



**JIMMA UNIVERSITY  
SCHOOL OF GRADUATE STUDIES  
JIMMA INSTITUTE OF TECHNOLOGY (JiT)  
SCHOOL OF COMPUTING  
DEPARTMENT OF INFORMATION TECHNOLOGY**

**Early Cancer Prediction and Hospital Recommendation using  
Integrated Fact Based Information and Opinion Summarization  
with the application of Knowledge Based System and Data Mining  
Techniques**

**By: Abdulkadir Ahmed**

**THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES OF THE  
JIMMA UNIVERSITY IN PARTIAL FULFILMENT FOR THE DEGREE OF  
MASTERS OF SCIENCE IN INFORMATION TECHNOLOGY**

April, 2017

---

**JIMMA UNIVERSITY  
SCHOOL OF GRADUATE STUDIES  
JIMMA INSTITUTE OF TECHNOLOGY (JiT)  
SCHOOL OF COMPUTING  
DEPARTMENT OF INFORMATION TECHNOLOGY**

**Early Cancer Prediction and Hospital Recommendation using  
Integrated Fact Based Information and Opinion Summarization  
with the application of Knowledge Based System and Data Mining  
Techniques**

**By: Abdulkadir Ahmed**

**THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES OF THE  
JIMMA UNIVERSITY IN PARTIAL FULFILMENT FOR THE DEGREE OF  
MASTERS OF SCIENCE IN INFORMATION TECHNOLOGY**

April, 2017

---

**JIMMA UNIVERSITY  
SCHOOL OF GRADUATE STUDIES  
JIMMA INSTITUTE OF TECHNOLOGY (JiT)  
SCHOOL OF COMPUTING  
DEPARTMENT OF INFORMATION TECHNOLOGY**

**Early Cancer Prediction and Hospital Recommendation using  
Integrated Fact Based Information and Opinion Summarization  
with the application of Knowledge Based System and Data Mining  
Techniques**

**By: Abdulkadir Ahmed**

**THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES OF THE  
JIMMA UNIVERSITY IN PARTIAL FULFILMENT FOR THE DEGREE OF  
MASTERS OF SCIENCE IN INFORMATION TECHNOLOGY**

**Approval Signature of the Board of Examiners**

Dabala Tesfay(Principal Advisor) \_\_\_\_\_

Teferi Kebebew (Co-Advisor) \_\_\_\_\_

Chair Person \_\_\_\_\_

External Examiner \_\_\_\_\_

Internal Examiner \_\_\_\_\_

# Contents

<b>1</b>	<b>INTRODUCTION</b>	<b>1</b>
1.1	Background of the Study . . . . .	1
1.1.1	Statement of the Problem . . . . .	4
1.1.2	Significance of the study . . . . .	7
1.2	Objective of the Study . . . . .	8
1.2.1	General Objective . . . . .	8
1.2.2	Specific Objectives . . . . .	8
1.3	Scope and limitation of the study . . . . .	8
1.4	Contributions of the study . . . . .	10
1.5	Methodology . . . . .	11
1.5.1	Empirical Research: . . . . .	11
1.5.2	Research Design: . . . . .	11
1.5.3	Literature review . . . . .	12
1.5.4	Opinion Summarization . . . . .	12
1.5.5	Knowledge Discovery Process . . . . .	13
1.6	Organization of the study: . . . . .	17
<b>2</b>	<b>LITERATURE REVIEW</b>	<b>19</b>
2.1	Introduction . . . . .	19
2.1.1	Opinion Mining . . . . .	19
2.1.2	Knowledge-Based System . . . . .	27
2.1.3	Overview of Data Mining . . . . .	32
2.1.4	Review of Related Works . . . . .	37
<b>3</b>	<b>DATA PREPARATION AND PREPROCESSING</b>	<b>40</b>
3.1	Overview of the real world data . . . . .	40
3.2	Understanding the problem . . . . .	40
3.3	Understanding the Data . . . . .	41
3.3.1	Description of the Collected Data . . . . .	42
3.4	Data Preprocessing . . . . .	43
3.4.1	WEKA Format . . . . .	49
3.5	Selecting Best classifiers . . . . .	49
3.5.1	Balancing Dataset . . . . .	52
3.5.2	Attribute Selection . . . . .	52
3.5.3	Experimental Setup . . . . .	52
3.6	Experiments and Results Using J48 . . . . .	53
3.6.1	The Experiment with the Highest Accuracy . . . . .	53
3.6.2	Confusion Matrix of the Selected Model . . . . .	53
<b>4</b>	<b>INTEGRATING OPINION SUMMARY WITH FACT BASED INFORMATION</b>	<b>58</b>
4.1	Opinion Summary . . . . .	58
4.2	General system architecture . . . . .	59
4.2.1	Proposed Techniques . . . . .	60
4.2.2	Sentiment Classification . . . . .	63
4.2.3	Part-of-Speech Tagger . . . . .	64
4.2.4	VSM (Vector Space Model) . . . . .	66
4.2.5	Cosine Similarity . . . . .	69
4.2.6	Corpus Preparation . . . . .	75
4.2.7	Manual collection of Opinion and Its Aspect . . . . .	76
4.3	Evaluation . . . . .	76
4.4	Experiment and Result Discussion . . . . .	77
4.5	Results . . . . .	77

4.6	Discussion . . . . .	78
4.6.1	Amharic Review discussion . . . . .	78
4.6.2	Afaan Oromo Review discussion . . . . .	82
4.6.3	the overall sentiment classification algorithm discussion . . . . .	83
4.7	Fact Based Information . . . . .	83
<b>5</b>	<b>INTEGRATING THE PREDICTIVE MODEL INTO KNOWLEDGE BASED SYSTEM</b>	<b>87</b>
5.0.1	Predictive Model . . . . .	87
5.0.2	Problem Understanding . . . . .	87
5.0.3	DataSet . . . . .	89
5.0.4	Data Understanding . . . . .	89
5.0.5	Data Preprocessing . . . . .	89
5.0.6	Applying Classifier Algorithm to Build Model . . . . .	89
5.0.7	Evaluating the Models . . . . .	89
5.0.8	Selecting the Model . . . . .	89
5.0.9	Selected Built Model . . . . .	90
5.0.10	Selected Built Model Extracted . . . . .	90
5.0.11	User Interface . . . . .	90
5.1	Knowledge Based System . . . . .	90
5.1.1	knowledge acquisition . . . . .	90
5.1.2	Knowledge Representation . . . . .	91
5.1.3	Integration . . . . .	91
5.2	TESTING AND EVALUATION . . . . .	93
<b>6</b>	<b>INTEGRATING OPINION SUMMARY AND FACT BASED INFORMATION WITH KNOWLEDGE BASED SYSTEM AND DATA MINING</b>	<b>94</b>
6.1	General Architecture of Integration . . . . .	94
<b>7</b>	<b>RESULT AND DISCUSSION</b>	<b>98</b>
7.1	Discussion . . . . .	98
<b>8</b>	<b>CONCLUSION AND RECOMMENDATIONS</b>	<b>100</b>
8.1	CONCLUSION . . . . .	100
8.2	Recommendation . . . . .	102
8.3	APPENDICES . . . . .	108
8.3.1	Appendix I:Interview Questions to Obtain Tacit Knowledge from Expert . . . . .	108
8.3.2	Appendix II:Policy Related Interview Questions: . . . . .	108
8.3.3	Appendix III: Sample Cervical cancer Dataset . . . . .	109
8.3.4	Appendix IV: Sample cervical cancer hospital Dataset . . . . .	109
8.3.5	Appendix V: Questioner to collect Afaan Oromo opinion from different hospital patient . . . . .	111
8.3.6	Appendix VI:Questioner to collect Amharic opinion from different hospital patient . . . . .	112
8.3.7	Appendix VII: System Prototype Evaluation Form . . . . .	113
8.3.8	Appendix VIII :sample Integration code . . . . .	114
8.3.9	Appendix IX:sample medical data request letter . . . . .	116

# List of Figures

1.1	Hybrid Knowledge Discovery Model adopted from [68]	12
1.2	Hybrid Knowledge Discovery Model adopted from [68]	14
2.1	Support Vector Machines Classifiers	25
2.2	Vector Space Model	26
2.3	Structure of Knowledge-Based System	29
3.1	Initial Dataset from Addis Ababa hospitals that provides Cervical Cancer Treatment	43
3.2	Attribute Name, Description, Data Type and Values	44
3.3	Initial Dataset from Addis Ababa hospitals that provides Patient History of follow up only	45
3.4	Attribute with missed values in cervical dataset	46
3.5	Attribute with missed values of Patient follow-up History	46
3.6	Total Number of registry of Cervical Cancer Dataset after removing missing Values	47
3.7	Data set of cervical cancer after cleansing	47
3.8	Total Number of registry of Patient History follow up Dataset after removing missing Values	48
3.9	Attributes Removed from both cervical cancer and patient history dataset	48
3.10	Transformation for Education Status Attribute	49
3.11	Transformation for Age Attribute	49
3.12	Sample of Arff file for Cervical Cancer Data	50
3.13	Sample of Arff file for Patient History of Hospital	51
3.14	Comparison between the Actual Data and Balanced Data	52
3.15	Sixteen Experiment that was conducted on actual,pruned,unpruned and balanced data	53
3.16	Confusion Matrix of J48 pruned algorithm upon all attributes of the actual data	54
3.17	Detail Analysis Result of J48 Algorithm	54
3.18	Partial decision tree for the selected J48 algorithm	56
3.19	Interesting Rule generated by weka rule for cervical cancer disease	56
4.1	Aspect Based Opinion Mining Model for Afaan Oromo and Amharic Language	59
4.2	Review Document for single hospital	60
4.3	Language Detection Model	62
4.4	Sample XML Parallel corpus for Amharic to English language Translation	63
4.5	Sample Java code to convert Afaan Oromo to English language Translation	64
4.6	Sample Review of Hospital Opinion	64
4.7	Sample Opinion Term and its Adjective Equivalent for Afaan Oromoo Language	65
4.8	Sample code for Pos-Tagging	65
4.9	Sentiment Classification Model for Afaan Oromo and Amharic Language	67
4.10	Computing similarity between bathtub and water adopted from[12].	68
4.11	Algorithm for Sentiment Classification	69
4.12	N-Gram	70
4.13	Equivalent Opinion for Afaan Oromo and Amharic Language Sentiment Orientation of reviews	70
4.14	Identifying Sentiment Orientation of reviews	71
4.15	Aspect Detection Model	72
4.16	Annotated Afaan Oromoo Aspect	74
4.17	Sample Detected Aspect by the System	74
4.18	Aspect based sentiment summarization Prototype	75
4.19	google translate for translation of online English opinion of patient corpus into Amharic Language corpus.	76
4.20	Model for fact based information	84
5.1	Architecture for Integration Data Mining Result with Knowledge based system	88
5.2	Step to Extract selected model for Integration	90
5.3	Sample code to Extract selected model for Integration	91
5.4	Java code to transform CSV file to Arff file format	92

5.5 sample code read by Java . . . . . 92  
5.6 .. CALLING PROLOG FROM JAVA . . . . . 93

# List of Tables

3.1	Total Number of Initial registry of Cervical Cancer Dataset from Addis Ababa hospitals. . . . .	42
3.2	Selected Experiment with the Highest Accuracy . . . . .	55
4.1	Aspect Detection Evaluation Comparison between various algorithms. . . . .	73
4.2	Aspect Detection: Exprimen1 (Amharic Language) . . . . .	78
4.3	Aspect Detection: Exprimen2 (Amharic Language) . . . . .	78
4.4	Sentiment Classification: Exprimen1 (Amharic Language) based on data collected for two months.	78
4.5	Sentiment Classification: Exprimen2 (Amharic Language) based on data collected for two months.)	79
4.6	Aspect Detection: Exprimen1 (Afaan Oromo Language Reviews) . . . . .	79
4.7	Aspect Detection: Exprimen2 (Afaan Oromo Language Reviews) . . . . .	80
4.8	Sentiment Classification: Exprimen1 (Afaan Oromo Language Reviews) based on data collected for two months.) . . . . .	80
4.9	Sentiment Classification: Exprimen2 (Afaan Oromo Language Reviews)) . . . . .	80
4.10	Aspect Detection: Exprimen1 (Afaan Oromo Language Reviews)) . . . . .	81
4.11	Aspect Detection: Exprimen2 (Afaan Oromo Language Reviews)) . . . . .	81
4.12	Aspect Detection: Exprimen2 (Afaan Oromo Language Reviews)) . . . . .	81
4.13	Human and Machine Summarizer ) . . . . .	82
4.14	Average Sentiment Classification Evaluation Comparison between various algorithms ) . . . . .	82
4.15	Fact Based Evaluation Comparison between various algorithms. . . . .	84
4.16	Fact Based Recommendation: Exprimen1 based on data collected for two months. . . . .	86
5.1	Summary of User Acceptance Testing. . . . .	93



## Dedication

The Memory of My Brother Ibrahim Ahmed.  
“Brother I Miss You Very Much”

# Acknowledgments

First Thing First, I would like to thank my family who give me love and encouragement while doing this research work. Next, I would like to give my special thanks to my respected advisers Debela Tesfay and Teferi Kebebew who has spent their whole time in providing their valuable comments, fruitful suggestions, brilliant ideas and magnificent support during time of this thesis work.

I would also like to thanks Julian McAuley who has helped me in getting hospital data source. My special thanks also goes to Seid Yusuf for his moral and encouragement. I wish to express my thanks to school of computing for their support in their career.

I also thank one of my best friends Yimam Nuriye who is currently working as an instructor in Jimma University School of Computing, who was behind my current situation and who has spent his time for notable advice and technical support throughout my work.

My thanks message also goes to my respected friends Mohammed Tune, Tesfaye Fufa, Fitsum Andersom, Abdellah Behri and Mohammed Nuru who are working as an Instructor in Debre Birhan University in computing school, Idris Mohammed working as Instructor in Debre Birhan University in Business and Economics Faculty , Idris Yimam working as Instructor in Mathematics Department, Ahmed Sali Bakar working as Civil Engineering Department and Husein Kadir working as instructors in department of philosophy for their knowledge sharing and technical help using their impressive skills and knowledge and I am very happy for having them.

I also thanks students with whom I made useful discussion. Finally, I would like to pass my thanks to all peoples as whole that has helped me throughout this thesis work either directly or indirectly.

# List of Acronyms

BOW:-Bag of Words

EM: -Expected Maximization

EDPHR:Early Disease Prediction and Hospital Recommendation

FBR: Fact Based Recommendation

IR:-Information retrieval

FMOH: Federal Ministry of Health

NLP:- Natural Language Processing

KBS: -Knowledge Based System

OMASS:- Opinion mining and Summarization System

JPL: -Java to Prolog Library

PART: - Partial Decision Tree

PROLOG:-Programming in Logic

SMOTE: - Synthetic Minority Over-sampling Technique

TF-IDF:-Term Frequency and inverted document frequency

UGC:-User generated Content

WP: -Wikipedia

WWW: World Wide Web

# Abstract

In this digital world, the support of health care system recently with the help of computer application are tremendously changing the way health care organization are currently delivering their services to the users. Recommender systems (RS) suggest items of interest to users of information systems or e-business systems and have evolved in recent decades. A typical and well known example is Amazons suggest service for products. Patient disease prediction, advising and suitable health care recommendations are some of the services that are provided in traditional means in hospitals that are found in Ethiopia. One of the problem currently facing healthcare/hospital organization is the provision of quality services at a reasonable price. Quality services can be related to customer satisfaction, correct diagnosis, and proper management of patient treatment in Developing countries like Ethiopia. This is due to absence of serious preventative mechanism that can be followed.

The main objective of this study was to develop cancer disease prediction and recommendation Model at Early Stage using integrated fact based information and opinion summarization with data mining and knowledge based system.

For prediction of cervical cancer J48 Algorithm was selected over 10127 records of cervical cancer medical history. Hand crafted rule with Vector Space Model was used for opinion summary. For fact based information only Vector Space Model was used. Afaan oromoo and Amharic dataset were collected in two versions: the first version is for training set and the second version is for testing set. In one hand, the dataset for training was Prepared in two forms: the first one is the data from online English patient opinion is translated into Afaan oromoo and Amharic using English to Afaan oromoo and Amharic language translators. The translated opinion was provided for Afaan oromoo and Amharic linguistic experts to correct its language meaning and the polarity of the sentiments. Then the corrected opinions were used to train the model with selected hybrid approach. An experiment was conducted on five hospital in Addis Ababa: ALERT, Black Lion Hospital, Teklehaimanot General Hospital, Zewditu Memorial Hospital, and St. Pauls Hospital. Lastly, as we understand from an experiment carried out in Amharic and Afaan oromo version Reviews St. Pauls hospitals registered less score in every experiment. Therefore, such types of output can't be considered as patient language complexity problem but there is some fundamental problem that should be addressed by the health care. For integration of data mining with knowledge based system JPL (java prolog language) was employed for integration. For this, User acceptance testing was carried out and registered 90.6 % of accuracy that was promising result.

The developed model was developed and applied on five hospitals data set. The recommender system is scalable and works for more than one disease if the EDPHR Model are supplied with adequate and proper data. The developed EDPHR Model can also help both patient who already know the disease and the new patients. For the patient who already know the disease it will recommend the best choice hospitals. For new patient it will detect the disease at its early stage and recommend the best choice among the hospitals. The developed system can minimize the disproportion between the patients and professionals. Generally, efficiency of opinion summarization depend on the frequency of summarization. The more input is summarized, the more efficient and effective the algorithm will be.

The future direction has been created for The developed Model can works only for medical domain, but in the future we planned to extend the system to develop model for other domain such as university, hotels and banks.

**Keywords:-** Cervical Cancer, Health Care, Opinion Mining, Fact Based Information, data mining, knowledge based system, Integrator

# Chapter 1

## INTRODUCTION

### 1.1 Background of the Study

In this digital world, the support of health care system recently with the help of computer application are tremendously changing the way health care organization are delivering their services to the users. There are several activities that can be accomplished in traditional means concerning health care services. WHO (World Health Organization) has put an improved provision of Health services as one of the six main component of health care system[1].According to an Interview made with FMOH(Federal Ministry of Health) administrator Patient disease prediction, advising and suitable health care recommendations are some of the services that are currently provided in Ethiopia in traditional means. Nowadays, getting people who has experience in diagnosing particular disease and served previously at the same hospital for particular disease is troublesome.

Early Disease detection is detection of disease at its early symptom. Early detection of cancer greatly increases the chances for successful treatment. There are two major components of early detection of cancer: education to promote early diagnosis and screening[3]. In order to provide an improved health care services, patient experience has seen as a very important factor[1][2][3].// According to [2][4][5], consumer satisfaction is a fundamental requirement for healthcare providers. Healthcare service are also one of the domain that use customer satisfaction as one factor to improve their service. To achieve this, an opinion (user generated contents) are manually collected via an opinion provided by the users of the health care toward a given services.

The inventions of web 2.0 have paved the way for the rapid growth of user generated contents(UGC). During collecting this opinion a large amount of user generated contents can be generated[6][7][8][9]. Analysing high volume of opinion manually is a challenging task. These user generated contents (UGC) has changed the world since they are easily abandoned online on the web in various domain. They can be posted online by an opinion holder from different review sites, blogs, discussion forums, social media review etc. This huge number of user generated content (UGC) can be used as source to mine the sentiment orientation (i.e. Sentiment orientation is the polarity of the opinions provided for example, positive, negative and neutral) of the opinion holder[10][11][12].

Now a days people often uses opinion, feeling, attitudes or emotion that can be expressed by users of the web online toward some particular entity. This helps for making decision like “what is the opinion of the opinion holder on some particular entity (i.e. object)”. In addition to this an opinion that can be supplied by users of the web online can helps a companies in improving their products and services. For example, most of us asks our friends for which Department we have to join, which Movie we have to watch, which Hotels can give us better service, what kind of Phone we have to buy, which Book we have to read. And which Hospital can give us better service. Considering the above example, an answer expected from friends toward these questions is called an opinion. The process of extracting the attitude of the holder of

an opinion toward something is called opinion mining or sentiment mining. Therefore sentiment miner tackles the problem of waiting an answer from friends. Furthermore, an opinion mining are known for its wide usability in real life. Thus, studying on this particular topic of sentiment mining have attracted an attention of many researchers. The object in which an opinions are expressed can be product, service, organization, events and certain issues[7][11][12][53].

Different researcher call sentiment mining as opinion mining, sentiment analysis, opinion extraction, emotion analysis, subjectivity analysis and review mining. Although the names are different they all comes under one field of study known as opinion mining with small task dissimilarity. This opinion mining is a hot research topic in the recent years in NLP (Natural Language Processing) although, it's broadly considered in field of data mining. This topic have inspired many researchers due to its availability of diverse application areas. Sentiment mining can be classified basically into three levels of granularity such as: sentence level, document level and aspect level[6][13][11].

Sentence level sentiment mining tasks are undertaken to classify reviews sentence into two concept subjective and objective. Subjective is a concept that can be classified in to either positive or negative sentiment orientation. It can answer the question whether the review is negative or positive while the objective ones are facts therefore, it can be categorized as neutral (i.e. no opinion).

***Example-1, “I bought an iPhone yesterday and its great!!”***

This sentence indicate subjective concept since, it contains opinionated words ‘great’ which is positive. Therefore the sentence has positive sentiment orientation. Example-2, ***“I bought an iPhone yesterday and it’s bad!”*** This sentence indicate subjective concept since it contains opinionated words “bad” which is negative. Therefore, this sentence has negative sentiment orientation[10][14][6].

Document level sentiment mining task is used to know the overall sentiment given over particular entity by an opinion holder. This can answer the question whether the review document has positive or negative polarity. In document level there is no way in which a user know an opinion orientation on each and every individual aspect.it can simply provide the general sentiment orientation for review document.

Let us see an example provided by an authors under one products reviews. Suppose this is the review document, (***“The battery of my htc phone is short, the Network of my htc phone is amazing, and the camera of my htc phone is bad!!”***).In document level there is no way to know in which features the reviews have expressed positive and negative opinion but the only thing it does is giving aggregate value negative since, the negative review exceeds the positive one in this particular example. Thus, at document level even though review has an amazing network features it doesn’t care about the component of the object but object itself. In both sentence and document level it is not possible to know which features are exactly loved or hated by an opinion holders[7][15].

In aspect level this things are possible since the aspect level opinion mining deals with the components or attributes of an object (e.g. Product, Movie,Hospital etc.) at finer grained level of analysis. See an example followed. Example-1 (“I bought an iPhone three days before and battery is still working!”). From this example we understand that an opinion is expressed on iPhone component battery that it’s good rather than iPhone itself. Therefore in aspect level one can able to know an opinion of an object somehow at detail level. This can answer the question whether this review sentence feature is positive or negative. Thus, this research focus is on automatic aspect summarization of hospital patient reviews[10][6].

Since opinion summary based on customer’s opinion presented to users can include both important and unimportant summary. The existing work in opinion summary can present users with both important and unimportant/unrefined summary. The integration of opinion

summary with knowledge based system and data mining is required because we do not have to make users busy/confused with unimportant summary instead it's better to present user with refined /important summary the user exactly seeking for. The integration of opinion summary with knowledge based system has never seen to the extent of our knowledge. This study can helps to extract opinion summary based on what user needs rather than opinion summary based on others customers data (i.e. opinion). In order to extract an important summary from an entire summary produced from customers opinion, knowledge about diseases are required, for this reason, knowledge based systems are integrated with opinion summarization.

According to [16][17] integration of knowledge based system with other technologies are currently used. [2] Also stated that, Knowledge based system is also named as expert system that are developed through coding, to represent knowledge or experience of professionals in a given domain. Knowledge based systems (KBS) or expert systems are system that acts as the human expert in particular domain knowledge.

Since, experts cannot share all the knowledge they have acquired via experience for diagnosing cancer disease, getting tacit knowledge through documents analysis, interview or questionnaires is very difficult. Without considering hidden knowledge, by using only knowledge based system developed based on questionnaires or interviews that has been carried out, simply diagnosing disease like cancer is very difficult. Therefore, to include tacit knowledge the integration of data mining with knowledge based system are used in this study.

Currently we are living in digital world in which knowledge is used as a powerful tool that can provide competitive advantages and knowledge management for a given organizations. The data in health care organizations are extremely increasing. Even though, the data are swiftly increasing in health care domain, the organizations doesn't get potentially maximum services data can able to provide for decision making[16][18].

This is due to availability of large amount of data. Manually, reading, understanding and analyzing every data and using it for making potentially useful decision is time consuming and boring. Therefore, using this data for making decision automatically can requires data mining techniques. Data mining is the process of acquiring potentially useful knowledge that are formerly unknown for decision making. The process of acquiring previously concealed information is possible with data mining algorithm that extract hidden knowledge[19][18].

This algorithm can able to extract meaningful and hidden patterns. According to [20], data mining can have importance in health care domain by providing numerous services such as disease prediction, hospital ranking, and drug recommendation. According to [21][18] data mining is one of the KDD (knowledge discovery in database) process that used to extract potentially useful information from high volume of data-set.

There are various techniques that are used in data mining, among this classification techniques are the most commonly used techniques. According to [18][22] data mining are classified into six main techniques such as: classification, regression, clustering, association, anomaly detection and summarization. Therefore, in this study, classification data mining techniques are used for cancer disease prediction, due to two class of screening result we use .

Often, people decide to buy a given object or serve in a given services from other people through judgment of others. But, decision based on a given entity based on judgment of others is not sufficient. Therefore, fact based information about a given object strengthen the recommendation on a given object or services. Some disease requires specialized experts who are experienced in diagnosing. According to interview we made from hospital administration and medical expert it is highly recommended to support personal experience with fact based information[10][6][7]. Cancer is one of the disease that requires specialized professionals for diagnosis.

There are various types of cancer that are appearing in this world. From all this Lung, breast, cervical, stomach, and prostate cancers account for the majority of cancer deaths[23][24][25]. Cancer is an increasing public health burden for Ethiopia and Sub-Saharan Africa at large.

Indeed, by the year 2030, cancer and other non-communicable diseases may overtake some infectious diseases as leading causes of death in the African Region. Currently cancer accounts for four per cent of all deaths in Ethiopia. Many of these deaths can be reduced if the cancer can be detected and treated early[24].

