37, d=74, (27+74) / (27+37+37+74) = 57.7%

2. **Precision/Recall measures:**

Precision measures the system's ability to reject any nonrelevant terms in the extracted set. Here the relevant terms are the terms found in the ideal clusters (clusters developed by an oncologist)

Precision= TP/ (FP+TP) where TP is true positive which are relevant terms extracted by the system and FP is false positive which irrelevant terms extracted (47).

For hierarchical: TP=42, FP=22, 42/ (42+22), precision= 65.6% and

For k means: TP=27, FP=37, 27/ (27+37), precision= 42%

Recall measures the system's ability to find all the relevant terms

Recall= TP/ (TP+FN) where FN is false negative which are relevant terms omitted (47)

For hierarchical: FN=22, 42/ (42+22), recall= 65.6 % and

For k means: FN=37, 27/ (27+37), recall= 42%

3. **Accuracy:** measures the proportion of true predictions (TP+TN) in the whole population (TP+FN+TN+FP) (46)

Accuracy= (TP+TN)/ (TP+FN+TN+FP)

For hierarchical: (42+0) / (22+42+0+22) = 48.8% and

For k means: (27+0) / (27+37+0+37) = 26.7%

Clustering Algorithm	Rand Index score	Precision	Accuracy	Recall
K means	57.7%	42.2%	26.7%	42.2%
Hierarchical	66.2%	65.6%	48.8%	65.6%

Table 5. Objective Evaluation results

When the evaluation result of this project was compared with the other articles reviewed, one can easily see that the precision and recall was the least because of different parameters mentioned below.

4.5. Parameters affecting the system performance

During evaluation of the extracted knowledge using the mentioned algorithms, as can be seen from table 5 it was not 100% accurate and the whole relevant texts are not retrieved as needed. There are some parameters that affect the result of the evaluation method such as:

- **The size of the corpus:** because of the physicians' hand writing illegibility, inability to determine the total number of the medical records in the oncology unit, the organization of the papers and missing papers found in the medical records and the number of staffs working in the record office is so low that they were unable to retrieve a large number of medical records. All these factors affect the size of the corpus.
- **Some missing texts:** some of the records have short patient notes that may miss the necessary parts which was also important for the text analysis.
- **Selection of initializations of the centroids:** even if selecting and running several initializations of the centroids in k means algorithm is expensive and time consuming. It has a great impact on the efficiency of the result.
- **Noisy text in the dataset:** texts which have poor spelling and punctuation.

4.6. Deployment

Even if the result of this project seems a lit bit simple, by improving the parameters affecting the system performance, the significance of the project can have a great value. And it will be included in the electronic health record for the purpose of clinical decision making process in the future.

4.6.Strength and Limitations of the Project

4.6.1. Strength of the Project

The strength of this project is being able to assess the condition of the cancer patients' medical records in Tikur Anbessa specialized hospital and the project tried to show the need for text mining in the medical domain especially in the context of medical records in order to improve the quality of patient care.

4.6.2.Limitations of the Project

Text mining in medical domain especially in the oncology unit can provide a lot of benefits but for this project the number of medical records taken as a corpus is minimum, it is limited only to 137, because of the challenges faced during the data collection. The other limitation of this project is that the clustering algorithms used were only K means and hierarchical and the comparisons of the algorithms were done with the same number of data sets. Even if there are many distance measuring methods, only elbow method was used.

4.7. Lessons Learnt and Challenges

4.7.1. Lessons Learnt

At the time of performing this project, the importance of having automated medical records has been observed. If the medical records were electronic, there would be a lot of benefits such as ease access of data for researchers, time saving and improving quality of patient care.

4.7.2. Challenges

During the time of data collection, I have faced a lot of challenges. Some of them are:

- The patients' medical records were not organized in such a way that any physicians or specially researchers will be able to get the necessay information that they want.
- The physicians hand writing was illegible to read and understand, also they usually use abberivations even some of the abbreviations are not medically standard.
- Most of the medical records are so bulky with lots of follow up form, history sheets and different kinds of investigation results which are updated every time they come. Because of this some of the necessary sheets are not there.
- Since the medical records were not automated, there are no data as to how many medical records for cancer patients are in the record room. So I was unable to determine the sample size for this project, I just took a random number.
- As the hospital has only two staffs for this record room and there are a lot of patients to be seen every day, I collected only at the staffs' convenient time. This made my data collection time longer.
- During the time of preprocessing, some of the redundant and unnecesaary terms which were supposed to be removed using remove words command in the transformation function were still there. So I was supposed to use the remove words repeatedly.
- Since the physicians hand writing was illegible and abberivated, it was difficult for text fairy (optical character recognizer) to recognize the characters. So I was forced to copy the texts in to the computer which was again a difficult thing to do and time taking.