According to WHO, Cancer is registered as the sixth cause of death in Ethiopia following Lower respiratory infections, Diarrhoeal Diseases, HIV, Tuberculosis, and Stroke respectively. Globally, the number of new cancer cases could grow to as many as 16 million, and the number of deaths could rise to as many as 10 million by the year 2020 [24][25][26].

Cancer, one of the non-communicable diseases, is among the major causes of illnesses and death in Ethiopia. Hospital records show that in Ethiopia there are more than 200,000 cancer cases per year [24][25][26].

*“Cervical cancer is one of the reproductive organ cancer that is commonly seen in women.it can start mostly from lower part of uterus.” Human Papilloma virus was seen as a cause of cervical cancer.it can also be transmitted from women to women sexually in genital areas if skin to skin contact was available [24][25][26].* Developing countries like Ethiopia was one of the victim of this types of cancer highly. This is due to absence of serious preventative mechanism that can be followed [23][24][25][26].

Furthermore, according to a World Health Organization (WHO) estimate, the number of mortalities from cervical cancer is expected to increase from the current 274,883 to 474,000 annually and over 95 % of these deaths are expected to be from developing countries. After a few years patient with cervical cancer can be doubled especially in sub-Saharan Africa since it's accounted for 34.5 of 100,000 women. According to World Health Organization around 4,648 cases and 3,235 diagnosis was occurred in the Ethiopia per every year. The diagnosis problem is also seen as one of the major problem. This is due to absence of sufficient diagnosis in Ethiopia[27][26]. Therefore, one of the aim of this study is to help cervical cancer diagnosis with computer tools.

This study is aimed at developing Model that automatically recommend hospital based on predicted cervical cancer disease results by using integrated opinion summarization and fact based information with developed knowledge based system for treatment of patient using the hidden knowledge extracted via data mining algorithm.

### 1.1.1 Statement of the Problem

Most commonly people asks their friends, relatives and other peoples of whom they know to recommend them the right hospital in which they get better services.In this busy,complex and dynamic world, getting people who had experienced every service to recommend you particular hospital is very difficult. Additionally, getting people who knows every services of the given hospitals at its finer grained level is very difficult[28][10][7].

Even though, you got an individuals who can tell you an information on services of a given hospital, its very difficult to get an individual who knows every hospital with the services currently they provides at finer grained level. Therefore, waiting time to search for people who knows the disease is also high. Decision made based on the recommendation that you have got from people you may know can be imbalanced. Since, the user recommendation of an individuals can be an opinion (i.e. individual judgment). Depending on individuals opinion deciding the hospitals in which ones should get services for particular disease is biased. Moreover, an individual from whom one can seek recommendation about particular services or products of a given entity might be fake. Most of the people can gives you the general services rather than based on its aspect[11].

Customer satisfaction has an important roles in keeping quality of health care service[29][2]. Hence, In order to know customer satisfaction collecting an opinion from client is compulsory. Even though, opinion gathered from client is used for decision making it doesn't mean getting decision based on the given opinion is always 100 % guarantee(some opinion are not matured). Since there is a factor which requires to be considered. For example let us see the following scenario.



*Scenario 1: suppose in 2016 one of the Ethiopian higher educational institute has accepted 1200 students and most of the student in this batch has scored Cgpa >3.5 or in other way most of the students that were graduated have scored very good results.*

*Scenario 2: suppose in 2017 one of the Ethiopian higher educational institute has accepted 1200 students and most of the student in this batch has scored Cgpa >2.00 in other way most of the students graduated have scored poor results.*

Based on scenario 1 the students opinion can favors to positive, in contrary to this in scenario 2 the students opinion can favours to negative opinion. From this context what needs to be considered is the wrong opinion leads to incorrect summary which causes incorrect decision that results in advising incorrect product or organizations or entities. Therefore, this scenarios are not dissimilar in hospital/health-care domain[15].

To handle the problem of incorrect decision, supporting opinion with fact based information is important to decrease the effect of biased opinion for decision makers. Basically, there are two information in this world fact based information and opinion based information.

Fact based information is an information that can be proven. Meanwhile, opinion based information is an information that cannot be proven either true or false or right or wrong. That means an opinion collected from different people have a probability that all can be true. Or in contrary to this all opinion could be false or the probabilities of some of them are correct and others are not. Decision making based on unverified opinion may lead to incorrect purchase of product or use of services. But it doesn't mean that opinion do not have place in customer decision. Customer satisfaction is important to judge a given services.

Fact is combined with opinion to increase the accuracy of an information for decision making. To the extent of my knowledge I have never seen the study that integrate opinion based and fact based information together to increase the accuracy of decision making. Therefore, this study focused on integrating opinion based and fact based information.

Traditionally, in order to rank hospitals an opinion polls are conducted to collect an opinion from users of the services. After an opinion has collected from every individuals, each opinion is going to be analyzed and finally, the hospital is ranked according to an opinion analyzed. But, reading, understanding and analyzing each opinion collected from thousands of users from different hospital is annoying and time consuming. By using, an opinion that are collected from users only recommending a given hospital is unfair. Moreover, Ethiopia is diverse country in which more than 85 languages are spoken. to collect free text opinion/comments from users of the hospital the language problems comes here.users can write an opinion by language they are familiar with. detecting comments provided by all this language is troublesome. Therefore we have limited our scope to summarize sentiment of two major language that are most commonly in Ethiopia, Afaan Oromo and Amharic Language. these comments are collected from five hospitals.

Often, peoples visit hospitals for diagnosing and treatment of particular disease while they are being healthy. This causes, financial burdens either implicitly or explicitly. Explicitly, healthy people face financial burdens such as transportation, shelter, food, diagnosis and waiting time cost for diagnosing of the disease. Implicitly, peoples lost earnings, productivity (i.e. income) and leisure time during their visit to hospitals. Since, there is shortage of health care services in the rural areas than urban relatively[19][16].

This problem highly affect especially the peoples coming from rural areas than urban. In Ethiopia most people travel long distance for diagnosis and searching for better treatment. 60 % of rural people travel 40 km to reach their nearest health care services 23 % travel more than 100 km and 13 % and above travel 180 km. Most Patients who already diagnosed for particular disease are suffering because of poor treatment (quality services) were given at particular health care. Therefore, system that recommend better health care for particular disease is required[19][16].

Getting access to health care services are restricted by factors such as less road infrastructure,

under education, financial problems, poor communication skills, poor technology coverage, less ICT technology usage and problems of getting diagnosing for particular disease[16][19].

Late diagnosis and treatment of patient for particular disease more or less can causes the disease to become chronic/recurrent or sudden death[24][25]. Moreover, it can also cause problem such as financial loose, social stigma, infecting other people, simply infection by other disease, psychological problems, difficulty in treatment and even death. Therefore this study helps patients in diagnosing the diseases timely. Therefore, diagnosing of disease in early stage at appropriate hospital helps patient to solve this problem from aggravation.

One of the problem currently facing health care/hospital organization is the provision of quality services at reasonable price. Quality services can be related to customer satisfaction, correct diagnosis, and proper management of patient treatment. Disproportion between health care workers and patients causes workload on workers. Due to overwhelming work load, doctors/health care workers are too busy, this can lead to ineffectiveness in their career[16].

This cause poor services such as poor diagnosis. Poor diagnosis causes poor decision that lead to disastrous effect which are unacceptable in medical profession. Additionally, since there is a shortage of senior medical expert who has experience in treating cancer in Ethiopia, health workers were trained and distributed in various state and remote regions of the country[25]. But, diagnosis of disease like cancer is very tough with knowledge acquired from short training period. Since aspect based opinion summary based on customers opinion presented to user can include both important and not important summary. The existing work in aspect based opinion summary can present users with both important and unimportant/unrefined summary. The integration of opinion summary with knowledge based system is required because we do not have to make users busy/confused with unimportant summary instead it's better to present user with refined /important summary the user exactly seeking for.

The integration of opinion summary with knowledge based system has never seen to the extent of our knowledge. Therefore, this study helps to extract opinion summary based on what user needs rather than opinion summary based on others customers data (i.e. opinion).

The data generated by the health organizations is very vast and complex due to which it is difficult to analyze the data in order to make important decision regarding patient health. This data contains details regarding hospitals, patients, medical claims, treatment cost etc. So, there is a need to generate a powerful tool for analyzing and extracting important information from this complex data.

Increasing health information needs and changes in information seeking behavior can be observed around the globe . According to recent studies 81% of U.S. adults use the Internet and 59% say they have looked online for health information regarding diseases, diagnoses and different treatments . Thus, patients tend to become active participants in the decision-making process. This change in the way of thinking is often referred to as patient empowerment . However, information overload and irrelevant information are major obstacles for drawing conclusions on the personal health status and taking adequate actions [8]. Faced with a large amount of medical information on different channels (e.g., news sites, web forums, etc.) users often get lost or feel uncertain when investigating on their own. In addition, a manifold and heterogeneous medical vocabulary poses another barrier for laymen. Therefore, improved personalized delivery of medical content can support users in nding relevant information. Medical information available for patient-oriented decision making has increased drastically but is often scattered across different sites .

Recommender systems (RS) suggest items of interest to users of information systems or e-business systems and have evolved in recent decades. A typical and well known example is Amazons suggest service for products. We believe the idea behind recommender systems can be adapted to cope with the special requirements of the health domain. The state of the art health recommender system does not provide customer with what they exactly seeking for since they start from summary of data rather than customer interest. This study is therefore, endeavors to fill gaps in Developing Model that recommend suitable Health care by predicting the disease

at first hand[8][9][10].

This study goes to address the following research questions:

- How it is possible to use rules of symptoms, hospital fact based information and aspect based opinion summarization to build rule based knowledge base system for disease prediction and hospital advising?
- How it is possible to develop system that will automatically recommend hospital based on predicted disease on the same interface?
- How it is possible to combine aspect based opinion summarization of local language Reviews(Afaan Oromo and Amharic Language) and hidden knowledge extracted from fact based information?

### 1.1.2 Significance of the study

Based on the significance the study can provide either implicitly or explicitly, an authors has classified into seven important section that are discussed below such as Health care Administrators, Patients, Medical experts, government, other people and other researchers.

- **Healthcare Administrators:** this study helps Health care Administrators to distinguish their strong side and weak side of the service that has been provided. This allow Healthcare Administrators to work hard on their weak side and continue with the same spirit on their strong side. Since, this system can make summary based on users rather than other customers data, it can helps to reduce an effort of analyzing summary. Therefore, this study plays a great role in helping them to provide quality services by producing summary based on users predicted disease aspect rather than others customer opinion data in shortly without confusing users.
- **Healthcare:** according to the researchers this study helps healthcare services where there is insufficient domain expert, -in balancing the distribution between healthcare workers (medical domain experts) and number of patients seeking the services. Additionally. This system helps health care where shortage of facility has seen.
- **Patients:** this study helps patients of two types, patient who already know their disease and new patient. For patient who already know their disease and suffering from getting quality services, this study will recommend them the health care in which they can get appropriate services according to their disease. For new patient, this study helps patients by diagnosing their disease at early stage. Identifying the disease at its early stage helps patients to get diagnosis and treatment timely. Getting treatment timely benefits chance of getting the disease to be converted into chronic/recurrent or sudden death.
- **Medical experts:** this study helps medical experts as long as an advice and consultation is required using hidden knowledge extracted from previous data. Additionally, this study helps medical experts to upgrade their knowledge and experience. Hence, this system consult medical expert by acting like partner in case some decision were required.
- **Government:** Cancer is the sixth most killer disease of productive age in Ethiopia. Late detection of the disease can affect these productive age highly. These can affect the growth domestic product (GDP) of the country. Therefore, early detection of the disease can reduce more or less the effect on GDP (Growth Domestic Product). In sub Saharan Africa Ethiopia has registered worst rank in comparison with other low-income country. The worldwide health problems have seen in these low-income country[30]. Hence, this study has significance in improving quality of health care services[29][3].
- **Dependent peoples:** according to Health Sector Development Program IV draft of Federal Democratic Republic of Ethiopia in 2010, the overall economic dependency ratio for the country is estimated at 93% dependents per 100 persons in the working age group of 15-64 years. According to this draft the number of economically dependent peoples

are very high. Hence, this study helps other dependent people who are economically dependent on others, by detecting the disease at its early stage. Moreover, According to study investigated by WHO in 2015 cervical cancer can affect most commonly productive age than other disease. cervical cancer most commonly affect peoples in working age. Therefore, this study helps in minimizing effect of the disease on other people more or less, by detecting the disease timely and recommending appropriate health care in which independent people can get better services[1][2][30].

- **Other researchers:** In addition, this study will inspire other researchers to work on integrating joint fact based information and opinion summary techniques with knowledge based system in other domain especially in areas where there is shortage of domain experts seen and recommendation is required such as Higher Education, Hotels, Banks, Movie, Computer and Network troubleshooting etc. to acquire knowledge but there exist tremendous data such that knowledge can be acquired and opinion is summarized automatically from the data.

## 1.2 Objective of the Study

### 1.2.1 General Objective

The general objective of this study is to develop Model that recommend hospital automatically based on predicted cancer disease results, by using integrated opinion summarization and fact based information with developed knowledge based system using the hidden knowledge extracted via data mining algorithm.

### 1.2.2 Specific Objectives

To achieve the general objective of this study, the specific objectives are:

- To investigate the feasibility of the study and collect data or prepare corpus for hospital reviews, fact based information of hospitals and patient medical data.
- To select, adopt and develop suitable techniques or algorithm for aspect based opinion Mining of local language( Amharic and Afaan Oromo ), data mining of pre-cervical cancer dataset and knowledge based system.
- To understand the nature of Amharic and Afaan Oromo reviews and develop model and prototype of hospital review domain.
- To combine unstructured aspect based opinion Mining with fact based information and integrate knowledge based system with data mining.
- To develop joint rules that integrate aspect based opinion summary and fact based information with knowledge based system and data mining.
- To evaluate the performance and user acceptance levels of the developed prototype.

## 1.3 Scope and limitation of the study

We select cervical cancer, since this types of cancer are currently increasing in Ethiopia and it attacks mainly productive age people. Additionally, for the reason that early diagnosis of this disease is very important to increase survival rate of the patients.

According to FMOH administrator cervical cancer disease treatment is currently provided in 5 hospitals in addis ababa: Black Lion, ALERT, St. pauls, Zewditu and Tekle Haimanot General Hospital. The study focused on 5 hospitals in Addis Ababa, one public hospitals and other two private hospitals.

We select hospitals in Addis Ababa based on two reasons: The first reason is because, Addis Ababa is capital city, and peoples expectation are high in seeking better service in Addis Ababa than other regional city that have shortage in some medical facility and other medical treatment. Beside this, its wide range of hospitals distribution in distinct sub city in providing different services makes the city preferable[30]. Additionally, patients who are seeking better health care service are high while comparing with other regional state. Therefore, comparing the quality of health care services is must. We select 5 hospitals in Addis Ababa because these five hospitals are the only hospital that are providing the cervical cancer treatment.

The second reason is since our study domain focus on cervical cancer, the treatment and medical equipment required for such types of cancer diseases are only available in Addis Ababa as compared with other regional state. Furthermore, since cancer is most dominant in urban while compared with rural, and based on the Ethiopian health statistics agency registry of 2010-2016 data this types of cancers are available more in Addis Ababa city than others. the study was carried out on five hospitals setting such as Black Lion,ALERT,Zewditu Memorial,TekleHaimanot General and St.Pauls hospital.

This study focus on diagnosing patient with only, cervical cancer. Diagnosing of other disease are out of the scope of this research work. Moreover, the study also focus on rule based knowledge based system other types of knowledge based system such as case based reasoning, Frame based, hybrid are out of the scope of this research work.We have selected rule based knowledge based system since, the algorithms employed in data mining are classification algorithm, and it can be better integrated with classification algorithm than clustering one. In one hand, the integration of classification algorithm with rule based system are efficient than other knowledge based system.in the other hand, integration of clustering algorithm are efficient with knowledge based system other than rules. rule-based was selected because of its easiness to integrate with J48 rules than other.

We have selected hybrid models since the nature of health care services requires both business understanding and data understanding. the classification techniques are used because we have two class for prediction pre-cancer positive and negative.

In this study classification decision trees were used by adopting J48. the study focus only on the cervical data collected from 20-10 up to 2016 of five aforementioned hospitals only. the study also focus on symptoms that is shown at early stage of cervical cancer(Pre-cancer signs).

This study gives an attention for regular opinions only, comparative opinion is out of the scope of this research work. Regular opinion can be expressed in two ways direct and in indirect. Direct expression is, for example, *“the camera of galaxy s7 is amazing!”* and indirect expression is for example, *“by closing the camera I boost the battery performance of my iPhone 6 phone”*. in direct opinion is out of the scope of this research work. Comparative like opinion is not considered for example, *“iPhone 6 is better than iPhone 5”*. Opinion spam detection or fake review detection is not considered in this research work, this opinions are false positive or false negative opinions. Opinion holder detection or the reason for positive and negative opinions are not included in this research work. This research can handles mixed sentence. Mixed sentences is sentence that can be positive in one sense and negative in other. For example, *“the battery life is ‘long’*, in this sense the term *‘long’*, has positive orientation. While in other sense For example, *“the starting time of this phone is ‘long’*. The term *‘long’*,is the only term that can express opinion but in this statement the orientation is negative. This study covers feature detection, polarity identification sentiment classification (i.e. positive and negative polarity) at aspect level and summary of features and sub features with its corresponding polarity.

In organization such as Health care services evaluation based on its aspect can influence increasing of quality services positively. Despite, the quality of health care services can be measured with its individual aspect at finer grained level other level such as sentence level and document level lacks to provide such types of evaluation summary. Hence, the nature of Health care institution force us to choose aspect based opinion summary.

We choose comments that are provided with the two most popular languages spoken in the country, Amharic that is the official language of the country and Afaan Oromo, the language that is spoken by over 35 million peoples of the country because of two reason: the first reason is, detection of opinion that is provided with other language requires the linguistic person who can support with corpus development of opinion and languages. The other is most of the patient who are served in the hospitals are the speaker of this two languages.

This study do not cover document and sentence level (i.e. subjective and objective) sentiment mining levels. This study focus on domain dependent kinds of opinions summary generation. This study do not cover cloud based opinion summarization services. This study also do not gives an attention for semantic oriented domain reviews. Context based sentiment mining and summarization task is considered in this research work. This study is not designed to cover opinion question answering query (*for example, which phone is best in network?*) mining and answering.

This study can't handles typographical Errors (i.e. errors occurred while typing), orthographic words (i.e. orthographic words are used for expressing excitement, happiness for example, *I'm tooooooo happpppppppy, you are so sweeeeeeeeeeeeeettttttt, KKKKKK.* ).

This study gives an attention in aspect extraction for explicit aspect (i.e. aspect that is noun and noun phrase only) and implicit aspect expression is out of the scope of this research (i.e. aspect that is other than noun and noun phrase). This research is not also designed to handle infrequent features for aspect based extraction. Single domain are implemented in this research work.

## 1.4 Contributions of the study

Major parts of the contribution of this study work can be seen as below.

- Data mining models and knowledge based system rules, for cervical cancer based on data collected from healthcare services was developed.
- Integration of Machine Learning (opinion summarization) with knowledge based system.
- Developing Model that provides disease prediction and hospitals recommendation for Ethiopian patients at single interface using both fact based information and opinion summary with knowledge based system Integration.
- Models are designed to provide automatic disease prediction and hospitals recommendation for Ethiopian patients at single interface using both fact based information and opinion summary with knowledge based system Integration.
- An algorithm is developed for aspect detection, opinion detection in hospital domain.
- Development of opinion Mining model for Afaan Oromo and Amharic languages.
- Rule based sentiment mining and summarizing algorithm was developed.
- Development of KBS with graphical user interface which is simple to use and attractive to users.
- Joint algorithm for combining unstructured opinion collected from patients/clients with fact based information collected from different files, documents and users was developed.
- Evaluation of the new proposed model.

- Direction for future work has been created for other researcher who has an interest to work in area of opinion mining integration with other study areas.

## 1.5 Methodology

The following methodologies have been used in the progress of this study.

### 1.5.1 Empirical Research:

Empirical methods enhance our observations and help us see more of the structure of the world. Visualization methods and techniques for drawing sound conclusions from data. Empirical methods cluster loosely into exploratory techniques for visualization, summarization, exploration and modeling and confirmatory procedures for testing hypotheses and predictions [14]. In short,

*Empirical= exploratory plus experimental*

Because empirical studied usually observe non-deterministic phenomena, which exhibit variability, both exploratory and experimental methods are based in statistics. Exploratory statistical methods are called, collectively, exploratory data analysis; whereas methods for confirmatory experiments go by the name statistical hypothesis testing. We have used this method by developing a prototype and we described all the application of or prototype also we have tested our work by using different experimental method.

### 1.5.2 Research Design:

In order to achieve the above stated objectives, the researcher as per the discussion made on model evaluation of DM methodologies in literature review section 2.3, Hybrid-DM process model was selected as a better model regarding its design to suit for academic researches. To understand a model that produces best possible classifier of immunization status of infant, a Hybrid DM process model was used. Hybrid-DM model is an intersection of KDD and CRISP-DM models [7].

The Hybrid DM presented in figure 1.1 consists of six-step Knowledge Discovery Process; i.e. understanding the problem domain, understanding data, preparation of data, Data mining, evaluation of the discovered knowledge and use of the discovered knowledge.

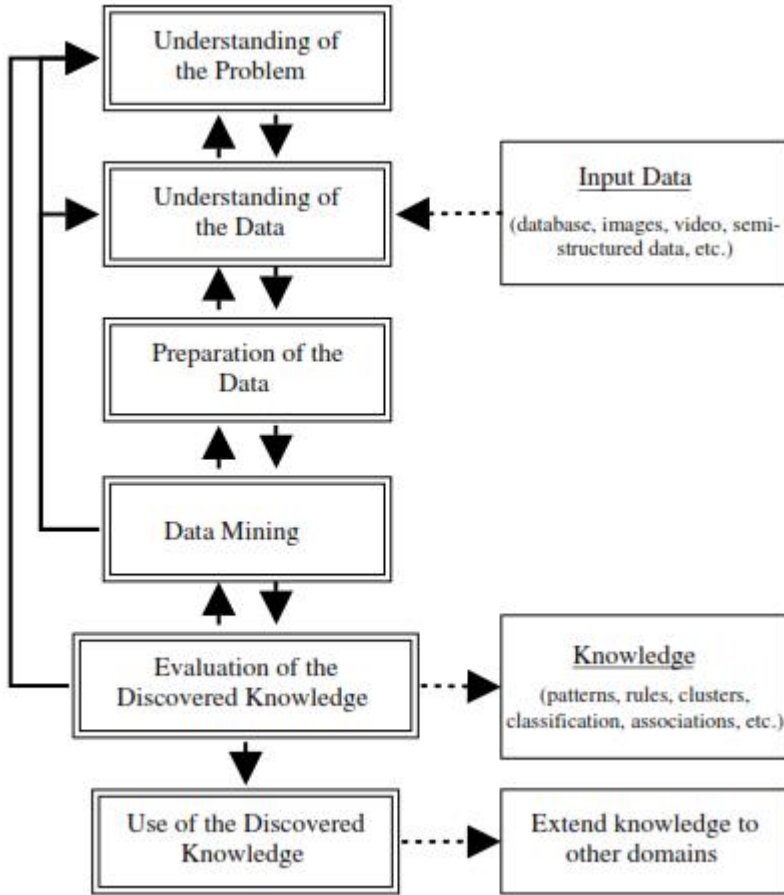


Figure 1.1: Hybrid Knowledge Discovery Model adopted from [68]

### 1.5.3 Literature review

State of the art works that were done by other researchers are referred and cited; based on the study areas, domain area and methodological component. Based on the methodological components, we referred aspect based opinion summarization tasks that were previously done for hospital recommendation, knowledge based system, data mining for factoid information and disease prediction. Additionally, based on our domain areas, we referred healthcare and disease domain that were previously done for hospital/ healthcare services recommendation and cancer disease to gain detail knowledge in the areas. Moreover, research works that were attempt to integrate data mining with knowledge based system are studied.

### 1.5.4 Opinion Summarization

Opinion summarization is one of the discipline comes under sentiment analysis. The goal of an opinion summarization is to make quick and concise summary of review which can be brought from large size of review corpus of entities. This allows any users of the system to make better decision on a given product through getting concise summary in fast manner. Opinion summarization has sub task that need to be accomplished such as aspect detection, opinion detection and actual summary generation.

#### Aspect based opinion summarization

Other classification such as sentence level or document level opinion mining were not provides a notion on what is exactly liked by the customer or what is not. Besides, having positive in polarity of the overall opinion of document does not mean that customer likes everything (all features or aspect) of a given object or entity. Similarly having negative opinion polarity for a given document doesn't also mean that every aspect or feature of an entity or object are hated by the customers.



In a common sense opinion holders can writes a review on a given object or its features. This review may have optimistic or pessimistic polarity. To get an opinion on product at finer grained level opinion mining at aspect level is must. For aspect based summarization there are three commonly used tasks, such as aspect detection, opinion detection and opinion summarization.

**Aspect detection** This task will determine aspect on a given objects of reviews. For example in a review statement that says “the script writer of the movie is amazing”. From this review the opinion “amazing” is given on feature “script writer”. Aspect extraction is under an umbrella of information extraction since information extraction is used for extracting structured information from unstructured sources in automatic manner (For example,3)[3]. Although, the regular used traditional information extraction techniques has already implemented the situation do not allow the same techniques used for opinion mining applications.

**Semantic Based Approach:** This methods can helps to group semantically similar terms in one. This is performed by calculating the similarity between opinion terms. It can be used in application that requires to generate model for verbs, adjectives and noun opinion lexicon and others. Often, in sentiment classification semantic are used in a combination with statistical methods to obtain good results[5].

**Vector Space Model:** Distributional models are used to place words in a fashion they appear inside the corpus. Vector space model is usually implemented under distributional model. The word that appear in the same context were assumed semantically the same. In vector space model the words can be represented as point in space. Vector space model is over predicting the meaning similarity since they use linear algebra[12].

### 1.5.5 Knowledge Discovery Process

This study was implemented by applying Hybrid model for the data mining task. We have selected this model since the nature of health care services requires both business understanding and data understanding.

**Data Mining** This step of hybrid model used for deciding the best data mining algorithm that can fit the model of the study, i.e. prediction or description. This part mostly depends on the former phases of understanding the domain and data. The main aim of the study is prediction of cervical cancer and suggested hospital. Therefore, classification algorithm are used to achieve the goal. Classification techniques are implemented via widely used algorithm such as decision tree, rule-based induction, Bayesian classifiers, neural networks, support vector machine, and k-nearest neighbor. Thus, the study applied, decision tree (J48) predictive algorithms because it is easily understandable by Knowledge Based System as in figure 1.1.

**Implementation tools, libraries and Programming language:** In this study, for implementation of opinion summarization, some programming and scripting language such as Java with UTF-8, JSP (Java Server Page), Servlet, Html, CSS, and JavaScript are employed. We choose this languages like java for two reason: the first one is based on their popularity and platform independency.

The second reason is we are very familiar with this types of languages than others. These languages are used for classifying sentiments, detecting aspects, summarizing opinions and creating interactive web Application GUI (Graphical User Interface). Moreover, tools such as Stanford Parsers, Jfreechart (i.e. used to draw charts for visualizing web server), (Apache Tomcat 8.0.27.0 and Glass Fish Server 4.1.1), <sup>Netbeans IDE 8.1 with JDK 1.8.0\_45</sup> are used for aspect detection

and sentiment classification.

For implementation of fact based information and disease prediction, WEKA 3.6 were used which is the current version of Weka. Weka tool is used for the data mining in order to extract hidden information for knowledge discovering from large data, Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. Weka is also used for data preprocessing[31][32]. This data can be kept in either CSV or (.xls) before conversion to ARFF file format. For storing this data it can be kept in excel file. Therefore, for keeping this data tools like Microsoft office Excels 2013 can be utilized. Weka is free software available under the GNU (General Public License). WEKA (Waikato Environment for Knowledge Analysis) can provides state of the art collection of machine learning and data mining techniques. Recently, this tool is well-known by its wide acceptance in different business and academic domain. Additionally, from the time its hosted online its downloaded in excess of 1.4 billion times by web users[31][32].

For implementation of Knowledge Based System, some programming languages and tools are used. Prolog (programming in logic), is declarative language that is used for knowledge representation. It is smart and robust tool that represent formal and natural language data through automatic memory utilization and its safe semantics. Prolog was developed by Alain Colmerauer and his associates at the University of Ai-x Marseille, in Marseilles, France, in 1972. Its named as programming in logic.

Prolog (programming in logic) is one of the most popular programming language, especially in the artificial intelligence research. SWI-Prolog is used to write rule based prolog programming. SWI-Prolog was first created as an open prolog environment for providing powerful bi-directional interface to C. it is developed with an agreement of object oriented GUI system known as XPCE[31][32].

XPCE is used for developing knowledge demanding graphical applications. SWI-Prolog is most popular tool for prolog implementation. Its most widely used tools that has been currently utilized for developing prototype of knowledge based system model. It comprises services such as back and forward chaining inference engines. We have chosen prolog because of two reason first, its widely available second, its familiar with us[31][32].

For integration of joint opinion Summary and fact based information with knowledge based system and user interface some programming languages and libraries are used. Programming language such as Java and prolog are employed. JPL 3.x Prolog API (Java Prolog library) and Weka API were among library that are used for integration of both. Basically, JPL is Bi-Directional (i.e. it allows you to call Prolog from Java and Java from Prolog)[31][32].

**Data Sources:** In every scientific community evaluation of a given system is required. Therefore, in order to develop and test the prototype of our developed model collecting data set for both training and testing is desired.

**Data source for Hospitals Recommendation:** Recommendation of hospitals mainly use fact based information and customer satisfaction information to minimize the one-sidedness of the recommendation. Therefore, for Recommendation of hospitals data of both opinion summarization and fact based informations are collected using different techniques of data collection as discussed below.

**Data source for opinion summarization:** For training semantic based approach selected yet, since the method we followed is semantic based approach we have collected 1500 review sentence that consists of 750 positive and Negative 750, which are composed of various aspect from online site: <https://www.patientopinion.org.uk/>. Afaanoromoo dataset were collected in two versions: the first version is for training set and the seconed version is for testing set. In one

hand, the dataset for training was Prepared in two forms:the first one is the data from online English patient opinion is translated into Afaanoromoo using English to Afaanoromoo language translator. The translated opinion was provided for afaanoromoo linguistic experts to correct its language meaning and the polarity of the sentiments. Then the corrected opinions were used to train the model with selected hybrid approach.

The other was collected manually from patient (customer) of the hospitals by using purposive sampling techniques on cancer department of hospitals.We have used this sampling technique because our focus domain is cancer. In the other hand,the datasets (reviews) used for conducting the experiment are collected manually from five hospitals follow-up patients.

The Evaluation or Test data were collected from five hospitals in Addis Ababa that were currently engaged in providing cancer care, the doctors were responsible for this particular activity. The involvement of doctors was due to the sensitiveness of the data regarding privacy of the patient. Totally 70 opinion from different customers of each hospitals were collected. The opinion collected from five hospital were 350. From this opinion data were provided in different language that were spoken in Ethiopia.

From this language, around 150 opinions were provided in Amharic language. 100 opinions were provided in Afaan Oromo language. And others opinions were provided in other Ethiopian local language. Since other opinion is out of the scope of the study we do not deal with it. Then Rule based translation approach were applied due to less resource in local language like Amharic and Afaan Oromo language. Hence, translation made to English language for collecting English Amharic corpus. Then this corpus showed to linguistic expert and approval was made. We have also showed Afaan Oromo English translation to linguistic expert for approval. This approval was also made for its opinion.

To accomplish this task we have collected training dataset from different websites with hospital review data. Most of dataset collected were raw data (unprocessed). Therefore, we have to process those unprocessed dataset to make our system efficient. Some of raw data set for training was sent by hospital owner through e-mail from website <https://www.medicare.gov.html> . Other raw dataset for evaluation were collected manually from users (patients) of the hospitals through questionnaires. This data set mainly target five hospitals that are found in capital Addis Ababa: two public and one private hospital.

**Data source for Fact based information:** For fact based information, we have collected training and testing dataset from different sources such as users, study s, archives and database. Raw data are collected from users manually using questionnaires and interviews techniques of data collection. The collection of different dataset with different data formats are combined manually by the researchers. This data set also target mainly five hospitals that are found in capital Addis Ababa: two public and one private hospital.

**Data source for Disease Prediction and Advice:** In order to get interesting rules (hidden knowledge) and domain knowledge of medical advice on treatment, data source are collected in two forms: data mining and knowledge based system forms as discussed below.

**Data source for Data Mining:** For cancer disease, raw data are collected from five different hospitals using different sources such as users, study s, archives and database. The collection of different dataset with different data formats are combined manually by the researchers. This data set also target mainly two cancer type that are killer in Ethiopia: cervical cancer.

**Data source for Knowledge based system:** In order to get domain knowledge for cancer disease, we have collected dataset from different sources such as experts, web documents, archives and books. Raw data are collected from domain experts (specialized doctors) manually using questionnaires and interviews techniques of data collection. The collection of different dataset

with different data formats are combined manually by the researchers. This data set focus mainly two cancer type that are killer in Ethiopia: cervical cancer.

**Evaluation methods** In every scientific community evaluation of a given system is compulsory. Hence, for this study two types of evaluation mechanism have been used: system performance Testing (SPT) and user acceptance testing (UAT)[33].

**System Performance Testing (SPT) for opinion summarization:** The common standard of information retrieval (IR) effectiveness evaluation techniques are used in order to evaluate the developed prototype of the system. Precision, recall and F-measure metrics are used in order to measure the retrieval effectiveness. Precision is a metric used to provide ratio of the number of relevant retrieved document to total number of document retrieved. Recalls is a metric used to provide ratio of the number of relevant retrieved document to total number of relevant document.

F-measure is another metric that considered both precision and recall this this metric often performed by harmonizing both precision and recall together. Hence, for evaluating the performance of developed opinion summarization system Precision, recall and F-measure metrics are used for this study.

**System Performance Testing (SPT) for Data Mining** Precision, recall and F-measure metrics are used in order to measure the retrieval effectiveness. Hence, for evaluating the performance of developed Data mining system Precision, recall and F-measure metrics are used for this study

**User Acceptance Testing (UAT):** The main objective of performing user acceptance testing is verifying to what extent the developed system satisfy users desires concerning their usability and acceptance. For this reason, test case are prepared for testing the performance of the whole system[63].

## 1.6 Organization of the study:

Brief description on the organization of document structure is given as follow. We cluster this study mainly into Eight chapters.

Chapter one: this section of the chapter provides brief discussion concerning introduction part of the study, by giving general overview about the focus area of the study. Additionally, it gives emphasis on sub parts such as opinion summarization, data mining, fact based information, knowledge based system, cervical cancer, healthcare services (hospitals), statement of the problem, objective of the study, significance of the study, motivation, contribution of the study, methodology used and evaluation methods.

Chapter two: this section of the chapter offers brief discussion of previous and current related literature that are related to the topic of the study.

In this chapter, related literature in opinion summarization for hospital/healthcare recommendation, customer satisfaction, fact based information, component of opinion mining, levels of opinion mining, challenges in opinion mining, aspect based opinion. Additionally, works related to hospital recommendation system using opinion summarization and cervical cancer disease prediction using data mining techniques are included in this chapter.

Chapter Three this section of the study deals with data preparation and preprocessing techniques and steps to produce quality data that is suitable for analysis carried out by weka tool. Chapter Four: this section of chapter briefly discuss about our idea in developing model for opinion mining in two language Afaan oromoo and Amharic language, developing model for fact based information and merging opinion summarization and fact based information.

Chapter Five: this section of chapter briefly discuss about developing integrated data mining techniques and knowledge based system for disease prediction.

Chapter Six: this section of chapter briefly discuss about integration of joint opinion summary and fact based information with data mining and knowledge based system.

Chapter Seven: This section of chapter briefly discuss about evaluation results and discussion.

Chapter Eight: This last chapter briefly discuss about conclusion and recommendation.

# Chapter 2

## LITERATURE REVIEW

### 2.1 Introduction

Every text based information available online can be classified into two major classes, facts and opinions. Facts are an objective information about the given object or event while opinions are subjective information about a given object or entities. Subjective is about reflecting ones perception or views on particular entities. Most of the work done previously were focused on facts such as web search, information retrieval etc. in contrary to this little attention were given for opinions compared to facts since, there was a few user generated content before the coming of web 2.0.

Previously, people used to make decision based on the opinion of friends or families. For an organization to decide about a given product or service conducting survey or target group is common. But, after coming of World Wide Web especially web 2.0 the world has transformed. Today, nobody is tired of asking a friends for opinions toward particular product or services to make decision since huge volume of information about a given entities are available online. In addition to this there is no organizations are participating in conducting survey or focus group in order to know what others say about themselves[89].

This is because of huge volume of review information that are produced online on different social networking sites, discussion forums, blogs etc. since the size of an information is very huge, it's very tough for ones to read, understand and produce useful summary of every review information provided on different aspect of the entity. Therefore, to tackle this problem an automated opinion mining and summarization system is required. This chapter discuss area of research on the general concept of opinion mining, sentiment classification, aspect based and non-aspect based opinion summarization with their techniques that were utilized[10].

#### 2.1.1 Opinion Mining

Opinion mining is an emerging area of study under the umbrella of text mining and natural language processing this day. After coming of the web, users can share their experience on particular product or services by reflecting reviews. The reviews can be posted online through blogs, movie, products by reviewers toward objects such as individuals, organizations, product, services events etc.[7].

The review posted online by web users can not only limited toward the entity it can also be given for component or attribute set of an entity. This reviews are increasing in an alarming rate this day. Hence, large volume of information can be created online (i.e. user generated content). This created user generated content needs to be mined to analyses. Therefore, the process of finding opinion, attitude, emotion, appraisals of opinion holders about a given topic from huge amount of created information is known as opinion mining. Since, huge volume of opinions are available on the web this day many researchers are inspired [15][30].

Opinion mining is also known as sentiment mining, sentiment analysis, emotion detection and review mining. Different researches are undertaken under this field of study such as: subjectivity detection, opinion summarization, comparative opinion mining, opinion spam detection and

opinion question answering, Entity and feature detection[30].

**Component of Opinion Mining** Basically there are three or four component of opinion mining as defined by other researchers[35]: Opinion orientation: opinion orientation defines the opinion given on particular object as positive, negative or neutral. *For example, “According to Miles the battery life of iPhone that he bought yesterday is no longer to stay”.* This statement state that the battery life is short. Thus, the word *“no longer”* in this statement shows sentiment orientation which is negative.

Opinion Holder: is writer or an author of the opinion that express feeling on some particular object. From the above statement Miles is an opinion holders or writer of the opinion. Opinion Object: is an object on which an opinions are expressed by writers. On the above example iPhone is an object on which an opinion is expressed. Other customer: are other individuals or organization that uses this opinions which are expressed by opinion holders to make decision.

**Levels of opinion mining** Sentiment mining can be classified in to three levels of granularity such as: sentence level, document level and aspect level. Sentence level sentiment mining task is undertaken to classify reviews sentence in to two concept subjective and objective. Subjective is one that can be classified in to either positive or negative sentiment orientation. This can answer the question that is this review sentence is positive or negative? While the objective ones are facts therefore, it can be categorized as neutral (i.e. no opinion). Example-1, *“I bought an iPhone yesterday and its great!”* This sentence shows subjective concept since it contains opinionated words ‘great’ which is positive. Therefore the sentence has positive sentiment orientation. Example-2, *“I bought an iPhone yesterday and its bad!”* This sentence shows subjective concept since it contains opinionated words ‘bad’ which is negative. Therefore this sentence has negative sentiment orientation. Document level sentiment mining task helps us to know the overall sentiment given over particular entity by an opinion holder. This can answer the question whether review document is positive or negative.

In document level there is no way to know opinion orientation on each and every individual aspect.it is simply providing the general sentiment orientation for review document. Let us see an example provided by an authors on one products reviews. Suppose this is the review document, (*“the battery of this Samsung phone is short”, “the Network of this Samsung phone is amazing”, and “the camera of this Samsung phone is bad!!”*).

In document level there is no way to know in which features the reviews have positive and negative opinion but the only thing it does is giving aggregate value negative since the negative review is greater than the positive one in this particular example. Thus, in document even though review has an amazing network features it doesn’t care about the component of the object instead object itself. In both sentence and document level it is not possible to know which features exactly is loved and hated by an opinion holders.

In aspect level this things are possible since the aspect level opinion mining deals with the components or attributes of an object (e.g. **Product, Movie,Hospital etc.**) at finer grained level of analysis. See an example followed. *Example-1 (“I bought an iPhone three days and battery is still working!”)*.From this example we understand that an opinion is expressed on iPhone component battery that it’s good rather than iPhone itself. Therefore in aspect level one is able to know an opinion of an object somehow at detail level. This can answer the question whether a given review sentence feature is positive or negative.

### **Opinion Summarization**

Opinion summarization is one of the discipline comes under sentiment analysis. The goal of an opinion summarization is to make quick and concise summary of review which can be brought from large size of review corpus of entities. This allows any users of the system to make better decision on a given product through getting concise summary in fast manner.

Opinion summarization has sub task that need to be accomplished such as aspect detection, opinion detection and actual summary generation. Most of existing work classify opinion summarization into two broad categories such as abstractive summarization and extractive summarization. In abstractive summarization understanding the concept that are expressed by an authors in a review document in well manner is needed.

In these, understanding the reviews in short format is required. In order to detect crucial idea abstractive summarization employ many natural language processing techniques on review source. In extractive summary, concatenation of important sentence together is carried out first by choosing crucial sentence from reviews source. After these concise summary will be generated[36][12].

**Aspect based opinion summarization** Other classification such as sentence level or document level opinion mining were not provides a notion on what is exactly liked by the customer or what is not. Besides, having positive in polarity of the overall opinion of document does not mean that customer likes everything (all features or aspect) of a given object or entity. Similarly having negative opinion polarity for a given document doesn't also mean that every aspect or feature of an entity or object are hated by the customers. In a common sense opinion holders can writes a review on a given object or its features. This review may have optimistic or pessimistic polarity. To get an opinion on product at finer grained level opinion mining at aspect level is must. For aspect based summarization there are three commonly used tasks, such as aspect detection, opinion detection and opinion summarization[36][12].

**Aspect detection** This task will determine aspect on a given objects of reviews. For example in a review statement that says **“the script writer of the movie is amazing”**. From this review the opinion **“amazing”** is given on feature “script writer”. Aspect extraction is under an umbrella of information extraction since information extraction is used for extracting structured information from unstructured sources in automatic manner (For example,3)[7].

Although, the regular used traditional information extraction techniques has already implemented the situation do not allow the same techniques used for opinion mining applications. There under, for opinion mining application it is used for opinion information such as reviews, blogs, forums etc. since, it has complex nature due to noisy and complex characteristics of reviews.

Thus, it's important to develop techniques that are particular to an opinion mining. This circumstance can be used also for entity extraction. Since most of applications are focused on online reviews ours are not exceptional. Basically, two common format of reviews are most commonly used online under this circumstance.

**Format 1: pros, cons and detailed description: an authors are required to provide some discussion on the advantages and disadvantages of a given topics or objects by separating them both.**

**Format 2: Free Format: here an authors are expected to write freely no more separation of advantages and disadvantages of a given topics are required. Since format 1 consists of one aspect in a segment it's simple. Therefore, our attention is toward format 2 that can be freely expressed by an authors[7].**

aspect based opinion summary is based on opinion from various opinion holder since opinion from single opinion holder has practically less importance in most of the opinion mining application[7].consequently, we need some form of summary which comprises both quantitative and qualitative summary. We can discusses an examples which is used in [6] as the following. This is an example that shows the use of aspect based opinion summary in most of an application used in an organizations.

**Opinion Prediction** Opinion Prediction is the most commonly studied area under natural language processing field of study. Having a set of review document D, it helps to predict the



polarity of the individual document  $d \in D$  on a given object or topic. For instance, Polarity prediction by system for movie review into two granularity such as positive and negative.

This kind of classification are distinct from the regular classification of topic based classification, since it cannot utilized to classify document under defined class topic e.g. Politics, science and sports. Here, for topic based classification the words that gives notion for topic should be there while for opinion prediction no need of words that gives notion to classify under particular topic classification.

But, for opinion classification there is a need of opinion terms that shows opinion either positive or negative e.g., great, amazing, stylish, best, bad, awesome etc. most of works done on sentiment classification utilize machine learning techniques for classification. The classification that is used there in before is at document level since individual documents are used as basic input for the system to classify it into given classification.

An assumption is given under sentiment classifications are said, each individual evaluative document are taken as opinion against single object  $O$  or topics that can be expressed from single author since it contain single opinion. Therefore, opinion prediction is utilized for classifying polarity of particular opinion expressed by opinion holder on a given object  $O$  under evaluative document which can satisfy an assumption made above.

Studies have also been done on sentiment classification at sentence level and aspect level. For sentence level sentence classifications are objective classification and subjective classification. Sentence level classification do not gives an attention for an object on which an opinions are expressed rather gives more consideration for objective and subjective classification. It also do not care of compound sentence (i.e. sentence that comprises more than one opinions). *For example, “The picture quality of this camera is amazing and so is the battery life, but the viewfinder is too small”*[10].

In opinion summarization, opinion prediction step come after sentiment aspect detection step. In fact it is a hot research area in opinion mining we are going to discuss sentiment prediction that were carried out under sentiment summarization. Under this circumstance sentiment prediction is not only limited to detecting the sentiment orientation of review instead the corresponding aspect of the review.

In here, different people could have different opinion on the same aspect. This happens since opinion reflected by different people on the same aspect can be different. For example, *“the camera quality of the Sony camera is amazing ”* and *“the camera quality of the Sony camera is awful”*. This nature differs sentiment predication under opinion summarization from others[14].

**Lexicon/Rule-based Methods for Sentiment Prediction** This methods have been used by various researchers in opinion summarization. The method requires dictionary of words or positive and negative seed list to predict opinion. This seed list/dictionary of words can be matched with corpus to determine the sentiment orientation of a given review.

In some cases this lexicon can be utilized with pre-defined set of rules and pos taggers or parser. This method were seen favorable under the product domain since an explicit expression has taken place. This method obtained poorer result under movie review domain since contextual meanings are unnoticed. Finally, this method were decided based on the quality of seed list. Each domain require separate dictionary for best case[14]

**Summary generation** This step is followed after sentiment prediction and aspect detection step has carried out. Based on the result obtained the summary will be generated. This summarization can help users to get brief and quick summary of the reviews supplied. There are various methods for summary generation. For summary generation parsed review sentence are required.

This review sentence can comprises an aspect on which an opinion is given and its corresponding opinion. In previous section, the task of classifying an opinion extracted into positive

and negative granularity were accomplished. Here, clustering can be made for an aspect that are synonyms. Then, after the positive and negative opinion with corresponding aspect are clustered the graphical and text summarization of an aspect on a given object with its corresponding opinion can be generated and displayed[1][9][34]. Some methods can be discussed in the subsequent section as below[14].

**Statistical Summary** This types of summary is the most commonly used for various business activities. The results that were obtained in sentiment and aspect prediction steps are supplied are presented statistically. This presented results can help users for comparing and contrasting product of their needs[14].

**Text Selection** Despite, providing the general overview on an opinion is crucial for users to make decision, text based opinion summaries are also some times important since it helps in reading the small substance of text. Under this circumstance different types of summary are included. This can be word, phrases and sentence. Sentence level can give better understanding on the topic[14].

**Aggregated Ratings** This types of summary are formed based on the combination of statistical summary and text selection[8].

**Summary with timelines** Opinions can be changed in accordance with time. Providing overall summary on the recent review is very crucial to have latest decision since reviews that can be provided by an online users before one day and after one day can be different[8].therefore, this study focus on this types of summary since the hospital domain requires recent recommendation based on changed facts.

### **Sentiment classification Techniques**

Commonly, there are three general classification approaches for sentiment: machine learning techniques, lexicon based techniques and hybrid techniques. The popular machine learning algorithms are used for machine learning approaches. For lexicon based techniques an opinion lexicons are used. This lexicon based approaches can be classified into two approaches dictionary based approaches and lexicon based approaches.

Statistical and semantic methods can be applied in order to predict sentiment orientation. Hybrid approaches are most commonly used approaches known in combining both lexicon and corpus based approaches. Machine learning approaches broadly classified into supervised and unsupervised learning approaches. The supervised one carried out with opinion words labelled with its corresponding polarity tags.

In a situation where there is challenge in getting this training dataset unsupervised learning techniques are used. In lexicon based approaches an opinion lexicons are used for analyzing an opinions. Dictionary based approach can perform the task of predicting sentiment by considering its synonyms, antonyms. While the corpus based approach use large size of sentiment lexicon. This sentiment lexicon can be also used to predict the sentiment of the given review and its context. This is performed using statistical and semantic practices[30].

There is general understanding assumed regarding sentiment classification. Some of an assumptions are discussed as below. For document level sentiment classification its assumed that an opinion document  $d$  consists of a review that has given under a particular entity  $e$ . This reviews can be forwarded from single opinion holder  $h$ . this review can be the review that can be expressed against a given products or services[11].

Usually this kind of reviews are expresses by single authors. This kind of reviews will exclude blogs and forums since the situation could allow writing reviews on different product and comparisons. Most of the paper has worked on sentiment classification at document level employ supervised techniques. This do not mean unsupervised learning were not used even though this techniques are used not as the supervised one[11].

Sentiment classification can be seen as problem of supervised learning since it comprises three

classes such as positive, negative and neutral. Review dataset such as product review or movie review can be used for training and testing system. In example [11] the review rated from 1-2 are considered as negative, the review rated 3 are considered as neutral while review rated from 4-5 are considered to be positive[11].

As discussed in example [7] classification techniques such as naive Bayes and support vector machines (SVM) were used for classification in previous work thus, no exception for sentiment classification. Some researchers use this methods to classify movie reviews into positive and negative. Earlier various features have been used for learning purpose. Some recently used features are discussed below:

**Terms and their frequency:** in one case, the frequency of terms in a given document, n-gram of word and word itself can be used as a features. In other case the position at which words are found also considered. Other than this other techniques that are effective in sentiment classification that used in information retrievals such as TF-IDF are also used[11].

**Part of Speech:** adjectives were used as indicator of an opinions in various research since most of the opinion terms are an adjectives in nature[11].

**Opinion words and phrases:** opinion words are words that are used as an indicators of either positive or negative opinions. For example words such as ‘beautiful’, ‘wonderful’, ‘nice’ and ‘amazing’ are positive opinion and ‘bad’, ‘useless’, ‘poor’ are negative opinions. Most of an indicators of opinion terms are adjectives but, there is also non-adjective words that are an indicators of an opinion[11].

These non-adjective can be words that are adverbs (e.g., clearly), verbs (e.g. hate and like) and nouns (e.g. rubbish, junk and crap).other than this there are also an opinion phrases and idioms that can be expressed e.g. it costs somebodies arm or legs[7].

**Negations:** since, the presence of negation terms can change the polarity of it is required to be handled. For example, the review statement that says I dont like the camera of this phone has negative polarity since the positive opinion terms are negated by term dont. Therefore, negation should be handled in a review sentence very carefully[11].

In one hand negation terms may come positive and in the other hand it comes as positive. For example, the review statement that says camera of this phone is not bad has positive polarity since the positive opinion terms are shifted by term not. Syntactic dependency: various researchers also seen dependency based word sentence parsing or creating parsed tree[11].

**Machine learning techniques** Machine learning systems are used to get and integrate new knowledge automatically. Analytical observation, training, experience, and other means allows system to learn. It allows the system to develop self-improvement, effectiveness and efficiency. The knowledge obtained from the system could also be examined through testing[10][11].

This approaches depends on the popular machine learning techniques to solve sentiment analysis problem. This task of text classification are performed using some syntactic and language features. According to [5] text classification problem is defined as given a set of records  $D=x_1, x_2, x_3, x_n$  are given a tagged polarity class.

Task of classifying text can be categorized under sentiment analysis into hard and soft classification problem. Hard classification problem is when one tag is given for given sentiment words while the soft one is when tag is determined as probabilistic value and assigned to a given sentiment terms[30].