5. CONCLUSIONS AND RECOMMENDATIONS

5.1. Conclusions

The purpose of this project was to demonstrate the application of text mining techniques to extract valuable knowledge from medical records. The project was done on cancer patients' medical records because cancer has become a challenging issue around the globe and it is increasing significantly in Ethiopia. Even if the medical record of Tikur Anbessa specialized hospital is not yet automated and retrieving the necessary medical records was very challenging.

Since cancer is a very costly chronic disease particularly in developing countries, proper care and management of resource allocated for this program is very mandatory but this can be done by having the necessary structured information at hand. This is done by using different information technologies especially text mining. It helps in describing pattern between sign and symptoms and type of treatments needed, predicting the prognosis, showing association between cases etc.

In case of Ethiopia the patients' medical records need a lot of organization and structure in order to have the necessary information and knowledge at hand and for this reason text mining will play a great role and makes the life of medical practitioners and users at ease.

During data collection, I was able to assess the nature of the texts found in the medical records. The physicians hand writing was illegible even sometimes it was very difficult to recognize the texts and there were a lot of abbreviations both medical and non-medical. After preprocessing the necessary texts patterns were extracted using both k means and hierarchical clustering algorithms.

From the result observed the cancer cases mainly covered in the corpus, are mostly breast and cervical cancer as well as cervical cancer involving the vaginal wall. The patterns extracted also showed that the most common ways of treating the cancer are chemotherapy and radiotherapy and the most commonly used chemotherapy drugs are clustered together. And according to the evaluation approach used, hierarchical algorithm performed better than the k means because of different parameters like the size of the data collected and missed texts. As some of the articles mentioned in the related literature, the hierarchical algorithm performs better whenever there is a small dataset is which the same as this project.

At last, we can conclude that text mining has a great benefit in the field of medical domain especially for extracting valuable knowledge from patient medical records which will ease the work load of health professionals, researchers and for the patient to have quality of care. Not only clustering can give different benefits in this domain but there are also other text mining applications that are very invaluable.

5.2. Recommendations

For Researchers

1. As different literatures have disscussed and showed text mining has various applications and these applications will have a lot of benefits in the medical health care. And different kinds of knowledge can be discovered and predicted not only from cancer medical records but also from other chronic illnesses (AIDS, cardiac, diabetes mellitus, hypertension) that need further researches and experiments. This new knowledge discovery will help in improving the quality of patient care and teaching/learning processes in the field of medical care. Also text mining decrease the medcial errors that can happen and save patient life.
2. In this project, the evaluation of the clustering algorithm was only concerned in the accuracy of the algorithms used but there are issues like time complexity, compactness, entropy which measure the quality of the algorithm which need further research to the context of medical records.
3. By varying the number of data set and using a large dataset, comparison of the other clustering algorithms will help to identify the best one in health care setting specially in the medical record.

For Health Practitioners and Tikur Anbessa Specialized hospital

1. Handling the patients medical records in an organized and proper way so that the physicians and any researchers will be able to have the necessary data/information whenever the need arises. By minimizing the number of formats used in the patients' medical records and discarding the laboratory results after recording them in a chronologically manner.
2. Writing in an eligible manner so that either a health practitioner or researcher can easily understand what has been written about the patient or his health care condition.

For Business Organizations

Software developers can extract different new knowledge in the medical domain and solve the problem of manual medical record handling by automating the patient medical record. The new knowledge extracted will be included in the electronic medical record for clinical decision making process. They can engage in the data preparation and conversion, database building, research, etc. Particularly big data in the health sector is a big thing these days. It

helps a lot in policy making as well as in clinical decision making. Therefore the businesses can and should invest to discover this virgin territory.