**Supervised Learning** Supervised learning approaches requires labelling of training data set with the given class. Previously, many researchers have been used SLA. In this section we are going to present some of the most commonly used classifiers in sentiment analysis. They can be categorized under probabilistic classifier and linear classifier

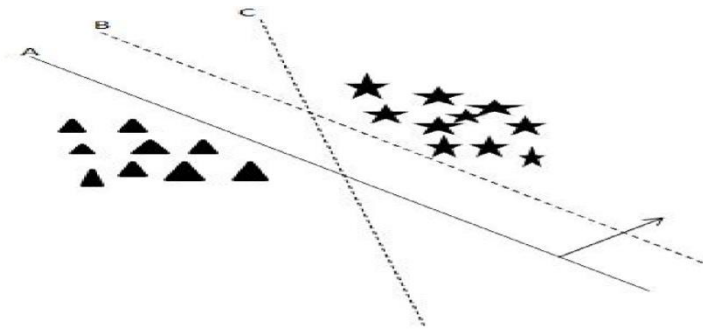


Figure 2.1: Support Vector Machines Classifiers

**Probabilistic Classifier** Here under, two commonly used probabilistic classifiers such as naive Bayes and Bayesian network are discussed below

**Naive Bayes Classifier (NB)** Naive Bayes classifier is used by many researcher since easily understandable. This method use Bayes rule in order to classify a given instance to its appropriate class. It works based on the lexical probability of a given document[5].

$$P(\text{label}|\text{features}) = \frac{P(\text{label}) * P(\text{features} | \text{label})}{P(\text{features})}$$

$P(\text{label})$  is defined as the probability of likelihood of a label.

$P(\text{features} | \text{label})$  is the likelihood probability of a given class appeared as label.

$P(\text{feature})$  is the likelihood probability of feature.

**Bayesian Network (NB)** NB classifier assumed mainly the features independence. NB has also assumed that every individual features are completely depend to each other. This a cause for the appearance of a model known as Bayesian Network model (BNM). It is directed but not cycle graph, it is node are stands for variables while an edges are stands for conditional dependencies. For variables and relationship it is reflected as complete model. This model is rarely used in text mining since, it is not efficient[30].

**Linear Classifiers** Linear classifier can be denoted with formula  $p = \bar{A} \cdot \bar{X} + b$  where, p is stands for prediction. It is used for separating hyper plane between distinct classes.  $\bar{A} = \{a_1 \dots a_n\}$  This represent linear coefficient vector and b is for scalar.  $\bar{X} = \{x_1 \dots x_n\}$  , is used to represent normalized word frequency. There are various types of linear classifiers were used in sentiment classification from them Support vector machine is better in separating linear between labels. In this section the two popular classifiers are seen below[30].

**Support Vector Machines Classifiers (SVM)** The challenge task of SVM is choosing where to separate linear classes. In figure below we have three hyper planes such as A, B , C and two classes x and o. the function of hyper plane is to determine the best separation of the classes[5].

**Weakly, Semi and Unsupervised Learning** The core objectives of this types of learning is to cluster a given document into many classes. In large document with many clusters it is difficult to obtain classified label, since, making label in situation where many clustering is available is not practical[5].

**Lexical Based Approach** Sentiment terms can be classified into two positive and negative. The positive ones are used to show encouragement while the negative ones are used to show discouragement. Additionally, phrases and idioms all are known as sentiment lexicon. There

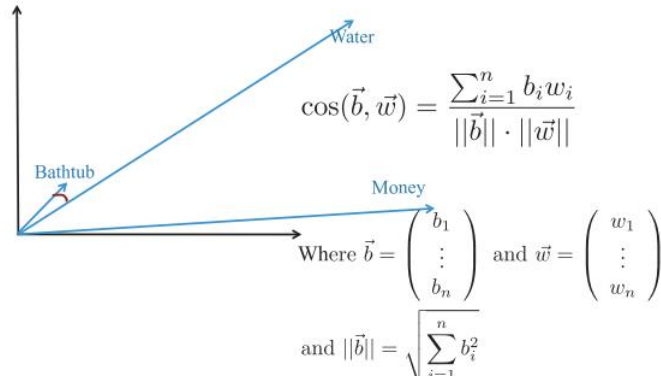


Figure 2.2: Vector Space Model

are various methods are employed in order to collect opinion lexicon. The challenge task is the manual one which is not practical[30].

**Dictionary Based Approach** In dictionary based approach an opinion word list is collected, then the process of finding for the term siblings using synonyms and antonyms in the Word Net and thesaurus is applied. The process of adding new opinion lexicon to the list ends when no more new words are found. Here, the existence of error is checked manually. The main shortcoming of this approach is, it doesn't work for the context words and domain dependent terms

**Corpus Based Approach** It is less effective than dictionary based approach when used isolated but in combination with other approach corpus based approach is effective especially for context and domain dependent opinion. This approach can be carried out in two ways, statistically and semantically.

**Statistical Based Approach** This approach consider the sentiment word list or the co-occurrence between them. It is carried out through probability of likelihood from the corpus. Using indexed document unseen lexicon is predicted. It is performed using occurrence of sentiment lexicon in a given document. If it appears frequently in positive document it is assigned positive sentiment orientation. If not it is assigned negative orientation. But if occurrence of positive lexicon and negative lexicons are equal, it is assigned neutral orientation. Therefore, the semantic orientation are determined by the frequency in a given document[30].

**Semantic Based Approach** This methods can helps to group semantically similar terms in one. This is performed by calculating the similarity between opinion terms. It can be used in application that requires to generate model for verbs, adjectives and noun opinion lexicon and others. Often, in sentiment classification semantic are used in a combination with statistical methods to obtain good results[30].Therefore, in this study the semantic based approach are used with combination of our developed rule based algorithm.

**Vector Space Model** Distributional models are used to place words in a fashion they appear inside the corpus. Vector space model is usually implemented under distributional model. The word that appear in the same context were assumed semantically the same. In vector space model the words can be represented as point in space. Vector space model is over predicting the meaning similarity since they use linear algebra[12].

**Non-Aspect Based Sentiment Summarization** NABSS requires different representative formats of sentiment summarization. This format can be integrated with aspect based opinion summa-

rization techniques. NABSS comprises various cluster such as advanced text summarization, basic sentiment summarization, visualization, and entity-based summarization[14].

**Basic Sentiment Summarization** This types of non-aspect based opinion summarization considers sentiment classification scores for presenting summary. This could be performed by counting the predicted polarity value of a given input review. Aspect identification steps is not included in this types of non-aspect based opinion summarization. Furthermore, the result generated cant show finer grained level of sentiment. Therefore, this kind of summary do not help customer to know about particular product at finer grained level.

### 2.1.2 Knowledge-Based System

Knowledge based system (KBS) is a discipline that has emerged from field of study known as Artificial Intelligence (AI). AI is the system that is used in order to solve particular problem. KBS comprises archive of expert knowledge with functionality planned to provide the knowledge retrieval if certain queries are available besides providing learning and explanation to make the expert know another knowledge domain[22][33][37][38].

Specifically KBS, give emphasis in providing knowledge based technique in order to make human decide, learn and use it for action. KBS can use knowledge generated from data, information and knowledge. This system knows the information that are being processed and it can use this information to make decision. Other system such as transaction processing, management information system do not know the information/data they are being processing.

KBS is a program that act as a domain expert and it can be used when required at anytime, anywhere without wasting time. According to [37][19] such systems are capable of cooperating with human users and are being used for problem solving, training , and assisting users and experts of the domain for which the system are developed, in some cases KBSs are even better than human, as they are enriched with the qualities of efficiency and effectiveness[16][17][18].

The main objectives of knowledge-based system is to act as a domain expertise in the place shortage of domain experts has seen, by providing quick response to make users decide on the output and justification provided by the system. Movable computers are enhanced with depth knowledge to solve specific problems at given domain. Knowledge based system can reduce memorization and discussion time by leveraging an experts, it can also make users to function at higher level and reduce consistency. It is useful tool that provides joint knowledge of one or more experts[37][19][17].

#### Objectives of Knowledge based system

According to [66] the objective of knowledge based system are discussed as below:

- Provide high level of intelligence by producing new situation
- Obtaining new observations
- Providing huge amount of knowledge in various domain application
- Providing software development productivity that is vital Mainly minimizing cost in terms of money and time in constructing the system that is automated.

**Advantage of knowledge based system** The basic advantages of knowledge based system are discussed as below: Stable knowledge documentation: knowledge engineer represent knowledge base based on knowledge extracted from domain expert and relevant document by using knowledge representation techniques. This knowledge can be stored permanently for the user to access and use this knowledge anytime its needed in the future[22][37][38][39].

Effectiveness and efficiency: knowledge based system are more efficient than the domain expert and as effective as domain experts. This is due to the knowledge base elements that are combined to make knowledge based system provide the right decision. Efficient, since knowledge based are deployed on computer[18][22][38][33][39].

Reliability and consistency: the occurrence of error rate decrease with the growing level of knowledge. Moreover, knowledge are combined to make knowledge based system as effective as domain expert. Knowledge can also be accessed speedily with reasonable justification[37][22][33].

Justification for decision made: the ability of domain expert can be judged by his ability to give an explanation for the decision made. Knowledge based system encompass aspect of providing explanation on decision made for convincing end-users. Self-Learning and easy update: this self-learning are carried out in knowledge based through the use of inference engine. Inference engine helps to make knowledge based system learn from experience especially new knowledge. Knowledge based system learn new knowledge from experience through automatic machine learning and manually by domain experts.

**Limitation of knowledge based system** Partial self-learning: knowledge based system can provide justification on decision made by itself and self-learn by updating its knowledge. The knowledge may not be represented totally, for this reason partial learning occurred. Domain experts can learn new circumstance by themselves but knowledge based system should be directly updated[18][22][38][33][39].

Creativity and innovation: knowledge based systems are not as creative as human being this is due to lack of five basic common sense. Human being can comprise of five common sense such as listening, smelling, testing, visioning and touching. Including five common sense to knowledge based system is not possible because computer can understand symbolic input[18][22][38][33][39].

Development methodology: the absence of commonly accepted standard make the system development challenging task. Even though, there is a common steps for development life cycle, there is no common development model that is provided for knowledge engineers to construct knowledge based system model.

Knowledge acquisition: representing knowledge using knowledge acquisition techniques in a way its utilized by the knowledge based system. Knowledge is subjective in nature and its very difficult to extract such type of knowledge that is hidden in domain expert mind.

Development of knowledge based system and testing strategies: Before the representation of knowledge into knowledge base using knowledge base representation techniques, the knowledge acquired from different source should be tested. Even though, acceptance is required during knowledge base representation the absence of standardized testing mechanism make it difficult for testing[18][22][38][33][39].

### **Structure of Knowledge-Based System**

Knowledge based system are responsible for acting like an expert that can make a decision and provide an appropriate explanation. Knowledge based system are used to represent particular domain area. It can be seen as an effective tool since it contains knowledge combined from various domain experts and documents. According to Sajja and Akerkar[37][18][22][38][33][39] knowledge based system are classified into five main component. This components are illustrated in figure 2.2 and are discussed as below.

**KNOWLEDGE BASE:** To make knowledge based system act as an expert the knowledge base should include particular knowledge under a given domain. This knowledge can be taken from human mind and presented in knowledge base in a way its understood by a machine.

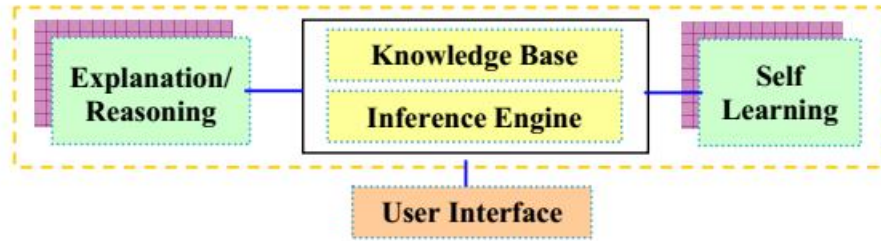


Figure 2.3: Structure of Knowledge-Based System

Various techniques are used in order to represent knowledge in knowledge base, one of the technique is rule. Rule can be represented in IF-THEN format. If certain conditions are met then a given consequence are executed. Conditions are expressed in IF part and consequences are expressed in THEN part[18][22][38][33][39].

Example of rule:

Rule: IF symptoms of diabetes= “diabetes related symptoms”  
 AND lab test result using FPG= “ $\leq 99$  mg/dL”,  
 THEN diabetes type= “Diabetes Free”

Given fact: the patient has symptoms related to diabetes and the laboratory result show us FPG test of 99 mg/dL.

Conclusion: the patient is diagnosed as free from diabetes.

This rule has knowledge for diabetes diagnosis. It checks the symptom indication result of the patient with the diabetes types. Therefore, using rule its possible to represent knowledge for specific application domain.

INFERENCE ENGINE: is a search program that is responsible for deducing conclusion from knowledge base based on provided set of facts from user. Primarily, three techniques are utilized for inferring facts or deducing conclusion from knowledge base such as forward chaining, backward chaining and hybrid chaining. Forward chaining technique: it starts with knowledge base rules and facts and attempt to draws every possible inference from the data. It is also known as data-directed inference. This techniques can be suggested when a goal is difficult to determine[18][22][38][33][39].

Backward chaining technique: It starts with conclusion or goals and goes back to the fact that make the goals inferred or conclusion deduced to search for the facts that confirm the facts. It also provide an explanation and provide justification on how the conclusion are inferred. This technique is also known as goal-directed inference. Hybrid chaining: this techniques are used in a complex problem domain by merging forward chaining and backward chaining to provide efficient program. Inference engine must merged using both techniques to solve the complex nature of domain problem for an expert[18][22][38][33][39].

Explanation Facility: besides providing the final result domain expert could have come up to providing brief justification with knowledge based system on ‘how’ they reached to the final result. This is essential since justification is preferred in knowledge based system when providing an answer to the end-users. Knowledge based system is capable of providing justification of ‘why’ question on a given issue[18][22][38][33][39]. For instance, if an automated banking machine (ABM) move toward a decision to accept the card of end-user, then the ABM can show justification message like:

*“Sorry”, password still incorrect after three trials. We withheld your card, for your protection. Please contact your bank during regular office hours for help”*

According to [18][22][38][33][39] justification for an output generated is important since, the generated output could be dangerous. This problem is serious especially in medical domain where everything is related to life. In medical domain if the knowledge based system used by



an expert are not able to justify for an output generated an expert have no means of getting justification for diagnosed patient.

### **User Interface**

User interface act as a bridge between knowledge based system and end-users. Hence, in order to make interaction with knowledge based system an interface is required to retrieve and display information in an easy way. In one hand, if the result is displayed by an inference engine, an explanation can be provided by the kbs[21][18][22][38][33][39].

In other hand, if no result can be displayed by the inference engine in case there is no information the user interface should make a mechanism for obtaining an information from the user. Hence, bad implementation of user interface can led to make knowledge base system seen us unimportant in the look of end-user. End-user often used to judge a given knowledge based system based on fineness of its user interface rather than kbs itself[65][67][71][81][77].

**Knowledge Acquisition** The process of gathering and documenting knowledge is known as knowledge acquisition. An expert who is responsible for knowledge acquisition technique is called knowledge engineer. Knowledge can be collected from various source such as books, database, images, maps, stories, flow of diagrams and sensor[21][18][22][38][33][39].

There are two types of knowledge source: Tacit knowledge which is kept in subconscious mind of experts and explicit knowledge which is comparatively very easy to place in terms of books, tables and diagram. Knowledge can be extracted either human sense or machine such as scanners, camera, pattern matchers and intelligent agents. There are three types' of knowledge acquisition methods manual, automatic and semi-automatic[21][18][22][33][39].

**Manual Knowledge Acquisition Methods** The manual acquisition methods of knowledge encompass: interview (structured and unstructured), tracking methods and observation[21][18][22][38][33][39].

**Interviews:** is the most commonly used types of knowledge acquisition techniques. It include message exchange that are conducted between human expert and knowledge engineer. It includes a direct exchange of concepts between the human expert and the knowledge engineer. Information is gathered with the help of instruments such as tape recorders, questionnaires, and so on and is subsequently transliterated, analyzed, and coded[37][38][33][39].

**Tracking methods:** is a collection of methods that try to track the cognitive process of an expert. It is a widely held approach among reasoning psychologists who are interested in realizing the expert's train of thought in reaching a conclusion[[37][38].

The knowledge engineer can use the tracking process to discover what information is utilized and how it is utilized.

**Observation:** is about information gathering through watching how an expert practically work. This types of knowledge acquisition mechanism is direct and obvious. Most of the time an experts may work in different place at a given time. Knowledge engineer ought to keep track of every activity. This types of methods produce large volume of knowledge but little amount of knowledge can be useful[37][38].

### **Automatic Acquisition Methods**

This types of knowledge acquisition techniques use computer-based tools to obtain knowledge automatically from existing data. This types of knowledge acquisition techniques is better as compared with semi-automatic and manual acquisition techniques since, an expert is either busy, uncooperative or less knowledge in a given domain. Data mining and machine learning approaches are used to facilitate automatic knowledge acquisition techniques[37][38].

## Semi-automatic Acquisition Methods

Semi-automatic methods also use computer-based tools to extract knowledge automatically from particular data set.

**Domain Expert:** - is an individual who is proficient in his/her field of study. For example a medical doctor is a domain expert in the medical domain.

**Knowledge Engineer:** - is an individual who is responsible for gathering knowledge using different knowledge acquisition techniques. It is not must for a Knowledge engineer to be professional in a given domain area in which knowledge based system is being developed. For knowledge engineer the general knowledge about problem yet to be solved and important key term is enough[19][37][38].

## Knowledge Representation

It is a set of agreement that simulate the real world. Knowledge representation is a way of encoding knowledge of domain expert in a suitable form. This domain knowledge are acquired using vocabulary, data structure and program. The most common knowledge representation techniques are logic, frames, rules, semantic, and cases. But, we discuss two of the most commonly used techniques such as rule and case based representation[37][38].

## Rule-Based Representation

Rules are one of the most commonly used form of knowledge representation techniques that simplifies the represented knowledge. Rules can be represented in a terms of IF-THEN form. IF part is used to represent the condition that validate the truthfulness of facts. If the condition met then part will execute conclusion. The conditions in IF part connected through logical connector such as **AND, OR, NOT**[37][38] . If sufficient rule are satisfied then the conclusion will be fired.

## Case-Based Representation

It is a technique of using correlated problems to find the solution for the existing problems. It consists of two steps: find the related cases in memory that previously solved problems similar to the existing problem, and use the former solutions to fit the existing problem. The case archive makes an additional vital component in case-based representation.

The inference cycle in case based representation includes adapting solutions, retrieving solutions and testing solutions. The critical step is to find and fetch a relevant case from the case archive. Cases are placed using indexes. The stored case consists of a solution to be used by converting the parameters of the previous problem to match the new situation ensuing in an advocated solution.

If the analyzed solution is effective, then it will be a contribution to the case archive. The knowledge acquisition process in case-based representation needs little effort for the reason that of the graininess of the knowledge. Knowledge is provided in precedent or resultant cases.

According to Sajja and Akerkar[37][38],knowledge can be classified as: Tacit knowledge, explicit knowledge, factual knowledge, procedural knowledge, commonsense knowledge, domain knowledge, Meta knowledge. Domain knowledge: is a representative knowledge about a given domain. Expert develop their own domain knowledge and use it for solving a certain problem.

**Meta Knowledge:** is knowledge that provides information about another knowledge.

**Commonsense knowledge:** the general knowledge that has seen in every normal human being.

**Heuristic knowledge:** is a knowledge one can obtain from previous experience.

**Explicit Knowledge:** This type of knowledge is formal and systematic. It can be conveyed in words/numbers and shared in the form of data, formula, manuals and universal principles.

**Tacit Knowledge:** is knowledge that is very difficult to document; it is stored in sub conscious mind of human being. Since this types of knowledge is subjective it is very difficult to represent in knowledge based system. Personal visions, perceptions, emotions, mental models, values and action are some examples of tacit knowledge.

**Applications of Knowledge-Based System** According to [37][38] the knowledge based system is known by its use in various application. The most commonly used application area of knowledge based systems are discussed as follows.

**Advisory Systems:** knowledge based systems are preferable than in advising system than any other computer-based system. This is due to its goal-directed aspect of knowledge based system. It has also customizable to different place and the capability to reason on a given issue.

**Health care and medical diagnosis systems:** the ability to identify the state of sickness based on the symptoms provided is known as medical diagnosis. The explanation of medical reports like dermatology reports and expert decision can be validated easily with the help of knowledge based system[16][27][37].

**Searching in huge data base and data warehouses:** retrieving information that is important from huge database is not an easy task. Relevant information are scattered online and found in different web server. Hence, knowledge based system helps user to fetch an important information efficiently[16][27][37].

**Control and monitoring:** Monitoring is an uninterrupted interpretation of signals and making important decisions if interference is necessary. For example, such kind of monitoring can be used for an artificial life care system connected to a person who needs medical support after the surgical process[16][27][37].

**Prediction:** knowledge based system can help us to support forecasting about the future issues based on the developed model using current and previous knowledge. For example, it can able to predict the status of fluctuating market[16][27][37].

### 2.1.3 Overview of Data Mining

Data Mining is the process of obtaining and presenting previously unknown useful knowledge from huge amount of data. In this huge amount of data, the probability of getting many hidden pattern is high. This knowledge is then combined with extracted pattern and helps for decision making. Data mining can helps in detecting fraud and abuse, it can provides better medical treatments at affordable price. It also helps for detecting disease at its early stage and providing intelligence to support health care decision support system[18][16][27][37].

Data mining can be also used in hospital domain for disease prediction. The main objectives of data mining method is prediction and description. Prediction methods gives emphasis on predicting unknown values of target class based on values of existing variables in a given database. While, descriptive method involves searching for comprehensible pattern by human being.

The task of description includes clustering, association, summarization techniques. Since one of our aim is to predict the person likelihood of infecting by either cervical cancer disease or not, prediction of target class is mandatory. Moreover, we need to predict the suggested hospital

exactly according to fact based information of all available hospitals. Therefore, in this study classification technique are used, in order to make predictive model of both[18][16][27][37].

### Classification Techniques

Classification techniques is one of the most commonly utilized techniques in HealthCare Industry. It is learning task that can classify an instance of sample data-set into its target class. It can classify data points in to target class by separating sample data[17].

There are two categories of classification: binary and multilevel. In binary classification, we have two target class for example, high and low. Multilevel classification occurs when we have more than two class for example, 'high', 'medium' and 'low'. Data set can be classified as training and testing set. Training set are used to predict the target class of particular data point. Testing set are used to evaluate the trained set, the testing set is unknown in training set. In order to extract the knowledge from pattern classification can use rules for prediction. This Prediction rules can be placed in IF-THEN format. The following are an example of training set and testing set[18][34][5].

**Training Set**

Age	Heart rate	Blood pressure	Heart problem
45	75	140/64	Yes
28	85	101/60	No
38	62	105/55	No

**Prediction Set**

Age	Heart rate	Blood pressure	Heart problem
33	89	142/82	?
45	52	102/56	?
87	83	138/61	?

The most commonly used data mining techniques are discussed as below.

**K-Nearest Neighbor (K-NN)** K-NN classifier is simplest data mining techniques. It can predict the new instance based on previously known instance. And instance are classified using voting system. This techniques assume an existence of various objects. It can determine the characteristic of a given object and then, the nearest neighbor behavior can be determined based on previously determined characteristics. It is also known as memory based classifier because, training instance must kept in memory at run-time.

The main drawback of K-NN classifier is when larger values come under the smaller one while Euclidean distance is calculated in case of continuous attribute. Therefore, we need to normalize it in order to overcome problem that arise due to continuous attribute. K-NN can be applied for health data set, online marketing, clustering analysis and pattern recognition. It can provide various benefit such as simplicity, effectiveness and intuitiveness[18][33][45].

K-NN classifier is preferred when large training data set is available. One of the pitfall of K-NN classifier is large memory requirement. If the sample data set is large its response time will also increase in sequential computer. The pseudo code for K-NN classifier algorithm are summarized and discussed as below:

**Decision Tree (DT)** Decision Tree is one of the most widely used classifier techniques. In decision tree non-leaf node is used for testing, branches are used to represent the result of test and leaf-node is used to represent class label. Node that is found at the top-most is known as root node. Decision makers can use maximum information gain to select the best transversal path from root to leaf that destines unique class[19][17]. Its diverse in field of data mining, because of the following benefits:

- Self-explanatory
- Rules can be inferred easily from decision tree
- It can support any discrete classifiers, either in nominal and numeric attributes
- It can tolerate missing or erroneous dataset for a given sample dataset
- Domain knowledge is not required to build decision tree

Some limitation of decision tree are discussed as below:

- Most of the target class have discrete values, this is due to divide and conquer method used by the decision tree
- Decision tree is low in performance when there is multiple interaction between different attributes for a given dataset.
- over-sensitivity to the training set
- replication problem

Though, decision tree can be made using different algorithms, C4.5 is the well-known one and J48 is the extension of it. J48 algorithm denotes the decision tree in divide-and-conquer method. Here, each node has set of weighted cases considering the unknown attribute values. Decision tree construction doesn't require domain knowledge; hence, convenient for knowledge discovery. It can perform on high dimensional data and takes less learning time[17].

The algorithm can be summarized as follows.

*Create a node N;*

*If samples are all of the same class, C then*

*Return N as a leaf node labeled with the class C;*

*If attribute-list is empty then*

*Return N as a leaf node labeled with the most common class in samples;*

*Select test-attribute, the attribute among attribute-list with the highest information gain;*

*label node N with test-attribute;*

*for each known value v of test-attribute*

*grow a branch from node N for the condition test-attribute= v;*

*let sa be the set of samples for which test-attribute= v;*

*If sa is empty then*

*attach a leaf labeled with the most common class in samples;*

*else attach the node returned by*

*Generate decision tree(*si*,*attribute-list test-attribute*)*

**J48 algorithm** The J48 is an ideal way to present information from a machine algorithm and provide a quick and powerful way to explain structures in data. It is essential to know the various options in this algorithm, as they can bring a substantial difference in the accuracy of results. Mostly, the default settings are enough[17].

The J48 algorithm provides many options regarding tree pruning. Several algorithms try to "prune", or simplify, their outputs. Pruning results in fewer, but easily interpretable results. Pruning can also be used as a tool to balance a potential over fitting. This algorithm recursively classifies till each leaf is determined as perfect as possible. This process ensures high quality on the training data, but it may end in excessive rules[19][17].

When it is tested on a new data, the rules effectiveness may be less. Pruning usually cut down the accuracy of built model. This is due to the several means to relax the particularity of the decision tree, enhancing its performance on the test data. The general concept is to gradually summarize a decision tree until it obtains a balance of accuracy and flexibility[20].

### **Rule Induction**

Rule induction is finding a set of rules to predict or classify unknown class from the training set (known class). It is also a way of discovering important rules in the form of 'if-then' from the records on statistical basis. It has a form of

**IF condition THEN conclusion** According to [64][87], a classification rule has a conjunction between attributes. Among the several classification rule, PART is known for obtaining rules from partial decision tree [2].

**PART algorithm** According to[45][18] [17], for those learning techniques descriptions that are complex can be typically represented as sets of rules. Since they can be easily understood by human beings, these descriptions serve to describe what has been learned for other new predictions. Rules are a popular substitute to decision trees standing for the structures that are produced from the learning methods.

The precondition of a rule is just like the tests at nodes in decision trees and the consequence provides the class that is applied to instances using that rule. In general, the preconditions are joined together by the logical operator "AND". A set of rules is also easily readable directly off a decision tree. Each leaf represents one rule. The precondition of the rule includes a condition for each node through the path from the root to the leaf and the consequence of the rule is the class labeled by the leaf. PART algorithm is a class to create decision list in WEKA[18].

In all the iteration of PART, it builds a part of C4.5 decision tree and converts the best leaf node into a rule. The rules, which are generated by PART algorithm, are more clear and understandable by people. Subsequently, the J48 and PART algorithms were selected to develop a predictive model for cervical cancer prediction. This algorithms are also utilized for suggesting particular hospital exactly according to fact based information of all available hospitals[18].

### **Support Vector Machine (SVM)**

Although SVM is developed for binary classification, it is known by its strength of extensibility to multiple class domain. It is effective and efficient in multidimensional space, because it can built multiple hyper planes. The main objectives of building hyper plane in SVM is to create separation between different data point. Mainly, there are two techniques of implementing support vector machine (SVM)[18].

The first technique is by using mathematical programming. The second one is by using kernel function. This technique helps to convert non-linear function into multi-dimensional space by

using training data set. In order to classify data points hyper planes are used. It can help to increase the boundary of data points. Support vectors are preferred due to these hyper planes[45]. SVM has the following benefits:

- effectiveness in multidimensional space
- effectiveness when number of dimension exceeds its number of sample dataset
- memory efficient since, sub part of training set can be used only
- versatility, due to its kernel function

Some limitation of SVM are discussed as below:

- Poor performance occurs when number of features exceed the number of sample data set
- It will not provide probability estimate without five-fold cross validation

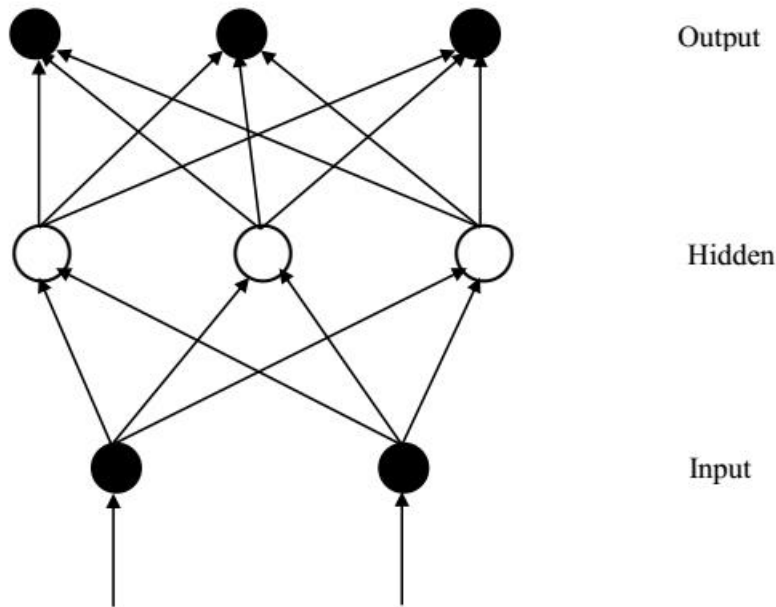
**Artificial Neural Network (ANN)** Artificial neural networks are highly parameterized models which can perform many of human tasks. For example, simultaneously analyzing large amounts of data (vision, hearing, touch) at the same time. It is inspired of complex human learning systems of highly interconnected neurons[17][18].

It is preferable to decision tree in that it can be used for robust and noisy data[61][92]. Nevertheless, its outputs are not easily interpretable. A neural network can generate output of one node providing into one or more nodes of the next layer. The Multi-layer Perceptron (MLP) a neural network using a back-propagation method which is during training, information spreads back via the network and modifies the connection weights. MLP training applies the back propagation learning method to come close to the optimal discriminant function used by Bayesian theory[45]. The Sigmoidal activation function is used to learn the supplied weight vectors in the training step as follows:

$$f(x) = \frac{1}{1+e^{(-x)}} \text{-----(2.1)}$$

Turning all of the learning rates, the number of hidden layers, the number of epochs, and the number of nodes in each hidden layer is a difficult task and all should be set suitably to arrive a good performance. In each epoch, all the input data are used with the current weights to identify the errors and then the back-propagated errors are calculated and the weights are modified. There is a bias presented for the hidden layers and the outputs. As depicted in Figure 2.2, there are input, hidden, and output layers. Each node in these layers has inputs and a single output. All the links in the network are accompanied by a weight[87][19][17].

As depicted in Figure 2.2, there are input, hidden, and output layers. Each node in these layers has inputs and a single output. All the links in the network are accompanied by a weight.



For instance, let us take the inputs  $x_1, \dots, x_n$  and the output  $a_i$  for the node  $N_i$  and then the weights  $w_{1i}, \dots, w_{ni}$  can be the input links of  $N_i$ . A node yields its output from the sum of its input weights, decreasing a threshold and passing the outcome to the activation function  $f$ . The outputs from the nodes in a layer are supplied as inputs to the next layer. Therefore, when the inputs  $(x_1, \dots, x_n)$  are fed to the input layer, the outputs  $(c_1, \dots, c_m)$  are obtained, where  $c_i$  has value 1 when the input belongs to class  $c_i$  and else 0.

**Bayesian Methods** Bayesian classifier is used for probabilistic learning methods. This method use bayes theorem of statistical field. This classifier assume that every attributes in a data set are do not depends to one another. Nave Bayes algorithm is also said called idiot's Bayes, simple Bayes, and independence Bayes[18][45].

Bayesian classifier has the following benefits:

- high performance are seen when attributes are independent to each other
- it is efficient and accurate for large number of sample dataset
- it can be easily calculated
- very easy to construct the rules
- Nave users can easily understand and justify it is classification.

Some limitation of Bayesian classifier are discussed as below:

- Poor performance occurs when attributes are dependent Bayesian classifier algorithm formula are also briefly discussed under previous reviews in opinion summarization. Bayesian classifier works by using the following assumption:

#### 2.1.4 Review of Related Works

In this section, previously conducted researches in areas related to the concept of opinion mining in hospital domain are presented and deeper discussion on various proposed cervical cancer prediction techniques are made. Also, specific works which is related to our proposed techniques and have relevant in terms of their objective were reviewed.



### Local Works:

- Local works under opinion mining are discussed under the table below

Author	Algorithm	Level of opinion mining	Language
Tulu	Rule-Based	Aspect-level	Amharic
Selama Gebre Meskel	Lexicon-Based	Document-level	Amharic

- Local works under data mining and knowledge based system are discussed under the table below.

Author	Area	Domain	Year
Mohammed A[17].	Integration	Network Intrusion detection	2013

- International Works:

Author	Area	Domain	Year
Narducci F.	Recommendation	Patient to Doctor connection	2015

### International Works:

**Opinion mining for hospital domain** In [46] an authors have advent a new approach to lexicon generation which is undertaken with bag of words(BOW) method in which BOW weak learner is consumed by stronger decision based learner and then lexicon based ensemble classification is carried out. This approaches were also applied to different domain such as Automotive, Book, Electronics, Health, Movies were used. The result that where obtained in this paper from both approach of Lexicon Based Ensemble and Supervised where the same.

**Fact based Information for hospital recommendation** Salunke A. and Kaser S. proposed medical assistance using keyword extraction. The similarity algorithm were used to make the similarity between person profiles and doctors profiles. Their system used top-k query to rank the doctors and his/her corresponding hospital. Using hybrid filtering, the result of 0.06 and 0.02 accuracy were achieved.

[47] presented a semantic recommender system able to suggest doctors with his/her corresponding hospital based on patient profile using similarity algorithm. This system was developed based on social network known as Health Net (HN). It can Rank hospital based on quality of doctors rather than various hospital quality factors such as material and medicine access, silence, health workers courtesy, sanitation and customer satisfaction.

This system is implemented only for patient who already know their disease (i.e. it ask patients to enter the disease) and searching for better doctors. It can recommend hospital based on patients profile only without including potentially useful variables/parameters that has to be considered during recommendation of a given health care services. Thus, this system judge the quality of a given hospital by only depending on doctor's quality and doctor's quality can also be judged based on patients profile.

They developed the system that connect patient with patient/doctors who are found online. But, recommendation doesn't works, if patient/doctors recommended are not found online due to busyness or other means. Thus, waiting time occurs by patient who are seeking for recommendation. The experience that were shared by online users do not tell you the death rate of a given hospital with particular disease, since mortality rate always remains the main factor that secure quality of hospitals.

However, this study is different from the works of others especially, Narducci[72] in:

- Its support of patients of two types, patient who already know their disease and new patient. For patient who already know their disease and suffering from getting quality services, this study will recommend them the health care in which they can get appropriate

services according to their disease. For new patient, this study helps patients by diagnosing their disease at early stage. Identifying the disease at its early stage helps patients to get diagnosis and treatment timely. Absence of getting treatment timely, can increase patients chance of getting the disease to be converted into chronic/recurrent or sudden death.

- Its support of Health care Administrators to use the system to distinguish their strong side and weak side of the service that has been provided. This allow Health care Administrators to work hard on their weak side and continue with the same spirit on their strong side. Since, this system can make summary based on users rather than other customer's data, it can helps to reduce an effort of analyzing summary. Therefore, this study plays great role in helping them to provide quality services by producing summary based on users predicted disease aspect rather than all summary of other customer opinion data in shortly without confusing users.
- Its support of health care services where there is insufficient domain expert, in balancing the distribution between health care workers (medical domain experts) and number of patients seeking the services. Additionally, this system helps health care where shortage of medical facility were seen.
- Its support of medical experts as long as an advice and consultation is required using hidden knowledge extracted from previous data. Additionally, this study helps medical experts to upgrade their knowledge and experience. Hence, this system consult medical expert by acting like partner in case some decision were required.
- Its approach that were utilized for the study to solve hospital recommendation, since both fact based and opinion based recommendation were carried out to solve local language problems particularly for Amharic and Afaan Oromo users.

Generally, several studies have been conducted on prediction of cervical cancer using data mining techniques. A few studies has been carried out in hospital recommendation and opinion mining for analyzing an opinion from hospital. As far as the researcher knowledge, there is no attempt that has been made:

- To make the domain tacit knowledge with treatment justification that use hidden knowledge of cervical cancer discovered by data mining techniques.
- To support hospital opinion mining with fact based information for hospital recommendation.
- To recommend hospital using opinion summarization and fact based information based on early predicted cervical using data mining with knowledge based system automatically.

## Chapter 3

# DATA PREPARATION AND PREPROCESSING

### 3.1 Overview of the real world data

Real world database are not pure in quality due to its noisiness, inconsistency and incompleteness nature of the data. Incompleteness can occurs due to various reason. The potentially useful data that are important for interpretation of cervical cancer diagnosis such as sexual contact starting age, after screening result may miss. Not only restricted to this, patient history that are used for hospital fact based information can also prone to the problem like Incompleteness. For example, most of after screening results shows as missing values due to Patients are absent after first treatment has been provide for them, they do not come again for the check up on the appointment day that was assigned by an expert due to various reasons.

This could make the result after screening empty/incomplete. Hence, based on this data recommendation of hospital is not possible since we do not know whether the patient get healthy or not after check-up. Although, the data are important, occasionally patients were not ready in responding to every query of the experts. The data are inconsistent while recording into excel from patient card this is due to poor checking mechanism for inserted data[33][39].

Others data that were not inserted due to misunderstanding for example, college is the value that is inserted for education status but it's not meaningful, we don't know whether college is college student ,college degree or college diploma. This type of problems led to poor quality of the data that are available for cervical cancer diagnosis. Lastly, the data that are seen as an inconsistent while interpreting can be removed to obtain high quality data.

Because of the reason that are discussed above data preparation and preprocessing is must to prevent garbage in garbage out (GIGO) problem. For the development of cervical cancer prediction model, the study use the seven step hybrid knowledge discovery model, we have selected these model since it is potentially enough in mingling both CRISP (Cross-Industry Standard Process for Data mining) and KDD (Knowledge Discovery in Database)[33][5][38].

### 3.2 Understanding the problem

The main activity in data mining is making the data available in well-formed manner, in such a way that it is suitable for the data mining tool or any other software that are going to be used. The major part of data mining task is data preparation. Data preparation consist of data selection, data construction, cleaning and integration.

The very first step in hybrid data mining process model is understanding the problem or domain area. In order to understand the problem domain we have reviewed the literature that has been done previously by the other researcher. The literature related to the area of cancer disease and hospital recommendation has been made using various source such as books, journals articles, web information and conference paper[5][22].

This has been carried out to understand different approach and technologies that were applied

in order to solve problem related to cancer patients. Domain understanding places its emphasis on three main issues: the first one is about working closely with the domain expert and defining the goal, the second one is about identifying the key people and learn the solution of the current problem. The third one is about learning the terminology that are specific to the domain area. Finally, it can be translated into data mining goals and an algorithm that can be applied on it to uncover useful knowledge that can be easily integrated with knowledge based system[5][22].

We have also identified and understand the topic that were not answered through review by using discussion, practical observation and interview techniques. Need assessments were also conducted to ensure the feasibility of the study. Useful discussion were also made with specialist in the cancer related issues. We have also made an interview and discussion with managers in Ethiopian Federal Ministry of Health(FMOE) for the hospital recommendation.

As we tried to understand from the talk of the managers of Federal Ministry of Health(FMOE)there is no benchmark system that are being used currently to recommend patient except the hospital quality checking mechanism that were based on annual reports coming from hospitals administrators at federal level. Moreover, from the experts that are engaged in the area of cancer diagnosis and treatment interview were made at different Ethiopian referral hospitals, we have understood that there is no any automated mechanism that were being utilized to help patients in hospitals in predicting or recommending hospitals other than recording raw data itself.

The patient history and the opinion of the patient that are potentially helpful in keeping quality of healthcare and recommendation of hospitals were abandoned without providing very useful service they can able to provide other than medical reports that were annually made at the federal level. After understanding the domain we have formulated basic research questions and objectives of the study.

### 3.3 Understanding the Data

Data is a Prerequisite for any data mining task. Data source is said to be better if the corporate data warehouse are kept in the same format for each attributes and their corresponding values. There is no cancer related data warehouse that are organized in most of Ethiopian referral hospitals except, the government had planned to open the cancer center for the first time in 2017, from the interview that was held between the researchers and manager of black lion hospital we understand that the cancer related data are planned to become centralized for the purpose of the research. Therefore, in upcoming few years there is a hope that data warehouse of various cancer types can be made.

These data warehouse will be centralized based on cancer registry of various types of cancer in different referral hospitals in Ethiopia. But, since there is no currently established cancer center in Ethiopia we have tried to collect data for cervical cancer from five recognized sites: Black Lion Hospital, Teklehaimanot General Hospital, Zewditu Memorial Hospital, St.Pauls Hospital and ALERT hospital.

In these hospitals, some of the hospitals kept the patient data in excel format but other data were registered and placed in hardcopy. The data that were registered in hardcopy format were copied into excel format for further preparation and getting the data in form that is suitable for data mining tool like WEKA 3.6 and 3.7.

After this phase the data that were copied from hardcopy and already in excel format are mingled to one consistent excel format. And finally these initial data set are defined and viewed in Excel 2013 format to observe the properties of the dataset in gross. As shown in Table 3.1 the initial data were collected from various Addis Ababa Hospitals: from black lion hospital the data with 2542 number of records with data set size of 128 KB were collected from period of 2010-2016 used.

From Teklehaimanot General Hospital the data with 2512 number of records with data set size of 124 KB were collected from period of 2010-2016. From St.Pauls Hospitals the data with

1200 number of records with data set size of 80 KB were collected from period of 2010-2016. From ALERT hospitals the data with 1300 number of records with data set size of 87 KB were collected from period of 2010-2016. From Zewditu Memorial hospitals the data with 2659 number of records with data set size of 133 KB were collected from period of 2010-2016. Totally, 10213 number of cervical cancer data set records with total of 552 KB were collected in a year 2010 to 2016 from five hospitals that are currently found in capital Addis Ababa. The total initial dataset size in excel format was 552 KB[34][33].

No	Station	Number of Records	Data Size	Period
1	Black Lion Hospital	2542	128 KB	2010-2016
2	Teklehaimanot General Hospital	2512	124 KB	2010-2016
3	St.Pauls Hospital	1200	80 KB	2010-2016
4	ALERT hospitals	1300	87 KB	2010-2016
5	Zewditu Memorial Hospitals	2659	133 KB	2010-2016
Total		10213	552 KB	

Table 3.1: Total Number of Initial registry of Cervical Cancer Dataset from Addis Ababa hospitals.

Similarly, for fact based information we have used the same station since cervical cancer diagnosis and treatments were provided in these aforementioned station. Therefore, for recommendation of hospital using fact based information we have used patient history dataset that are available in every hospitals that were providing these cervical cancer treatment service. Patient history comprises the history of patient before treatment and the result after treatment. Based on this information we can easily perceive that, if the number of patients in a given hospital shows all result after treatment positive, we can say that these hospital is not good at cervical cancer treatment.

These conclusion made based on the result of the patient since, no patient that took service in the hospital results negative. Finally, we can sum up that these hospital should not be recommended according to the result. But, the patient history data are not limited to this it can also include attributes such as Age, Number of Births, Marital Status, Education Status, history of sexually transmitted disease, cigarette smoking, early sexual intercourse started, screened on VIA, contraceptive medicine taken for a long time, counseled with VIA and CryoX, follow up-rescreening result.

These data was the same with dataset that were used for prediction of cervical cancer except a few attributes like follow up-rescreening result was also used for recommendation of hospitals these could also plays in recommending patients with the symptom at its early stages. However, for fact based information we have used the formula below in figure 3.1 as we can calculate by dividing the number of negative follow-up re screening results to the total number of patient that are served in a given hospitals.

$$FBR = \frac{NNRAS}{(NPRAS + NNRAS)}$$

FBR:-stands for fact based recommendation

NNRAS:-number of negative result after screening

NPRAS:- number of positive result after screening

### 3.3.1 Description of the Collected Data

As we have discussed previously the data collected were in two version in soft copy and hard copy format. Some of the hospital kept their data in excel format and others can use their data in hard copy format. Unfortunately all of the hospital were used the same attributes for cervical cancer patient records. While collecting these data some of the attributes were jargon to medical domain and clarification was made by an experts on that point.

	A	B	C	D	E	F	G	H	I	J	K	L
1	<b>age</b>	<b>mstatus</b>	<b>estatus</b>	<b>nob</b>	<b>hsti</b>	<b>ESS</b>	<b>CS</b>	<b>CM</b>	<b>VIAC</b>	<b>BAS</b>	<b>BP</b>	<b>Scresult</b>
2	adult	married	elementary	high	no	no	no	no	yes	no	no	postive
3	young	single	preparatory	low	no	no	no	no	yes	no	no	postive
4	adult	married	diploma	high	yes	no	no	no	no	no	no	postive
5	adult	married	elementary	high	yes	no	no	no	yes	no	no	postive
6	young	married	diploma	low	no	no	no	no	yes	no	no	postive
7	adult	single	ue	low	no	no	no	no	yes	no	no	postive
8	young	married	preparatory	high	no	no	no	no	yes	no	no	postive
9	young	married	preparatory	low	no	no	no	no	yes	no	no	postive
10	young	married	elementary	high	no	no	no	no	yes	no	no	postive
11	young	single	ue	low	no	no	no	no	yes	no	no	postive
12	young	married	highschool	low	no	no	no	no	yes	no	no	postive
13	young	married	highschool	low	yes	no	no	no	yes	no	no	postive
14	young	married	diploma	low	yes	no	no	no	yes	no	no	postive
15	young	married	elementary	low	no	no	no	no	yes	no	no	postive
16	young	single	elementary	low	yes	no	no	no	yes	no	no	postive
17	young	divorced	diploma	low	no	no	no	no	yes	no	no	postive
18	young	married	elementary	low	yes	no	no	no	yes	no	no	postive
19	young	married	preparatory	low	yes	no	no	no	yes	no	no	postive

Figure 3.1: Initial Dataset from Addis Ababa hospitals that provides Cervical Cancer Treatment

For fact based recommendation follow-up data was used from three types of data that were available such as normal, follow-up and died. Because, it is very difficult to get normal and died patient since they do not come to hospital for re-screening.

### 3.4 Data Preprocessing

The main part of data preprocessing is data cleansing, data reduction and data transformation. These task is important to make data available in a way that is appropriate for analysis. The primary goal of data mining is to uncover hidden knowledge that were not yet been known for years. Therefore, quality data should be there to obtain these knowledge after analyzing it. The fundamental requirement of data preprocessing is to minimize or even to eliminate GIGO (Garbage in Garbage out) if possible. Making the data prepared for analysis is the most time consuming and critical task in data mining process. Its criticality occurs due to the problems that occurs due to GIGO.To solve this problem Microsoft Excel and Weka plays an important role.

#### Data Cleansing

The data that are available actually in real-world were not always be complete, consistent and free from being noisy. The analysis that can be made based on the data that is not complete, inconsistent and noisy led to inaccurate output that causes incorrect decision by users. Thus, to solve this problem data cleansing is must. The main aim of data cleansing phase is to fill missing values, smoothing out noise, correcting inconsistency and identifying outliers that are available in data set.

One of the most commonly used technique of handling missing values is ignoring the records, especially if the values like predicted class is not available. Not only this but for example, if potentially important values were not available the tuples with missing values will be deleted for the sake of minimizing inaccuracy. Missed values can also filled using automatic technique such as expected maximization techniques (EM).Expected Maximization (EM) works by searching the missing values throughout the dataset and obtaining the optimal value to fill the missed place.

In this study both manual and automatic techniques of handling missing values were applied.

<b>NO</b>	<b>Attribute</b>	<b>Description</b>	<b>Datatype</b>	<b>Values</b>
1	Sno	Patient Record Number	Numeric	1,2,3...N
2	AGE	Patient age	Nominal	Young or Adult
3	MSTATUS	Marital status of the patient	Nominal	Maried, Single,Widow and Divorced
4	ESTATUS	Education status of the patient	Nominal	Degree,Diploma,Masters, ,Certificate,Preparatory ,HighSchool,Elementary, ,Uneducated
5	NOB	Number of birth by the patient	Nominal	High(>3) or low(<3)
6	HSTI	History of sexually Transmitted Disease	Nominal	Yes or No
7	ESS	Early Sexual Intercourse Started	Nominal	Yes or No
8	CS	Cigarate Smoking	Nominal	Yes or No
9	VIAC	VIA counselling and CryoTx	Nominal	Yes or No
10	CM	Contraceptive Medicine	Nominal	Yes or No
11	BAS	Blood After Sexuall intercourse	Nominal	Yes or No
12	BP	Back Pain	Nominal	Yes or No
13	SCRESULT	Screening Result	Nominal	Postive or Negative

Figure 3.2: Attribute Name, Description, Data Type and Values

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	age	mstatus	estatus	nob	hsti	ESS	CS	CM	VIAC	BAS	BP	Scresult	Ras	hospital
2	adult	married	elementary	high	no	no	no	no	yes	no	no	postive	negative	BLH
3	young	single	preparatory	low	no	no	no	no	yes	no	no	postive	negative	BLH
4	adult	married	diploma	high	yes	no	no	no	no	no	no	postive	negative	BLH
5	adult	married	elementary	high	yes	no	no	no	yes	no	no	postive	negative	BLH
6	young	married	diploma	low	no	no	no	no	yes	no	no	postive	negative	BLH
7	adult	single	ue	low	no	no	no	no	yes	no	no	postive	negative	BLH
8	young	married	preparatory	high	no	no	no	no	yes	no	no	postive	negative	BLH
9	young	married	preparatory	low	no	no	no	no	yes	no	no	postive	negative	BLH
10	young	married	elementary	high	no	no	no	no	yes	no	no	postive	negative	BLH
11	young	single	ue	low	no	no	no	no	yes	no	no	postive	negative	BLH
12	young	married	highschool	low	no	no	no	no	yes	no	no	postive	negative	BLH
13	young	married	highschool	low	yes	no	no	no	yes	no	no	postive	negative	BLH
14	young	married	diploma	low	yes	no	no	no	yes	no	no	postive	negative	BLH
15	young	married	elementary	low	no	no	no	no	yes	no	no	postive	negative	BLH
16	young	single	elementary	low	yes	no	no	no	yes	no	no	postive	negative	BLH
17	young	divorced	diploma	low	no	no	no	no	yes	no	no	postive	negative	BLH
18	young	married	elementary	low	yes	no	no	no	yes	no	no	postive	negative	BLH
19	young	married	preparatory	low	yes	no	no	no	yes	no	no	postive	negative	BLH

Figure 3.3: Initial Dataset from Addis Ababa hospitals that provides Patient History of follow up only

For attribute that were missed in cervical dataset such as marital status, number of births and history of sexually transmitted disease were filled by using automatic expected maximization. These were applied for missed values of 67 of marital status, 17 for number of births and 21 for history of sexually transmitted disease.

While for attributes that missed their predicted class manual techniques of removing the missing values were made. These deletion of predicted class were applied on 73 screening result. These process of ignoring the missing tuples were done before expected maximization takes place. Thus, expected maximization works after tuples with missed values of merely predicted class such as screening result were ignored.