BIBLOGRAPHIES

1. Hsinchun Chen, Sherrilynne S.Fuller , Carol Friedman and William Hersh. Knowledge management,data mining and text mining in medical informatics Boston: Springer; 2005.

2. Moty Ben-Dov, Wendy Wu, Ronen Feldman and Paul A. Carins. Improving Knowledge discovery by combining text mining and link analysis techniques. 2004.

3. wikipedia. [Online].; 2017 [cited 2018 January 5. Available from: https://en.wikipedia.org>wiki>healthinformatics

4. Ning Zhong, Yuefeng Li and Sheng-Tang Wu. Effective pattern discovery for text mining. IEEE transactions on knowledge and data engineering. 2012.

5. Micah J.Crowsey Amanda R. Ramstad, David H.Gutierez, Gregoru W.PaLadino and K.P.White Jr. An evaluation of unstructured text mining software. IEEE. 2007.

6. Gary Miner, Dursun Delen, John Elder, Andrew Fast, Thomas Hill, Robert A.Nisbet. Practical text mining and statstical analysis for non structured text data application. 1st ed. USA: Academic press; 2012.

7. Si Yan and Yangling Qi. Apply text mining to advance cancer research. International journal of pharma medicine and biological sciences. 2015.

8. Agnieszka Mykowiecka, Malgorzata Marciniak and Anna Kupsc .Rule based information extraction from patients' clinical data. Journal of bimedical informatics. 2009.

9. Kathy Lee, Ankit Agrawal and Alok Choudhary. Real time disease surveillance using twitter data:Demonstration on flu and cancer. ACM. ; 2013.

10. Niloofa Moahammadzadeh(phd) Reza Safdari(phd) and Farshid Mohammedzadeh. Using intelligent data analysis in cance,r care:benefits and challenge. Journal of health informatics in developing countries. 2014.

11. Yohannes W.Woldeamanuel Belayneh Girma and Alula M.Teklu. Cancer in Ethiopia. Article in the lancet oncology. 2013;: p. 1.

12. Health sector transformation plan(HSTP) 2008-2012. Addis Ababa.

13. Zwollo Kim. Accelerating Scientific discovery with Text mining. European Biotechnology. 2015.

14. Sahami Ashok N.Srivastava and Mehran. Text mining: Classification, Clustering and applications USA; 2009.

15. Matthew Herland, Taghi M. Khoshgoftaar and Randall Wald. A review of data mining

using big data in health informatics. Journal of big data. 2014; 1(2).

16. Zhai Charu C. Aggarwal and Cheng Xiang. Mining text data USA: Springer; 2012.

17. Henk Harkema, Ian Roberts, Robert Gaizauskas and Mark Hepple. Information etxraction from clinical records. 2005.

18. Grant S.Ingersoll, Thomas S.Mortaon and Andrew L.Faris. Taming text: How to find organize and manipulate it New York; 2013.

19. Feinerer Ingo. Introduction to the tm package: text mining in R. Journal of statistical software. 2017 March 2.

20. Abate Sm, Yilma Z.Assefa,.and Tigeneh W. Trends of breast cancer in Ethiopia. International journal of cancer research and molecular mechanisms. 2016.

21. Irena Spasic, Jacqueline Livsey, John A.Keane, Goran Nendaic. Text mining of cancer related information: Review of current status and future directions. International journal of medical informatics. 2014.

22. Robert J., S. Thomas, Ryan Callahan, Richard Bartlett and Jodie Geissier. Delivering Affordable cancer care: a value change to health systems. Forum Report. ; 2015.

23. National cancer control plan Ethiopia 2016-2020. ; 2015.

24. Francis Laura. The patient centered oncoolgy information system-automating oncology work flow for optimal patient care. European oncology and hematology. 2012.

25. Wondemagagnhu Tigeneh, Abera Molla, Aynalem Abreha and Mathwose Assefa. Pattern of cancer in Tikur Anbessa Specialized hospital oncology center in Ethiopia from 1998 to 2010. International journal of cancer research and molecular Mechanisms. 2015.

26. averbis text analytics. [Online]. [cited 2017 July 25. Available from: **Error! Hyperlink reference not valid.**.

27. Nikaksu Olegas. Development and application of data mining methods in medical diagnostics and healthcare management. 2015..

28. Bannie Dockham, Jill Vanak, Erica Fischer-Cartiledge, Allison Harvey and Rhea Suarez. Frankly speaking about cancer:Treatments and side effects; 2014.

29. Lekha. Information extraction in medical domain. 2015..

30. Delvecchio Alex. Clinical decision support system definition. Health IT purchasing report ; 2017.

31. R.Balamurugan Dr.S.Pushpa. A review on various text mining techniques and algorithms. Second international conference on recent innovations in science, enginering and management. 2015.

32. Demigha Souad. Mining knowledge of the patient record:"the Bayesian classfication to predict and detect anomalies in Breast cancer. The electronic journal of knowledge management. ; 4(3).

33. Shearer Colin. The CRISP-DM Model: The New Blue print for Data Mining. Journal of data warehousing. 2000 Fall; 5(4).

34. Michal Rogalewicz, Robert Sika. Methodologies of knowledge discovery from data and data mining methods in mechanical engineering. Management and production engineering review. 2016 December: p. 97-108.

35. goopta Chandan. The NEWSTACK. [Online].; 2004 [cited 2018 January 5. Available from: **Error! Hyperlink reference not valid.**.

36. Raghav Bali, Dipanjan Sarkar, Brett Lantz and Cory Lesmeister. R: unleash machine learning technique Birmingham, UK: Packt Publishing Ltd.; 2016.

37. W.H.Veanbles D.M.Smith and the core team. An introduction to R. Notes on R:A programming environment for data analysis and graphics version 3.4.0. 2017 Apr 21.

38. Kassambara Alboukadel. Practical guide to cluster analysis in R. 1st ed.: STHDA; 2017.

39. Ritu Chauhan, Harleen Kaur and M.Afshea Alama. Data clustering method for discovering clusters in spatial cancer database. International journal of computer applications. 2010 November; 10(6).

40. Rego Felipe. Example of k means clustering with R. [Online].; 2015 [cited 2017 August 15. Available from: https://rstudio-pubs-static.s3.amazonews.com.

41. Hssain Prof. Prajna Bodapti and Abdul Ameer. Pattern discovery and document clustering using k means, PAM and HAC. International journal of technical research and applications. 2016.

42. Ali Seyed Shirkhorshidi, Saeed Aghabozorgi, The Ying Wah. A comparison study on similarity and dissimilarity measures in clustering continous data. PLoS ONE. 2015 May; 10(12).

43. UC Business Analytics R programming GUide. [Online]. [cited 2017 July 12. Available from: uc-r.github.io>hc_clustering.

44. Daniel Crabtree, Peter Andra and Xiaoying Glao. A clustering evaluation method Newzealand; 2007.

45. Micheal Steinbach, George Karypis and Vipin Kumar. A comparison of document clustering techniques. .

46. Arguello Jaimie. Evaluation Metrics. 2015..

47. Uzma Raja, Tara Mitcheu and J.Michael Hardin. Text mining in health care: applications and opportunities. Article in journal of health care information management JHIM. 2008.

48. P.Matykiewicz, W.Duch and J.Pestian Nicolaus. Clustering semantic spaces of suicide notes and newsgroups articles. In Proceedings of the workshop on BioNLP; 2009.

49. Abbas Osama Abu. Comparisons between data clustering algorithms. IAJIT. 2008; 5(3).

50. Lousie Francis, FCAS,MAAA and Matt Flynn(phd). text mining handbook. In Casualty Actuarial society E-forum; 2010: spring.

51. Mrs.Sayantani Ghosh, Mr.Sudipta Rov and Prof. Samir K.Bandyopadhyay. A tutorial review on text mining algorithms. International journal of advanced research in computer and communication engineering. 2012.

52. Ingo Feinerer, Kurt Hornik and David Meyer. Text mining infrastructure in R. Journal of statistical software. 2008 March: p. 8.

53. Maceli Monica. Introduction to text mining with R for information professionals. The code4Lib journal. 2016.

54. Yonjing Lin, Wenyuan Li, Keke Chen and Ying Liu. A document clustering and ranking system for exploring MEDLINE citations. Journal of the American medical informatics associations. 2007.

55. Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saied Safael, Elizabeth D.Trippe, Juan B.Gutierrez and Krys Kochut. A brief survey of text mining: classfication, clustering and extraction techniques. KDD Bigdas. 2017.