For patient follow up history 37 missed values of result after screening attribute from various hospitals were ignored using automatic expected maximization technique of handling missing values. While 17 missed values of patient hospitals attribute were also ignored using manual techniques of handling missing values.

As discussed in Table 3.3 the total number of 10127 records were obtained after manual and automatic missing values handling mechanism were applied on cervical cancer dataset. After these techniques were applied the following dataset that is shown as below in Figure 3.3 Were obtained.

As discussed in Table 3.4 the total number of 10151 records were obtained after manual and automatic missing values handling mechanism were applied on patient history dataset. After these techniques were applied the following dataset that is shown as below in Figure 3.3 Were obtained.

Figure 3.3 Data set of patient follow-up history after cleansing

### Data Reduction

The main aim of data reduction is increasing efficiency of the model by reducing the data that do not affect the effectiveness of the system. Data that remains inside the dataset but add no values can minimize the efficiency of the model since it occupies particular space. Data reduction is the process of ignoring the data that can't make changes on the predicted results. Therefore, the attributes that are discussed in the Table 3.5 were removed by simply selecting and deleting using Microsoft Excel. The attributes that were removed and their reason were discussed as below.



	A	B	C	D	E	F	G	H	I	J	K	L
1	<b>age</b>	<b>mstatus</b>	<b>estatus</b>	<b>nob</b>	<b>hsti</b>	<b>ESS</b>	<b>CS</b>	<b>CM</b>	<b>VIAC</b>	<b>BAS</b>	<b>BP</b>	<b>Scresult</b>
2	adult	married	elementary	high	no	no	no	no	yes	no	no	postive
3	young	single	preparatory	low	no	no	no	no	yes	no	no	postive
4	adult	married	diploma	high	yes	no	no	no	no	no	no	postive
5	adult	married	elementary	high		no	no	no		no	no	postive
6	young	married	diploma	low	no	no	no	no	yes	no	no	postive
7	adult	single	ue	low	no	no	no	no	yes	no	no	postive
8	young	married	preparatory	high	no	no	no	no	yes	no	no	postive
9	young	married	preparatory	low	no	no	no	no		no	no	postive
10	young	married	elementary	high	no	no	no	no	yes	no	no	postive
11	young	single	ue	low	no	no	no	no	yes	no	no	postive
12	young		highschool	low	no	no	no	no	yes	no	no	postive
13	young	married	highschool	low	yes	no	no	no	yes	no	no	
14	young	married	diploma	low	yes	no	no	no	yes	no	no	postive
15	young	married	elementary	low	no	no	no	no	yes	no	no	postive
16	young	single	elementary	low	yes	no	no	no	yes	no	no	postive
17	young	divorced	diploma	low	no	no	no	no	yes	no	no	postive
18	young	married	elementary	low	yes	no	no	no	yes	no	no	postive
19	young	married	preparatory	low	yes	no	no	no	yes	no	no	postive

Figure 3.4: Attribute with missed values in cervical dataset

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	<b>age</b>	<b>mstatus</b>	<b>estatus</b>	<b>nob</b>	<b>hsti</b>	<b>ESS</b>	<b>CS</b>	<b>CM</b>	<b>VIAC</b>	<b>BAS</b>	<b>BP</b>	<b>Scresult</b>	<b>Ras</b>	<b>hospital</b>
2	adult	married	elementary	high	no	no	no	no	yes	no	no	postive	negative	BLH
3	young	single	preparatory	low	no	no	no	no	yes	no	no	postive	negative	BLH
4	adult	married	diploma	high	yes	no	no	no	no	no	no	postive	negative	BLH
5	adult	married	elementary	high		no	no	no		no	no	postive	negative	BLH
6	young	married	diploma	low	no	no	no	no	yes	no	no	postive	negative	BLH
7	adult	single	ue	low	no	no	no	no	yes	no	no	postive	negative	BLH
8	young	married	preparatory	high	no	no	no	no	yes	no	no	postive	negative	BLH
9	young	married	preparatory	low	no	no	no	no		no	no	postive	negative	BLH
10	young	married	elementary	high	no	no	no	no	yes	no	no	postive	negative	BLH
11	young	single	ue	low	no	no	no	no	yes	no	no	postive	negative	BLH
12	young		highschool	low	no	no	no	no	yes	no	no	postive	negative	BLH
13	young	married	highschool	low	yes	no	no	no	yes	no	no		negative	BLH
14	young	married	diploma	low	yes	no	no	no	yes	no	no	postive	negative	BLH
15	young	married	elementary	low	no	no	no	no	yes	no	no	postive	negative	BLH
16	young	single	elementary	low	yes	no	no	no	yes	no	no	postive	negative	BLH
17	young	divorced	diploma	low	no	no	no	no	yes	no	no	postive	negative	BLH
18	young	married	elementary	low	yes	no	no	no	yes	no	no	postive	negative	BLH
19	young	married	preparatory	low	yes	no	no	no	yes	no	no	postive	negative	BLH

Figure 3.5: Attribute with missed values of Patient follow-up History

No	Station	Number of Records	Number of Missed Values	Number of Records After Handling Missing Values
1	Black Lion Hospital	2542	78	2513
2	Teklehaimanot General Hospital	2512	27	2505
3	St.Pauls Hospital	1200	39	1190
4	ALERT hospitals	1300	22	1275
5	Zewditu Memorial Hospitals	2659	12	2644
Total		10213	178	10127

Figure 3.6: Total Number of registry of Cervical Cancer Dataset after removing missing Values

	A	B	C	D	E	F	G	H	I	J	K	L
1	age	mstatus	estatus	nob	hsti	ESS	CS	CM	VIAC	BAS	BP	Scresult
2	adult	married	elementary	high	no	no	no	no	yes	no	no	postive
3	young	single	preparatory	low	no	no	no	no	yes	no	no	postive
4	adult	married	diploma	high	yes	no	no	no	no	no	no	postive
5	adult	married	elementary	high	yes	no	no	no	yes	no	no	postive
6	young	married	diploma	low	no	no	no	no	yes	no	no	postive
7	adult	single	ue	low	no	no	no	no	yes	no	no	postive
8	young	married	preparatory	high	no	no	no	no	yes	no	no	postive
9	young	married	preparatory	low	no	no	no	no	yes	no	no	postive
10	young	married	elementary	high	no	no	no	no	yes	no	no	postive
11	young	single	ue	low	no	no	no	no	yes	no	no	postive
12	young	married	highschool	low	no	no	no	no	yes	no	no	postive
13	young	married	highschool	low	yes	no	no	no	yes	no	no	postive
14	young	married	diploma	low	yes	no	no	no	yes	no	no	postive
15	young	married	elementary	low	no	no	no	no	yes	no	no	postive
16	young	single	elementary	low	yes	no	no	no	yes	no	no	postive
17	young	divorced	diploma	low	no	no	no	no	yes	no	no	postive
18	young	married	elementary	low	yes	no	no	no	yes	no	no	postive
19	young	married	preparatory	low	yes	no	no	no	yes	no	no	postive

Figure 3.7: Data set of cervical cancer after cleansing

No	Station	Number of Records	Number of Missed Values	Number of Records After Handling Missing Values
1	Black Lion Hospital	4241	21	4230
2	Teklehaimanot General Hospital	2595	11	2585
3	St.Pauls Hospital	1520	6	1517
4	ALERT hospitals	701	2	701
5	Zewditu Memorial Hospitals	1128	14	1118
Total	-----	10185	54	10151

Figure 3.8: Total Number of registry of Patient History follow up Dataset after removing missing Values

Removed Attribute	Reason to be removed
Registration Date	The registration date do not affect the result of prediction
Registration Number	The number that is given for patient while registration, therefore since it doesn't affect the result its removed
Date Appointed	Patient appointment date ; to check patient progress Since, patient appointment date is used to check the progress of the patient it is not important for prediction
Treatment	Treatment option such as Cryotherapy, LEEP and others do not' helps for prediction as well as recommendation
Result After VIA screening	This is removed since, it is laboratory experiment not used for early diagnosis of patient

Figure 3.9: Attributes Removed from both cervical cancer and patient history dataset

### Data Transformation

Data transformation is the process of converting possible attributes values into the form that is convenient for data mining. For this study, all the data in the dataset were consolidated except the education status that were discretized by ministry of education plan and patient age that was considered by care for cervical cancer.

In addition to this transformation were also made for predicted class for the original class: noacitowhite lesion, acetowhite lesion eligible for cryo , acetowhite lesion not eligible for cryo and suspicious to cancer. These four class were converted into three class Normal, pre-cancer and cancer. This, study mainly focus on the pre-cancer symptom for prediction of cervical cancer.

Original Education Status	New converted Education Status
0 or – or illiterate	Represents UE-uneducated
1-8	Elementary
9-10	High School
11-12	Preparatory
10+1,10+2,10+3	certificate
College	Diploma
University	Degree
MA, MSc	Masters
Doctorate	PHD

Figure 3.10: Transformation for Education Status Attribute

Original Age	New Age
>40	Adult
<40	Young

Figure 3.11: Transformation for Age Attribute

### 3.4.1 WEKA Format

After preprocessing step has completed or after getting quality data the subsequent step is arranging the data in a way that is understandable and explored by Weka tool. This is possible by converting the quality data that were available in spreadsheet format to ARFF file format. Most commonly the process of converting excel (.xls) format can be accomplished by converting data in this format to comma delimited (.csv) file format and then these could be converted into ARFF file format.

The converted format were represented in three granularity: Relation, Attributes and Data. The Relation is used to represent the name of the project, Attributes is used to handle an attributes that were available in dataset while data were used to keep the values that are available in dataset.

These values could have different nature Nominal and Numeric are two of them. Nominal values are represented in curly brace. Dataset in ARFF file format do not gives any notion about the predicted class in fact it can provides the dataset itself merely. Therefore, this helps to know how much the other class helps to be predicted by the predicted class itself ; equal opportunity are provided for all class to be used for prediction. The sample ARFF file format for both cervical cancer and patient history were showed as below in listing 3.1 and 3.2 respectively.

## 3.5 Selecting Best classifiers

After the data were produced in WEKA understandable format, the next step is feeding data to it. In the intention of selecting best algorithm to build model that could produce better accuracy (i.e. results more hidden knowledge) for integrating the predictive model with knowledge based system, an experiment were conducted upon most commonly used data mining techniques for classification or class prediction. These algorithm were decision tree induction and rule-based induction.

The models were built with four different supervised machine learning algorithm. Two algorithm that were commonly utilized under Decision Tree Classification Algorithm and Two rule based classifier algorithm. In order to obtain better result four classifier algorithms were applied and comparison has made using different evaluation parameters. But, before performing an experiment to select best classifiers, two important operation would be carried out on the

```

@RELATION CervicalCancer
@ATTRIBUTE attribute_0 {adult,age,young}
@ATTRIBUTE attribute_1 {divorced,married,mstatus,single,unmarried,widow}
@ATTRIBUTE attribute_2 {degre,degree,diploma,elementary,estatus,highschool,preparatory,ue}
@ATTRIBUTE attribute_3 {high,low,nob}
@ATTRIBUTE attribute_4 {hsti,no,yes}
@ATTRIBUTE attribute_5 {ESS,no,yes}
@ATTRIBUTE attribute_6 {CS,no,yes}
@ATTRIBUTE attribute_7 {CM,no,yes}
@ATTRIBUTE attribute_8 {VIAC,no,yes}
@ATTRIBUTE attribute_9 {BAS,no,yes}
@ATTRIBUTE attribute_10 {BP,no,yes}
@ATTRIBUTE attribute_11 {Scresult,postive,negative}
@DATA
age,mstatus,estatus,nob,hsti,ESS,CS,CM,VIAC,BAS,BP,Scresult
adult,married,elementary,high,no,no,no,no,yes,no,no,postive
young,single,preparatory,low,no,no,no,no,yes,no,no,postive
adult,married,diploma,high,yes,no,no,no,no,no,no,postive
adult,married,elementary,high,yes,no,no,no,yes,no,no,postive
young,married,diploma,low,no,no,no,no,yes,no,no,postive
adult,single,ue,low,no,no,no,no,yes,no,no,postive
young,married,preparatory,high,no,no,no,no,yes,no,no,postive
young,married,preparatory,low,no,no,no,no,yes,no,no,postive
young,married,elementary,high,no,no,no,no,yes,no,no,postive
young,single,ue,low,no,no,no,no,yes,no,no,postive
young,married,highschool,low,no,no,no,no,yes,no,no,postive
young,married,highschool,low,yes,no,no,no,yes,no,no,postive
young,married,diploma,low,yes,no,no,no,yes,no,no,postive
young,married,elementary,low,no,no,no,no,yes,no,no,postive
young,single,elementary,low,yes,no,no,no,yes,no,no,postive

```

Figure 3.12: Sample of Arff file for Cervical Cancer Data

```

@RELATION FBHR
@ATTRIBUTE attribute_0 {adult,age,young}
@ATTRIBUTE attribute_1 {divorced,married,mstatus,single,unmarried,widow}
@ATTRIBUTE attribute_2 {degre,degree,diploma,elementary,estatus,highschool,preparatory,ue}
@ATTRIBUTE attribute_3 {high,low,nob}
@ATTRIBUTE attribute_4 {hsti,no,yes}
@ATTRIBUTE attribute_5 {ESS,no,yes}
@ATTRIBUTE attribute_6 {CS,no,yes}
@ATTRIBUTE attribute_7 {CM,no,yes}
@ATTRIBUTE attribute_8 {VIAC,no,yes}
@ATTRIBUTE attribute_9 {BAS,no,yes}
@ATTRIBUTE attribute_10 {BP,no,yes}
@ATTRIBUTE attribute_11 {Scresult,postive,negative}
@DATA
age,mstatus,estatus,nob,hsti,ESS,CS,CM,VIAC,BAS,BP,Scresult,Ras,hospital
adult,married,elementary,high,no,no,no,no,yes,no,no,postive,negative,BLH
young,single,preparatory,low,no,no,no,no,yes,no,no,postive,negative,BLH
adult,married,diploma,high,yes,no,no,no,no,no,postive,negative,BLH
adult,married,elementary,high,yes,no,no,no,yes,no,no,postive,negative,BLH
young,married,diploma,low,no,no,no,no,yes,no,no,postive,negative,BLH
adult,single,ue,low,no,no,no,no,yes,no,no,postive,negative,BLH
young,married,preparatory,high,no,no,no,no,yes,no,no,postive,negative,BLH
young,married,preparatory,low,no,no,no,no,yes,no,no,postive,negative,BLH
young,married,elementary,high,no,no,no,no,yes,no,no,postive,negative,BLH
young,single,ue,low,no,no,no,no,yes,no,no,postive,negative,BLH
young,married,highschool,low,no,no,no,no,yes,no,no,postive,negative,BLH
young,married,highschool,low,yes,no,no,no,yes,no,no,postive,negative,BLH
young,married,diploma,low,yes,no,no,no,yes,no,no,postive,negative,BLH
young,married,elementary,low,no,no,no,no,yes,no,no,postive,negative,BLH
young,single,elementary,low,yes,no,no,no,yes,no,no,postive,negative,BLH

```

Figure 3.13: Sample of Arff file for Patient History of Hospital

Classifiers Algorithm	Description	Data	Attributes	Experiment
Un-pruned J48		Actual	All	I
			Selected	II
		Balanced	All	III
			Selected	IV
JRIP		Actual	All	V
			Selected	VI
		Balanced	All	VII
			Selected	VIII
PART		Actual	All	IX
			Selected	X
		Balanced	All	XI
			Selected	XII
REP Tree		Actual	All	XIII
			Selected	XIV
		Balanced	All	XV
			Selected	XVI

Figure 3.14: Comparison between the Actual Data and Balanced Data

dataset: balancing dataset and attribute selection.

### 3.5.1 Balancing Dataset

According to [68], if the predicted target class were very less in number than other class, it is recommended to balance them in order to reduce the over fitting of the confusion matrix. Therefore, to balance huge variation between predicted classes, SMOTE (Synthetic Minority Over-sampling Technique) was applied to overcome the problem due to skew of the most available classes in the dataset. As it is shown in Figure 4.1, from the total of 10127 of original cervical cancer dataset the actual data contains 9293 “Negative” and 920 “positive” classes. However, after SMOTE was applied on it, the Positive class was obtained as 5001 instances.

### 3.5.2 Attribute Selection

According to [2] Attribute selection is one of the trivial practice followed in WEKA during evaluation. Even though, the records can be evaluated based on individual attribute value, all attribute are not relevant for the goal of the study, hence it has to be removed. This was done by using, the default parameter (i.e. “CfsSubsetEval” evaluator and “BestFirst” search) for attribute selection in WEKA was used. Out of the 16 attributes of the original data set, 11 attributes (including the class attribute) were selected. Those attributes age mstatus ,estatus,nob,hsti ,ESS ,CS ,CM ,VIAC,BAS,sresult.

### 3.5.3 Experimental Setup

In this study, sixteen experiments were conducted by applying decision tree and rule-based induction algorithm. Each experiment was conducted using the data before and after balancing, and for all attribute and for selected attributes. These experiment were shown as below in Table

Classifiers Algorithm	Description	Data	Attributes	Experiment
Un-pruned J48		Actual	All	I
			Selected	II
		Balanced	All	III
			Selected	IV
JRIP		Actual	All	V
			Selected	VI
		Balanced	All	VII
			Selected	VIII
PART		Actual	All	IX
			Selected	X
		Balanced	All	XI
			Selected	XII
REP Tree		Actual	All	XIII
			Selected	XIV
		Balanced	All	XV
			Selected	XVI

Figure 3.15: Sixteen Experiment that was conducted on actual,pruned,unpruned and balanced data

### 3.6 Experiments and Results Using J48

As depicted in Table 4.1, the experiment was conducted upon both selected attributes and all Attributes, using the actual data and the balanced data, and pruned and unpruned tree parameters as well. The detail analysis of all the J48 experiments is presented in Table 4.3 below, but here only the first three experiments with the higher accuracy and one experiment with the lowest accuracy are discussed.

#### 3.6.1 The Experiment with the Highest Accuracy

From all experiment carried out, J48 pruned algorithm upon all attributes of the actual data is run. This experiment scores (98.93%) of correctly classified instances and (1.066%) % of incorrectly classified instances. The true positive (TP) Rate in this experiment depicts that this scenario succeeds by 0.995 average weights. The average precision and recall of the model are 0.993 and 0.995 respectively. At the same time, the F-Measure value is 0.994 which is significantly balanced with the precision. As it is seen from the two tables, the first experiment that is J48 pruned algorithm upon all attributes of the actual data has a better accuracy than the other experiments. Hence, this model is selected to be integrated to the Knowledge Based System.

#### 3.6.2 Confusion Matrix of the Selected Model

The confusion matrix can show the selected model prediction performance by comparing it to the Actual value. Table 4.2 depicts the confusion matrix of the selected model.

As shown in the table, the selected model J48 pruned algorithm upon all attributes of the ac-



ACTUAL	Predicted	
	Negative	3122
Positive	21	313

Figure 3.16: Confusion Matrix of J48 pruned algorithm upon all attributes of the actual data

```

Time taken to build model: 0.22 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances      3435          98.9343 %
Incorrectly Classified Instances    37            1.0657 %
Kappa statistic                    0.9383
Mean absolute error                 0.0144
Root mean squared error             0.0888
Relative absolute error              8.6391 %
Root relative squared error         30.0876 %
Total Number of Instances          3472

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
          0.995   0.063   0.993     0.995   0.994     0.985   negative
          0.937   0.005   0.951     0.937   0.944     0.985   positive
Weighted Avg.  0.989   0.057   0.989     0.989   0.989     0.985

=== Confusion Matrix ===

  a  b  <-- classified as
3122 16 |  a = negative
 21 313 |  b = positive

```

Figure 3.17: Detail Analysis Result of J48 Algorithm

Classifier	Class		
		Positive	Negative
<b>PART</b>	Precision	94.8%	95.7%
	Recall	94.5%	95.6%
	F-measure	94.6%	95.6%
<b>JRIP</b>	Precision	96.6%	97.6%
	Recall	96.7%	97.5%
	F-measure	96.6%	97.5%
<b>J48</b>	<b>Precision</b>	<b>98.9%</b>	<b>98.9%</b>
	<b>Recall</b>	<b>98.9%</b>	<b>98.9%</b>
	<b>F-measure</b>	<b>98.9%</b>	<b>98.9%</b>
<b>J48graft</b>	Precision	98.01%	98.01%
	Recall	98.03%	98.03%
	F-measure	98.02%	98.02%
<b>Random Forest</b>	Precision	95.8%	95.8%
	Recall	95.6%	95.6%
	F-measure	95.7%	95.7%
<b>Random Tree</b>	Precision	96.2%	96.2%
	Recall	96.5%	96.5%
	F-measure	96.3%	96.3%

Table 3.2: Selected Experiment with the Highest Accuracy

tual data generally classified 3435 (98.93%) instances correctly as well as 37 (1.066%) instances Incorrectly. Moreover, the model classified 313 of Positive patients correctly but classified 16 Positive Patient incorrectly mean that it puts 16 positive patients as they are negative inversely. Meanwhile, the model also identified 3122 of negative patients correctly but unfortunately confused 16 negative as positive. The main reason for this algorithm to classify the instances 1.066%) % incorrectly is the convergence of attribute values of screening result noacitowhite lesion, acetowhite lesion eligible for cryo , acetowhite lesion not eligible for cryo and suspicious to cancer.

The rule can be extracted easily passing from the root node (in this case ESS) through other nodes to any leaves (in the figure positive and Negative). The first numbers in the leaves indicates the occurrence of that particular path (rule).The numbers after a slash reveal the wrongly classified instances and therefore it is possible to dictate the probability of occurrence of the rule [62]. Some of the interesting rules generated from the tree are:

Depending on the above generated rule there are of course many interesting rule from that

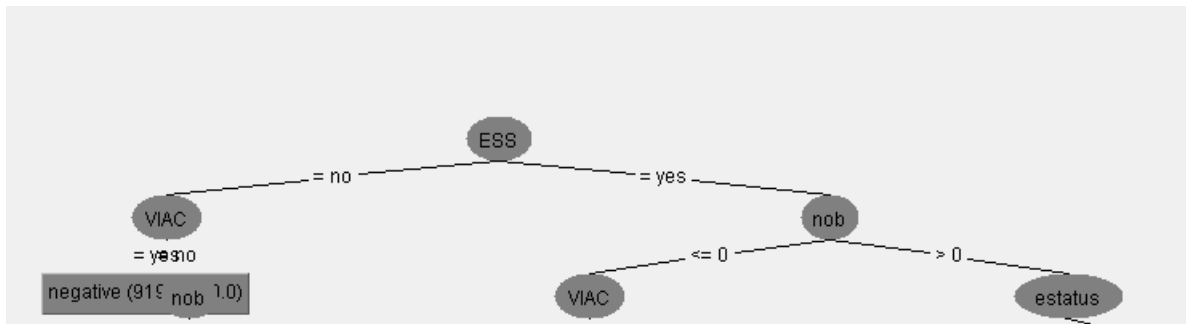


Figure 3.18: Partial decision tree for the selected J48 algorithm

- Rule1: IF ESS=no and VIAC=yes: negative
- Rule2: IF ESS=no and nob<=0: negative
- Rule3: IF Ess=no and nob>0 and hsti=no: positive
- Rule4: IF ESS=no and VIAC=no and nob>0 and hsti=yes and nob<=1: positive
- Rule5: IF ESS=no and VIAC=no and nob>0 and hsti=yes and nob>=1: positive
- Rule6: IF ESS=yes and VIAC=no and nob<=0 and mstatus=married and estatus=elementary: positive
- Rule7: IF ESS=yes and nob>0 and mstatus=divorced and estatus=preparatory and hsti=yes: positive
- Rule8: IF ESS=yes and nob>0 and mstatus=divorced hsti=no and nob>=1: negative
- Rule9: IF ESS=yes and nob>0 and mstatus=widow and age=adult: positive
- Rule10: IF ESS=yes and nob>0 and mstatus=widow and age=young and viac=no and hsti=yes and nob<=2: negative

Figure 3.19: Interesting Rule generated by weka rule for cervical cancer disease

one can be described as: if anyone(female) made sexual intercourse at her early age(age $\leq$ 18) and divorced and history of sexually transmitted disease is yes and number of birth greater than zero and education status is preparatory then positive. But if history of sexually transmitted disease is no then patient will be negative for pre-cancer stage. This is due to other sexually transmitted disease can be associated with HPV (Human Papilloma Virus) will occur but with other factor as it is.

## Chapter 4

# INTEGRATING OPINION SUMMARY WITH FACT BASED INFORMATION

### 4.1 Opinion Summary

In this chapter the brief explanation on the detail component of opinion mining and summarization system model for hospital review are discussed. The model that have been proposed for OMASS (opinion mining and summarization system) is given below with some very crucial steps and it is components. The first step of opinion summarization is file uploading or providing the reviews document collection. The second step is document pre-processing step. This step consists of activities such as sentence splitting, stop word removing, tokenizing and stemming. The review document generated from previous can be utilized for sentence splitting activity.

This activity can accept review document and split into individual sentence review based on punctuation. After splitting activity has been carried out the generated sentence review is utilized for tokenization activity. This activity is responsible for chunking each individual review sentence in to its small parts of sentence called token. Then stop word removal activity has followed by removing stop words such as punctuation, numbers etc. Finally, stemming activity can be undertaken. Then after preprocessing has carried out, then opinion summarization followed.

The first step in opinion summarization is aspect detection. This step is responsible for detecting feature and sub feature from each individual segment of sentence in a given document. These are carried out with the help of English parser and some rule based algorithm. This rule based algorithm is designed based on English common futures tags and its underlying relationship. The second step is sentiment orientation detection step. In this step n-gram is used for detecting the co-occurrence of opinionated sentence words together. With the same step again score weighting algorithm is used for calculating individual reviews score. In addition to this its made to include rule based semantic based sentiment detection algorithm. This is used to handle semantic based opinion reviews. The third step is sentiment summarization step. In this step aggregate polarity for each product feature and its sub feature with its corresponding score value is supplied for an algorithm which used to draw charts for visualization purpose.