56. Sarkar Raghav Bali and Dipanjan. R machine learning by example; 2016.

57. Singh Mr. Hardeep. Clustering of text documents by implementation k means. Streamed Info-ocean. 2016 January-June; 1(1).

58. Abbas Osama Mahmoud Abu. Comparison between data clustering algorithms. International Arab Journal of information technology. 2008 July: p. 321.

59. Kafkas Jee Hyubkim and Sehay. Key concepts in tm and its usage in service provision. 2016..

60. Taranu Ionut Data Mining in healthcare: Decision making and precision. Database

Systems Journal. 2015.

61. P.Diehl, Lise Geeoor and Chirostophe. Link mining. ACM SIGKDD explorations newsletter. 2005.

62. Scharber W. Evaluation of open source text mining tools for cancer surveillance. Evaluation of text mining tools. 2007.

63. Eight to late. [Online]. [cited 2017 August 29. Available from: https://eight2late.wordpress.com.

64 Burner F. Cluster analysis in R-Unsupervised machine learning. [Online]. [cited 2017 September 23. Available from: www.sthda.com>english.

APPENDIX

Sample R Code

```
library(tm)

library(NLP)
docs<-Corpus(DirSource("c:/users/ACER/Desktop/pro"))
getTransformations()
docs<-tm_map(docs,content_transformer(tolower))
docs<-tm_map(docs,removeNumbers)
docs<-tm_map(docs,removeWords,stopwords("english"))
mystopwords<-
C("age","sex","femal","region","oromiya","occup","merchant","marit","statuspartn","marri","
religion","orthodox","protest","patient","assessment","tomographi","comput","normal","num
ber","children","habit","smoke","ethanol","kchat","comorbid","ill","subject","complaint","mo
nth","durat","object","find","perform","status","eastern","cooper","oncolog","group","diagnos
i","specif","blood","pressur","puls","rate","non","icter","sclera","lymph","node","system","in
vestig","result","complet","blood","count","organ","function","test",
"chest","xray","pelvic","ultrasound","tumor","marker","radiolog","imag","histolog","biopsi","
treatment","plan","primari","secondari","summari","main","p","stage","iva","give","vital","si
gnstabl","headeyeearnosethroatpink","conjunctiva","gland","systemno","chestclear","central
nervous","conscious","countorgan","testultrasoundchest","amhara","othodox","year","back","
pink","conjunctiva","x cm","civil","servant","iv","mg","poday")
docs<-tm_map(docs,removeWords,mystopwords)
docs<-tm_map(docs,stripWhitespace)
docs<-tm_map(docs,stemDocument)
tdm<-TermDocumentMatrix(docs)
dim(tdm)
dtms
dtms<-removeSparseTerms(tdm,0.85)
dtms
m<-as.matrix(dtms)
library(ggplot2)
wss<-(nrow(tdm)-1*sum(apply(tdm,2,var))
for(i in 2:15)wss[i]<-sum(kmeans(tdm,centers=i)$withinss)
```

```
plot(1:15,wss,type="b",xlab="no of clusters",ylab="with in group sum of squares", main="assessing the optimal point of clusters with the elbow method", pch=20, cex=2)
library(stats)
library(cluster)
d<-dist(scale(m))
set.seed(1234)
k<-kmeans(d,6)
k
clusplot(m,k$cluster,color=T,shades=T,labels=2,lines=0)
hc<-hclust(d,method="ward.D")
plot(hc)
plot(hc,hang=-1)
library(ape)
colors<-c("red","blue","green","black","purple","brown")
clust<-cutree(hc,6)
plot(as.phylo(hc),type="fan",tip.color=colors[clust],label.offset=1,cex=0.7)
```

DECLARATION

I, the undersigned, declare that this project is my original work in partial fulfillment of the requirement for the Masters of Science in Health Informatics program and has not been presented for a degree in this or any other university. All source of material used for this project and all people and institutions who gave support for this work have been duly acknowledged.

Name: Bethelhem Abebe

Signature:

Place: Health Informatics Program, Faculty of Informatics, Addis Ababa University

Date of submission: 10 January 2018

This project has been submitted for examination with our approval as the university advisors.

Name of the advisors

Dr. Martha Yifiru

Ato Mengistu Yilma