In this chapter some of the already presented resource in the area of opinion mining for English text that can help us in implementing and testing the new proposed model prototype and its challenge is going to be discussed with brief explanation. This resource can be freely available resources than can be found online. Additionally it is made to have for the review with semantic nature. For this it is made to have Wikipedia corpus which is freely available online. There are also a tools which is considered to be helpful for this model. Some of the

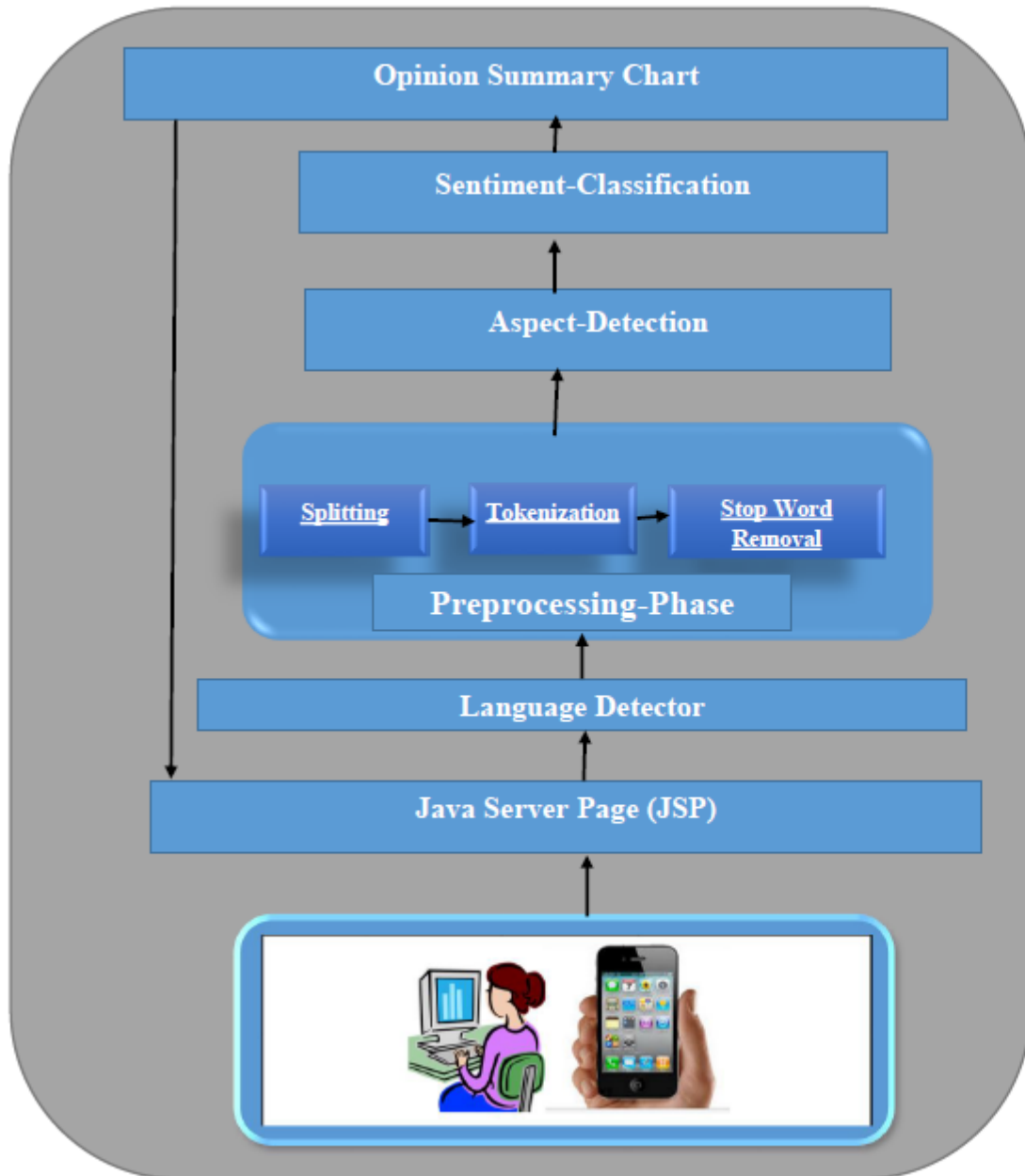


Figure 4.1: Aspect Based Opinion Mining Model for Afaan Oromo and Amharic Language

tools that are used in this model is: freely available English text parser is incorporated to an algorithm. The other is some of programming language tools used.

## 4.2 General system architecture

In this the general system architecture of AOMASS (automatic opinion mining and summarization system) is going to be briefly explained with its corresponding figure.as shown in figure 4.1. AOMASS is composed of various phases in which each and every individual phases has its own proposed role. In this phase we have structure to unstructured text formation phases, text preprocessing phases, the three commonly used aspect detection and classification phases. There is also dataset which are used for semantic and syntactic purpose.

1. ሆስፒታል የ ማገገን ከንብር እስኞ ላይ እስገራሚ ው: (Amharic)
2. The hospital was amazing at cancer diagnosis(English)
3. ውሆስፒታል የ ማገገን ከንብር እስኞ ላይ እስገራሚ ው: (Amharic)
4. tajajili hospitala kana bayee gaari dha.(Afaan Oromo Language)

Figure 4.2: Review Document for single hospital

### 4.2.1 Proposed Techniques

In this section we are going to discuss the general architectural design and functioning of the system in detail. The general architecture of the proposed model were discussed therein before for opinion mining and summarization System. This system can be carried out in four main phases. The input to the systems can be opinion document or web URL for hospitals. The output result is opinion summary of a single document

#### Pre-Processing

This step consists of activities such as sentence splitting, stop word removing, tokenizing and stemming. Each individual activities are discussed as mentioned below.

#### Sentence Splitter

Sentence splitting is the first activity in pre-processing step. This activity is responsible for splitting up the review document supplied from pervious step. This document review is pre request for the task of splitting up the document review in to individual review. This is carried out by considering some of the punctuations that is used at the end of sentence. This is called delimiter. It could be full stop (.) or exclamation mark (!) or space ( ).

The review document mentioned above could be changed in to individual sentence using regular expression

#### Tokenization

This is the second activity in pre-processing step, in this activity the sentence that were generated from pervious step were supplied to this tokenizer. Then.Tokenizer will segment this sentences in to chunk of sentence called token. This could be carried out through delimiters such as spaces or new lines.

#### Stop Word Removal

This is the third activity in pre-processing step, the token generated are supplied to stop word remover. This stop word remover can remove those stop words that do not have meaning such as: is, and that, the, there, which etc. For this purpose, stop word lists are prepared and the token that matches the stop words are removed. But, the stop word that have been created are different from other regular kind of stop word removal since it contain term like not, un- etc. which can change the opinion of the terms. (For example: *“I bought an iPhone yesterday and the screen light is wonderful but the camera light is not good!!”*). From the above review ‘an’, ‘and’, ‘is’, ‘but’, ‘the’, can be removed since it doesn’t show any meaning or features.

## **Stemming**

This activity is followed after stop word removal has been carried out. This activity are performed by receiving an input from stop word remover. The activity is used to minimize the morphologically variant words into their stem form. It's important in increasing the performance of the system by converting the words that are the same in meaning but differs in their morphology into common stem or root term. (For example: *amazingly and amazing can be converted into amaz-.*)

## **Language Detector**

The main objective of this corpus was to detect whether a given language is Afaan Oromo language, Amharic language, and English language. This corpus have no connection with opinion classification. But, language detector was used to detect whether a given language that are coming from particular document is Afaan Oromo language, Amharic language, English language. We have used the same algorithm (Vector Space Model) that was used for sentiment classification.

## **Language Corpus**

For preparation of Language corpus, data were collected from online from the website of Afaan Oromo VOA (Voice of America) from online site [www.voa.com](http://www.voa.com), ORTVO (Oromia Radio and Television organization), YouTube and Facebook. Data were chosen from various organization this is to obtain language representative data. Document were kept in such a way that its easily usable by the machine learning algorithm. To prepare language corpus for Amharic language, we have visited and then collected various website that contain Amharic content to make Amharic representative corpus. We have collected data from different sites to make language representative of various domain. Data were collected from website such as [www.youtube.com](http://www.youtube.com), [www.facebook.com](http://www.facebook.com), VOA Amharic version, [www.voa.com](http://www.voa.com), [www.diretube.com](http://www.diretube.com) .

## **Machine Translation (MT)**

Machine translation remains as hot research areas many years ago in Natural language processing. Machine translation are used to translating particular in a given language (source language) into other language (target language). Machine translation were used in this study for using an advantage of over resourced language like an English. Since, English language has resource such as Wikipedia and parser.

The translation were made to use an advantage that was available by highly effective parser and abundantly available Wikipedia dump. Basically, machine translation can be classified into various categories such as Rule-based, statistical based, example based, hybrid based, principle based, knowledge based, and online interactive based methods. For this study we are rule based approach were followed. Rule based approach were followed due to less availability of Afaan Oromo language to train language model for machine learning. The data of 500 sentence were collected for Afaan Oromo and Amharic Language.

## **Rule Based**

According to [12], rule based machine translation was the first machine translation that was used for translating source text to its target text. Effectiveness of rule based machine translation system less and provide nice grammatical result if appropriate parse was found. Rule based approach can also classified into three granularity such as direct, transfer and Interlingua. Direct approach has no any intermediary for translation. Transfer approach used in three steps analysis, transfer and generation.



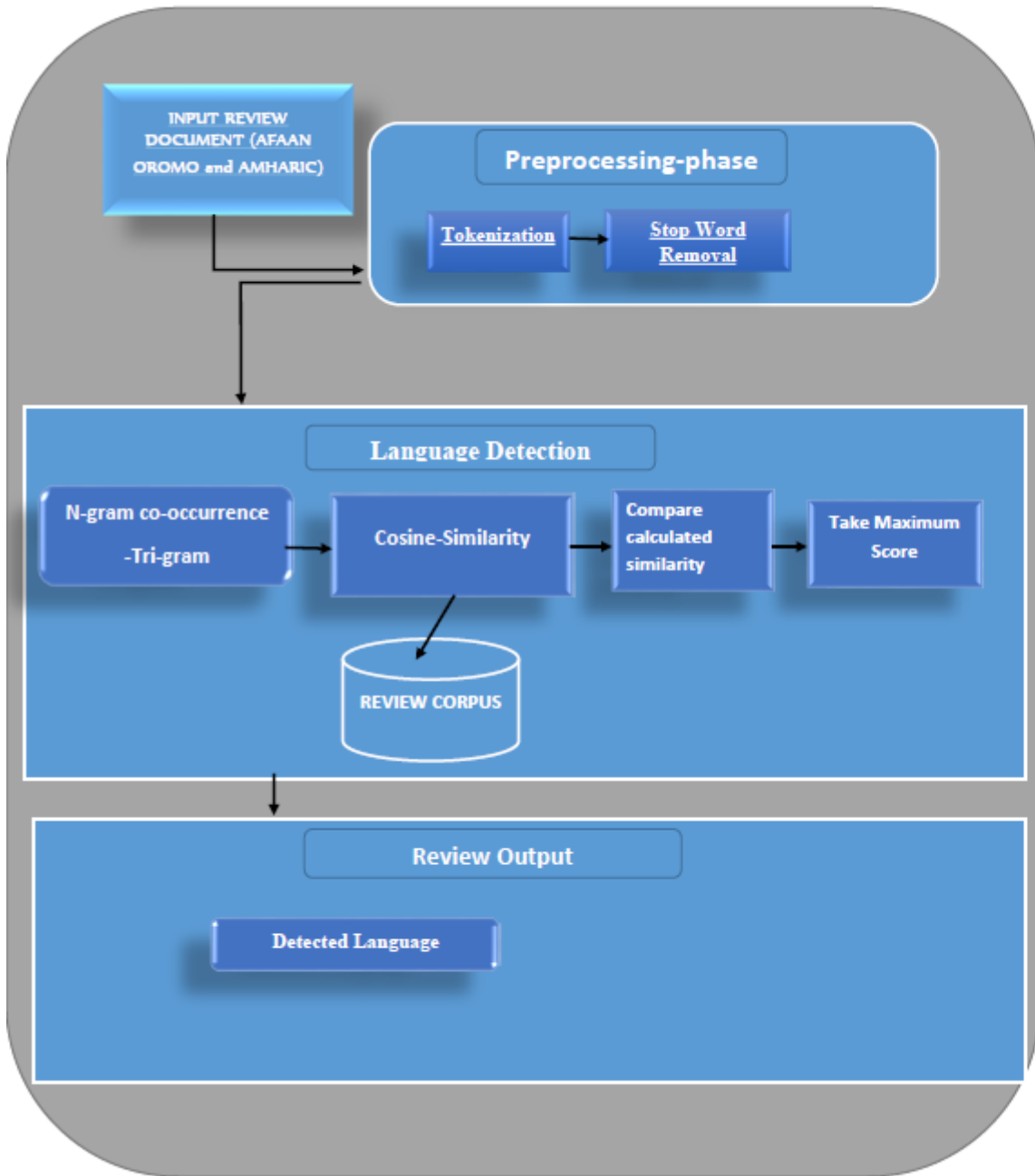


Figure 4.3: Language Detection Model

```

<lang id="1001">
<amharic>ሆስፒታል የሰርቪካል ነቀርሳ ምርመራ ላይ የተሻለ ነው.</amharic>
  <english> this hospital is best at cervix cancer diagnosis.
  </english>
</lang>

<lang id="1002">
<amharic>ሆስፒታል የማጎጸን ካንሰር ሕክምና ላይ የተሻለ ነው.</amharic>
  <english>the hospital is best at cervical cancer treatment.</english>
</lang>

<lang id="1003">
<amharic>ሆስፒታል የማጎጸን ካንሰር ሕክምና ላይ አስገራሚ ነው.</amharic>
  <english>the hospital is amazing at cervical cancer treatment.</english>
</lang>

<lang id="1004">
<amharic>ዋው ሆስፒታል የማጎጸን ካንሰር ሕክምና ላይ አስገራሚ ነው.</amharic>
  <english>wow the hospital is smart at cervical cancer treatment.</english>
</lang>

```

Figure 4.4: Sample XML Parallel corpus for Amharic to English language Translation

Additionally, Transfer approach are also composed of complex rules. Interlingua approach can also performed into two steps. The first step was the conversion of source language text into Interlingua. Then after converted into Interlingua the semantic of the text is going to be analyzed. For this study we are going to apply direct approach of translation. This approach was selected due to less translation data available online and less number of linguistic expert available For Afaan Oromo and Amharic language.

### The proposed algorithm

After preprocessing steps has been carried out, there are generally four steps are followed. The first one is feature detection steps, for this algorithm 4.1 is used. The second one is sentiment classification steps for this, algorithm 4.2 and 4.3 are used. The third one is feature and sentiment clustering. The fourth one is polarity detection, for this algorithm 5 is used.

### 4.2.2 Sentiment Classification

According to [2] Sentiment classification is the most commonly studied area under natural language processing field of study. Having a set of review document  $D$ , it helps to predict the polarity of the individual document  $d \in D$  on a given object or topic. For instance, Polarity prediction by system for movie review into two granularity such as positive and negative. This kind of classification are distinct from the regular classification of topic based classification, since it cannot utilized to classify document under defined class topic e.g. Politics, science and sports[56].

```

try {
    File fXmlFile = new File("C:\\Users\\robot\\Desktop\\parellel.xml");
    DocumentBuilderFactory dbFactory = DocumentBuilderFactory.newInstance();
    DocumentBuilder dBuilder = dbFactory.newDocumentBuilder();
    Document doc = dBuilder.parse(fXmlFile);
    doc.getDocumentElement().normalize();
    System.out.println("Root element : " + doc.getDocumentElement().getNodeName());
    NodeList nList = doc.getElementsByTagName("lang");
    for(int i=0; i<document_tobechecked1.size(); i++){
        String af=document_tobechecked1.get(i);
        for (int temp = 0; temp < nList.getLength(); temp++) {
            Node nNode = nList.item(temp);
            if (nNode.getNodeType() == Node.ELEMENT_NODE ) {
                Element eElement = (Element) nNode;
                if(eElement.getElementsByTagName("afaanoromo").item(0).getTextContent().contentEquals(af)){
                    goesforopinion.add(eElement.getElementsByTagName("english").item(0).getTextContent());
                }
            }
        }
    }
}

```

Figure 4.5: Sample Java code to convert Afaan Oromo to English language Translation

Review	English	Amharic	Afaan Oromo
1	This hospital is good at cervical cancer treatment.	ይህ ሆስፒታል የማገገን ካንሰር ሕክምና ላይ ጥሩ ነው።	Kaanseri gadamessa irrati hospitali kun mishadha.
2	Amazing hospital!!	አስደናቂ ሆስፒታል ።	Hospital kun bayee nama ajaibsisa
3	The doctor of this hospital is clever.	በዚህ ሆስፒታል ውስጥ ያለው ሐኪም ገበዳ ነው።	Doctori hospital kana bayee cimaa dha.
4	The hospital is better at treating cervical cancer.	የሆስፒታሉ የማገገን ነቀርሳ ለመያዝ የተሻለ ነው።	Hospitali kun kaanseri gadamesa irrati wayaa dha.
5	I got better treatment at this hospital.	እኔ በዚህ ሆስፒታል ውስጥ የተሻለ ህክምና አገኘሁ።	Hospital kana irrati fayee.
6	Something is better than nothing!!	ከቶላ ይሻላል ጭምዳ።	Waan hin jire irra waan xino waya.

Figure 4.6: Sample Review of Hospital Opinion

In this study, we have developed hand crafted rule on to vector space model (VSM) to classify the reviews that were collected from patients. The step to classify the sentiment are shown and discussed under figure 4.5 and algorithm below. An English resource such as POS-Tagger are utilized through translation to increase the effectiveness of the local language such as Afaan Oromoo and Amharic Language if they are not found in opinion equivalent corpus.

### 4.2.3 Part-of-Speech Tagger

Most of researcher indicate that adjectives are the main indicators of the opinion. Moreover, in our research we can also use adjectives as an opinion indicators tag[7].Part-of-speech tagging is the process of allocating a part of speech or other lexical class marker for each token in a given sentence. The part of speech tags can be assigned to words in a given sentence based on its context. The most commonly used and applied tag set in opinion mining for sentiment classification are article, noun, verb, adjective, and preposition, number and proper noun.

To detect opinion term we also select Adjective to identify an opinion terms. For this study our algorithm can understand an opinion as adjective if it is different from adjective,

Number #	Equivalent Verb and Adjectives (Afaan Oromoo)
1	{midhaagsuu: tolchuu: sireessuu}
2	{bareeda:midhaagaa:mishaa:sirraayaa}
3	Dallanuu:Aaruu:dheekamuu:loluu
4	{midhaagsuu: tolchuu: sireessuu}
5	gadde:qaana'e:mataa-buuse
6	{soba:dhara:fakkeessuu}
7	{ajjeesuu:lubbaasuu:dhabamsiisuu}
8	{sodaata:dabeessa:luguna}
9	{keessumeessuu:simachuu:offitti-fudhachuu:yaayefachuu }
10	{ moofaa:turaa:dullacha }

Figure 4.7: Sample Opinion Term and its Adjective Equivalent for Afaan Oromoo Language

opinion terms should be converted into adjective or Verb equivalent opinion and provided for vector space model. This step helps an algorithm to handle other tags such as noun and verb by converting into equivalent adjective terms. For Pos-tagging we have used online available Stanford Tagger that is found on website <http://nlp.stanford.edu/software/lex-parser.shtml>.

Examples (Amharic):

- The sentence ሆስፒታል የሚገኝ ከንሰር ሕክምና ላይ አስገራሚ ነው :: for this Amharic review sentence, the term “አስገራሚ” is an adjective. it is determined by using Pos-Tagger after translated in to an English language if its not found in opinion equivalent corpus.. This “አስገራሚ” term will be converted in to amazing then it will be provided for Pos-Tagger and pos tagger tag using tagset JJ or adjective. Then Vsm (Vector Space Model) applied

```

MaxentTagger tagger = new MaxentTagger
("C:\\stanford-postagger-full-2015-12-09\\models
\\english-left3words-distsim.tagger");
if(wordtag[j].endsWith("_JJ"))
{
polarity=VSM.VectorSpaceModel(wordtag[j]);
}

```

Figure 4.8: Sample code for Pos-Tagging

on it. And finally, the polarity of the review sentence would be assigned as positive and assigned in Amharic as “አዎንታዊ”.

- The sentence ሆስፒታሉ አገልግሎት የማጸን ካንሰር በማኮረገጅ የተለየ ነው: for this Amharic review sentence, the term “የተለየ” is not an adjective. Hence this term with verb tag set will be converted into other adjective equivalent term after it is determined by using PosTagger and translated in to an English language. This “የተለየ” term will be converted in to “distinct then it will be provided for Pos-Tagger and pos tagger tag using tagset “JJ” or adjective. Then Vsm (Vector Space Model) applied on it. And finally, the polarity of the review sentence would be assigned as “positive and assigned in Amharic as “አዎንታዊ”.

Examples (Afaan Oromo):

- The sentence “Hospitali kun kaanseri gadamessa irrati nama ajaibsis. for this Afaan Oromo review sentence, the term “ajaibsis is an adjective. it is determined by using Pos-Tagger after translated in to an English language if its not found opinion equivalent corpus. This “ajaibsis term will be converted in to amazing then it will be provided for Pos-Tagger and pos tagger tag using tagset JJ or adjective. Then Vsm (Vector Space Model) applied on it. And finally, the polarity of the review sentence would be assigned as positive and assigned in Afaan Oromo as posativi.
- The sentence Hospitali kun kanseeri gadamessa irrati kan isaan qixaaxu hin jiru for this Amharic review sentence, the term qixaaxu is not an adjective. Hence this term with verb tag set will be converted into other adjective equivalent term after it is determined by using Pos-Tagger and translated in to an English language if its not found in opinion equivalent corpus. This qixaaxu term will be converted in to distinct then it will be provided for Pos-Tagger and pos tagger tag using tagset JJ or adjective. Then Vsm (Vector Space Model) applied on it. And finally, the polarity of the review sentence would be assigned as positive and assigned in Afaan Oromo as postivi.

#### 4.2.4 VSM (Vector Space Model)

According to [12,5], this methods can helps to group semantically similar terms in one. This is performed by calculating the similarity between opinion terms. It can be used in application that requires to generate model for verbs, adjectives and noun opinion lexicon and others. Often, in sentiment classification semantic are used in a combination with statistical methods to obtain good results[5].

Distributional models are used to place words in a fashion they appear inside the corpus. Vector space model is usually implemented under distributional model. The word that appear in the same context were assumed semantically the same. In vector space model the words can be represented as point in space. Vector space model is over predicting the meaning similarity since they use linear algebra[12].

Several recent studies [5, 12,] have suggested that VSM is recommended for semantic opinion by another researcher’s is selected due to its ability of detecting semantic sentiment. Vector Space Model (VSM) is applied using N-gram and Cosine Similarity in combination. VSM was selected since it goes with developed hand crafted rules.

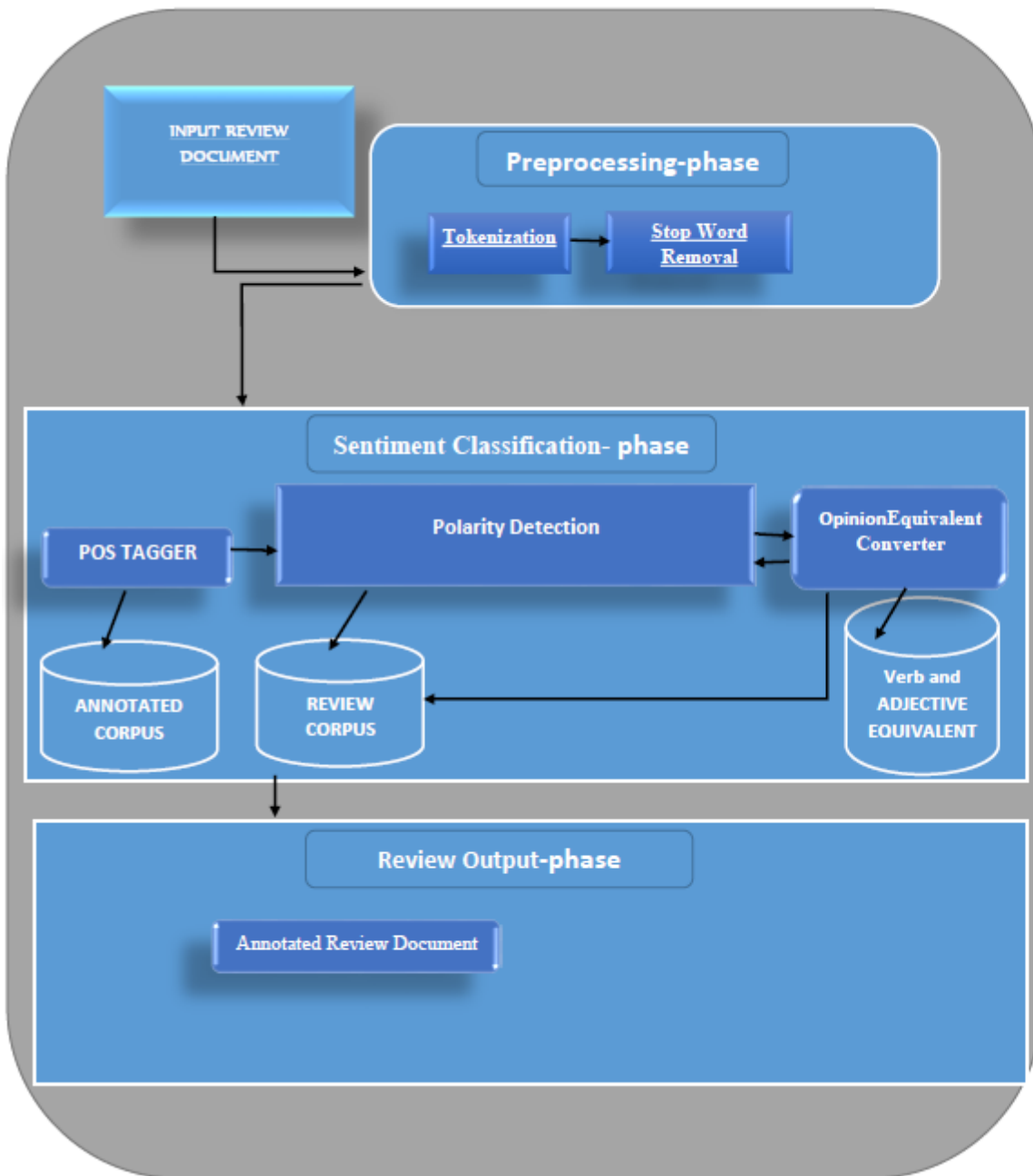


Figure 4.9: Sentiment Classification Model for Afaan Oromo and Amharic Language

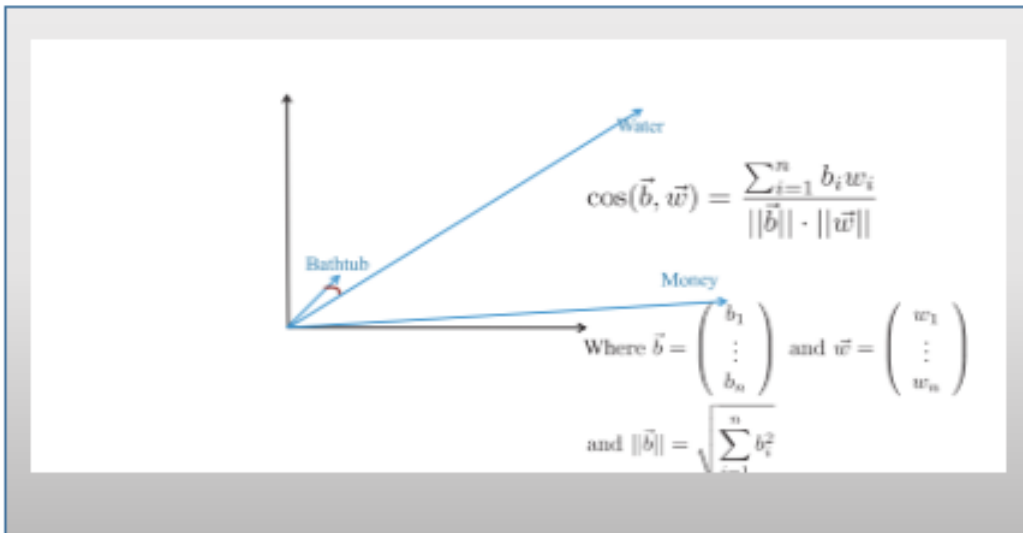


Figure 4.10: Computing similarity between bathtub and water adopted from[12].

**Step 1: START**

Step 2: Initialize **POLARITY\_SCORE**→0

Step 3: input hospital review document

Step 4: convert document into sentence

Step 5: annotate the converted sentence

Step 6: find the opinionated Term with */Opinion Tag*

Step7: assign the variable **OPT** with opinionated term

Step 8: check the existence of **OPT** variable in both positive and negative corpus

Step 9: **IF OPT** is found

**Then** create an **N-GRAM** of Input Sentence

Apply **VSM** (Vector Space Model)

Assign **POLARITY\_SCORE**

**ELSE**

Search for **Adjective/Verb equivalent** of **OPT** variable in Adjective equivalent corpus

**IF OPT Adjective/Verb equivalent** is found

**Then** replace **Adjective/Verb equivalent New OPT** with **Old OPT**

Replace **Old Input Sentence** with **New OPT**

Create an **N-GRAM** of **NEW Input Sentence**

Apply **VSM** (Vector Space Model)

```

Assign POLARITY_SCORE
IF POLARITY_SCORE >0.75
    Add NEW Input Sentence into corresponding corpus
ELSE
    Don't Add NEW Input Sentence into corresponding corpus
ELSE
    Initialize  $I \rightarrow 0$ 
    Assign new Adjective/Verb equivalent [I] until maximum Number of Equivalent is available
     $I=I+1$ 
Step 10: STOP

```

Figure 4.11: Algorithm for Sentiment Classification

#### N-gram:

N-gram is utilized as one major part of vector space model. It is used to convert query and text in corpus in to various vectors. Once it is converted into a given vector, the generated vectors are supplied to cosine similarity. This generated vectors are created based on co-occurrence count of corpus and query,the study use N-gram to handle context based sentiment detection.hence,the study employed Bi-gram to handle “Afaan oromo” and “Amharic” ambiguous opinion term, since, there are opinion term that have different meaning at various context, the term “badaa” has different meaning in different context.

Example:1, “*hospitalli kun tajaajila wal’aansa gadaamessa irratti badaa miti.*”

Example :2, “*hojjattonni hospitaala kana tajaajila wal’aansa gadaamessa irratti badaa hojjatu.*”

in Example:1 the term “*badaa*” has “**positive**” connotation in that context while in in Example:2 the term “*badaa*” has “**negative**” connotation. Hence,Bi-gram was used to solve such types of problem.

#### 4.2.5 Cosine Similarity

The vectors that were generated in previous steps are supplied to cosine similarity. This cosine similarity helps to calculate an angle between query and positive or negative text document. In one hand, the minimum angle between query and document means document and query match. Therefore, the query polarity is assigned based on angle between query and document. In other hand, the maximum an angle between query and document means the document is not selected as an appropriate cluster. Else the document is selected as an appropriate cluster.

#### Identifying the polarity of the Reviews

After vector space model was carried out the value will be feed to polarity detector this step is responsible for identifying polarity terms. The polarity weight of the individual sentiment terms in the review by the formula given in equation below.



```

while ((str = in.readLine()) != null) {
    // System.out.println(str);
    //if(str.endsWith(".")){
    //    tokenList.add(str);
    //    negative1 .add(str);

    // System.out.println(str);
    for (int n = 2; n <=2; n++) {
        for (String ngram : ngrams(n, str))
            System.out.println(ngram);

        System.out.println();
    }
}
}

```

Figure 4.12: N-Gram

**Step1: Intialize I from 0**  
**Step2: Intialize J from 0**  
**Step3:OPTERM**  
**Step4: create matrix for equivalent opinion**  
**Step5: For i>3**  
**Step6: IF WORD[I][J] NEXT IS EQUAL TO NULL**  
**THEN**  
    **EQOP=WORD[I-1][J-1]**  
    **Replace EQOP with OPTERM in a agiven sentence**  
    **Pass Replaced Sentence to VSM**

**ELSE IF WORD[I][J] NEXT IS NOT NULL AND WORD[I-1][J-1] NEXT IS START**  
**THEN**  
    **EQOP=WORD[I+1][J+1] OR EQOP=WORD[I+1][J+1]**  
    **Replace EQOP with OPTERM in a agiven sentence**  
    **Pass Replaced Sentence to VSM**

**ELSE IF WORD[I-2][J-2] OR WORD[I-1][J-1] IS START**  
    **EQOP=WORD[I+1][J+1] OR WORD[I+2][J+2]**

**ELSE**

**pass to machine translation**  
**END**  
    **END**  
        **END**

Figure 4.13: Equivalent Opinion for Afaan Oromo and Amharic Language Sentiment Orientation of reviews

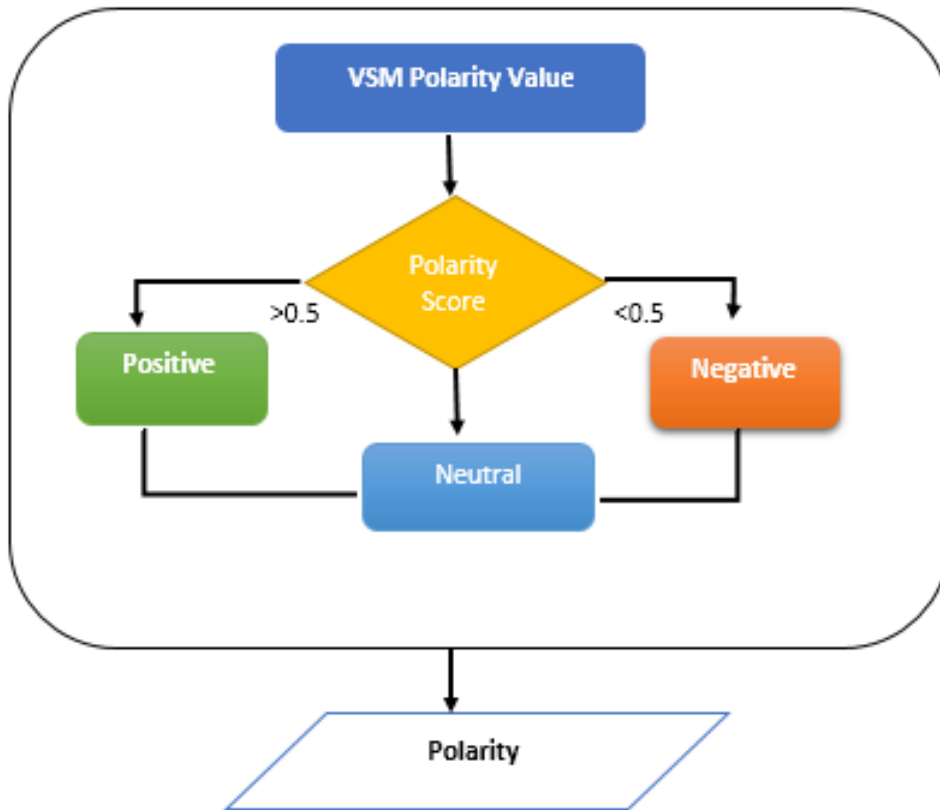


Figure 4.14: Identifying Sentiment Orientation of reviews

$$R_p = \sum_{i=0}^n T p_i \dots \dots \dots \text{Equation 1}$$

Where,  $R_p$  is review polarity value returned by VSM,  $T_p$  is sentiment term polarity value,  $n$  is number of sentiment terms within the given review and  $i$  is term instance. According to the result of the equation, if the value of  $R_p$  is greater than 0.5 then the review is categorized into a predefined category “positive”. Similarly if the value of  $R_p$  is less than zero then the review is categorized in to a predefined category “negative”. Finally if the total average weight of all the individual terms is equal to zero or equal to 0.5, the review is categorized in to the category “neutral”. but, neutral don’t have significance to know the strong side and weak side of Ethiopian hospital.hence, the study don’t include such types of polarity classification.

**Aspect Detection**

This step is planned to discuss some activities that are carried out for detecting aspect and its underlying algorithm. This step has carried out through Rule based Aspect detection. Most of the time an aspect extraction technique with Noun and Noun Phrase are employed in various research. For this study, aspect detection other than Noun and Noun Phrase were tried using algorithm shown below. for aspect detection Nave classifier is selected to annotate the input review document after evaluation of various algorithm was carried out.

Algorithm 4.4 Algorithm 4.4.1. Aspect detection

1. for every Review sentence in rs d
2. Apply Annotator(naive bayes ) 3. if SASP or SSASP tag exist in sentence
  - 3.1. then assign word as Aspect
  - 3.2.1 Assign Aspect

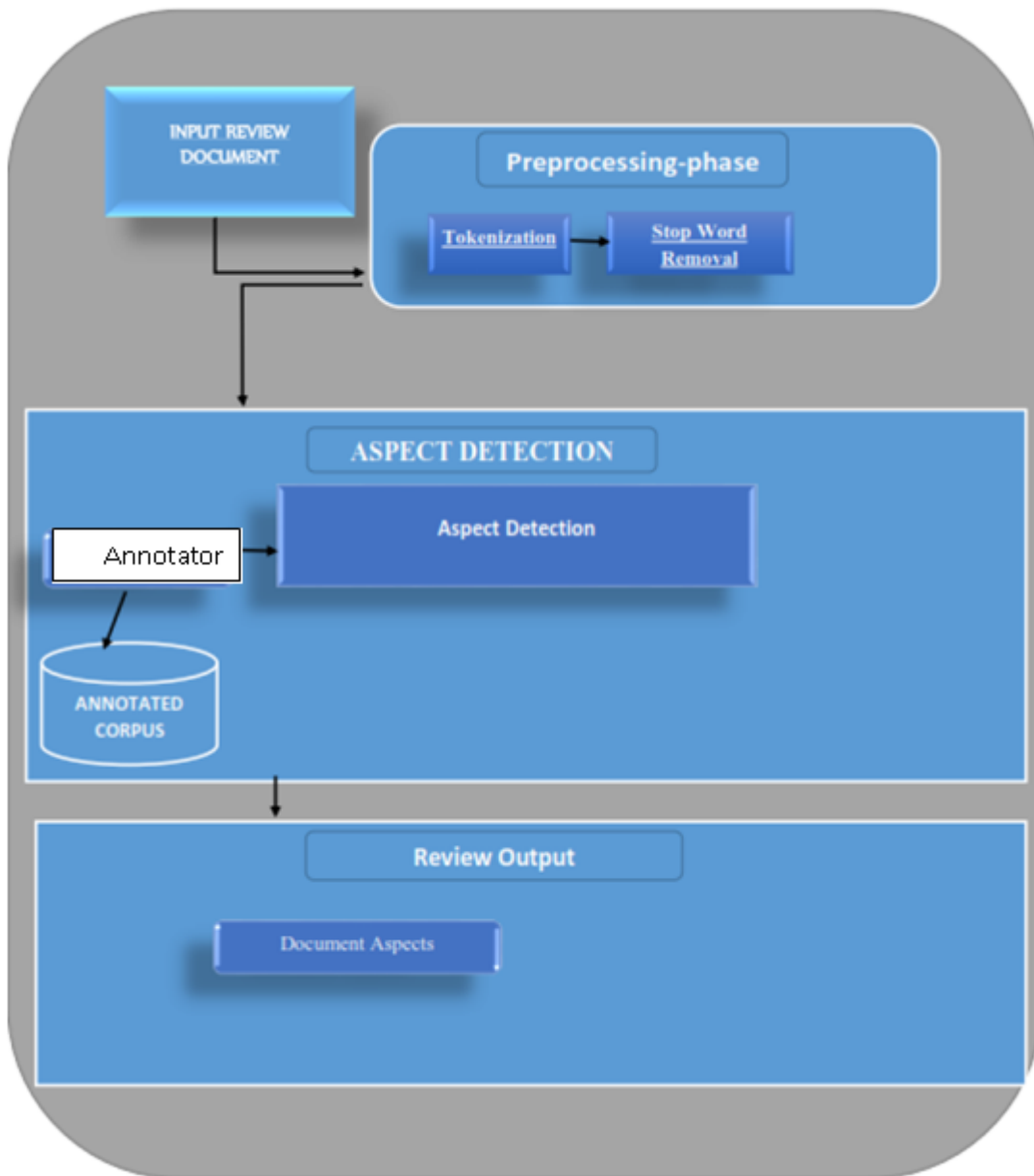


Figure 4.15: Aspect Detection Model

Evaluation	Cosine Similarity with N-gram(VSM)	Cosine Similarity with TF-IDF(VSM)	Naïve Bays	HMM(Hidden Markov Model)	Cosine Similarity with N-gram(VSM) +Rule Based
<b>Precision</b>	<b>52%</b>	<b>49%</b>	<b>84.2%</b>	<b>52%</b>	<b>73%</b>
<b>Recall</b>	<b>64%</b>	<b>49%</b>	<b>74.2%</b>	<b>48%</b>	<b>62%</b>
<b>F-Measure</b>	<b>58%</b>	<b>49%</b>	<b>79.1%</b>	<b>50%</b>	<b>68%</b>

Table 4.1: Aspect Detection Evaluation Comparison between various algorithms.

**Probabilistic Classifier** Here under, two commonly used probabilistic classifiers such as naive Bayes and Bayesian network are discussed below

**Naive Bayes Classifier (NB)** Naive Bayes classifier is used by many researcher since easily understandable. This method use Bayes rule in order to classify a given instance to its appropriate class. It works based on the lexical probability of a given document[5].

$$P(\text{label}|\text{features}) = \frac{P(\text{label}) * P(\text{features}|\text{label})}{P(\text{features})}$$

P (label) is defined as the probability of likelihood of a label.

P (features |label) is the likelihood probability of a given class appeared as label.

P (feature) is the likelihood probability of feature.

Sample Aspect of review in corpus

Text Based Summary Algorithm 4.5. Opinion summary

1. For every Review sentence in rs d
2. Apply sentiment classification and store as key
3. Apply aspect detection store as value
4. Iterate through the aspect and its corresponding sentiment
5. feed to chart
6. display chart

### Chart based Summary

Visualization helps to make an opinion presented easily readable and understandable format. This visualization do not only benefit customers who use an opinion to make decision but it can also benefit other researcher to show the result on the polarity of an opinion generated. We have selected visualization way of summarizing an opinion because of its easily readability and understandability. In addition to this its most popularity in various business activity also make it preferable for the study.

**Hospitali /ASP Kun /NA kanseeri /SASP gadameesa /SSASP irrati/NA bayee/NA gaaridha/OP. /PUNC**

**Kanseeri/SASP gadameesa /SSASP waldhanu/NA irrati /NA hospitalich/ASP kun/NA mootidha/OP**

**Hospitalich/ ASP Kun kanseeri /SASP gadameesa/SSASP irrati/NA tajjili/NA isaa/NA bayee/NA gaari/OP dha/ NA. /PUNC**

**Hospitalich /ASP kun/NA kanseeri /SASP gadameesa/ SSASP irrati/NA tajjili/NA isaa/NA bayee/NA misha /OP dha/NA. /PUNC**

Figure 4.16: Annotated Afaan Oromoo Aspect

Aspect	English	Amharic	Afaan Oromo
1	cervix	የሜንጸን	Gadamessa
2	Lung	ሳምባ	Somba
3	Blood	ደም	Dhiiga
4	Breast	ጡት	Harma
5	Boone	አጥንት	Laafe

Figure 4.17: Sample Detected Aspect by the System

## Aspect Based Sentiment Summarization for Bilingual (Afaan Oromo,Amharic&English) Text

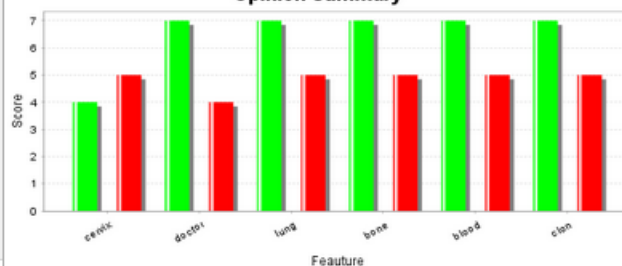
Please enter text to see its sentiment prediction and summary results:

```
this hospital is bad at cervix.  
this hospital is good at cervix cancer  
diagnosis amazing.  
this hospital is good at cervix cancer  
diagnosis amazing.  
this hospital is good at cervix cancer  
diagnosis amazing.  
this hospital is good at cervix cancer  
diagnosis amazing.  
this hospital is bad at cervix diagnosis..  
this hospital is bad at service diagnosis..  
this hospital is smart at service diagnosis..  
this hospital is bad at cervix.  
this hospital is good at cervix cancer  
diagnosis amazing.  
this hospital is good at cervix cancer
```

You can also upload a file (limit 200 lines):

No file chosen

Opinion Summary



Show trees in binary form

implemented for the seek of research evaluation

Figure 4.18: Aspect based sentiment summarization Prototype

### 4.2.6 Corpus Preparation

For training semantic based approach selected yet, since the method we followed is semantic based approach we have collected 1500 review sentence that consists of 750 positive and Negative 750, which are composed of various aspect from online site: <https://www.patientopinion.org.uk/>. Afaanoromoo dataset were collected in two versions: the first version is for training set and the second version is for testing set. In one hand, the dataset for training was Prepared in two forms:the first one is the data from online English patient opinion is translated into Afaanoromoo using English to Afaanoromoo language translator. The translated opinion was provided for afaanoromoo linguistic experts to correct its language meaning and the polarity of the sentiments. Then the corrected opinions were used to train the model with selected hybrid approach.

The other was collected manually from patient (customer) of the hospitals by using purposive sampling techniques on cancer department of hospitals.We have used this sampling technique because our focus domain is cancer. In the other hand,the datasets (reviews) used for conducting the experiment are collected manually from five hospitals follow-up patients.

The Evaluation or Test data were collected from five hospitals in Addis Ababa that were currently engaged in providing cancer care, the doctors were responsible for this particular activity. The involvement of doctors was due to the sensitiveness of the data regarding privacy of the patient. Totally 70 opinion from different customers of each hospitals were collected. The opinion collected from five hospital were 350. From this opinion data were provided in different language that were spoken in Ethiopia.

From this language, around 150 opinions were provided in Amharic language. 100 opinions were provided in Afaan Oromo language. And others opinions were provided in other Ethiopian local language. Since other opinion is out of the scope of the study we do not deal with it. Then Rule based translation approach were applied due to less resource in local language like Amharic and Afaan Oromo language. Hence, translation made to English language for collecting English Amharic corpus. Then this corpus showed to linguistic expert and approval was made. We have also showed Afaan Oromo English translation to linguistic expert for approval.

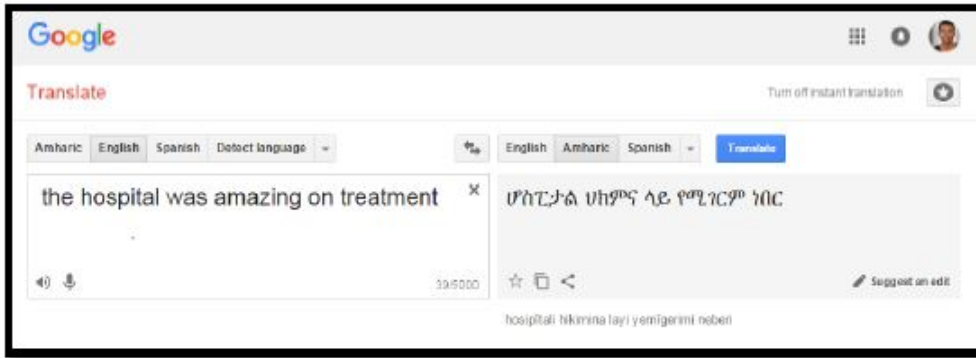


Figure 4.19: google translate for translation of online English opinion of patient corpus into Amharic Language corpus.

This approval was also made for its opinion.

#### 4.2.7 Manual collection of Opinion and Its Aspect

Since the algorithm we have used was semantic based algorithm for training opinion was manually collected from various sites told above for both Afaan Oromo and Amharic Language with the help of linguistic expert. Machine translation plays an important role for translating Afaan Oromo and Amharic Language aspect as well as opinion to English language. An opinion were gathered from five hospitals in Addis Ababa such as: ALERT, Black Lion Hospital, Teklehaimanot General Hospital, Zewditu Memorial Hospital, and St.Pauls Hospital.

Those opinion were provided to linguistic expert to annotate its polarity and check the aspect of the reviews in its test set. The reviews collected from five hospital was exactly 350. Then, annotation was made by linguistic expert for both opinion and aspect. Finally, annotated opinion and aspect was evaluated against the system output of the same reviews.

### 4.3 Evaluation

The common standard of information retrieval (IR) effectiveness evaluation techniques are used in order to evaluate the developed prototype of the system. Precision, recall and F-measure metrics are used in order to measure the retrieval effectiveness. Precision is a metric used to provide ratio of the number of relevant retrieved document to total number of document retrieved.

$$\text{Precision} = \frac{\text{Number of relevant documents retrieved}}{\text{Total Number of retrieved documents}} \dots \text{equation1}$$

Recalls is a metric used to provide ratio of the number of relevant retrieved document to total number of relevant document.

$$\text{Recall} = \frac{\text{Number of relevant documents retrieved}}{\text{Total Number of relevant documents in database}} \dots \text{equation2}$$

F-measure is another metric that considered both precision and recall this this metric often performed by harmonizing both precision and recall together. F-measure is defined as:

$$F\text{-measure}=(2*\text{percision}*\text{recall})/(\text{Percision}+\text{Recall})\text{..equation3}$$

## 4.4 Experiment and Result Discussion

The common standard of information retrieval (IR) effectiveness evaluation techniques are used in order to evaluate the developed prototype of the system. Precision, recall and F-measure metrics are used in order to measure the retrieval effectiveness. Precision is a metric used to provide ratio of the number of relevant retrieved document to total number of document retrieved.

$$\text{Precision}=(\text{Number of relevant documents retrieved})/(\text{Total Number of retrieved documents})\text{..equation1}$$

Recalls is a metric used to provide ratio of the number of relevant retrieved document to total number of relevant document.

$$\text{Recall}=(\text{Number of relevant documents retrieved})/(\text{Total Number of relevant documents in database})\text{.equation2}$$

F-measure is another metric that considered both precision and recall this this metric often performed by harmonizing both precision and recall together. F-measure is defined as:

$$F\text{-measure}=(2*\text{percision}*\text{recall})/(\text{Percision}+\text{Recall})\text{..equation3}$$

## 4.5 Results

For sentiment classification evaluation experiment were conducted in two different experiment were performed on five hospitals that are found in Addis Ababa and currently serving public such as ALERT, Black Lion Hospital, Teklehaimanot General Hospital, Zewditu Memorial Hospital, St.Pauls Hospital .

Every system is developed to solve a given problem and to know the extent to which the system attain the required goal is going to be evaluated toward the real situation through effectiveness and efficiency of the system. Efficiency can be the amount of time taken by the system to produce a given output and effectiveness related to the extent to which the problem is solved. The study going to use precision, recall and F-measure to evaluate an aspect detection and sentiment classification.

For testing purpose purposive sampling technique were applied. We have manually collected testing reviews by providing form to doctors in hospital because of the privacy purpose and the collected reviews were from cervical cancer follow-up patient (customer). The table below show an experiment that was conducted on five hospitals that are currently providing cervical cancer treatment.



<b>Hospital</b>	<b>Manually identified Aspect</b>	<b>Identified by the system</b>	<b>Identified by the system correctly</b>	<b>Precision</b>	<b>Recall</b>	<b>F-Measure</b>
Hospital	50	40	35	0.875	0.7	0.78

Table 4.2: Aspect Detection: Expriment1 (Amharic Language)

<b>Hospital</b>	<b>Manually identified opinion</b>	<b>Identified by the system</b>	<b>Identified by the system correctly</b>	<b>Precision</b>	<b>Recall</b>	<b>F-Measure</b>
Hospital	50	44	39	0.89	0.78	0.84

Table 4.3: Aspect Detection: Expriment2 (Amharic Language)

<b>Hospital</b>	<b>Manually identified opinion</b>	<b>Identified by the system</b>	<b>Identified by the system correctly</b>	<b>Precision</b>	<b>Recall</b>	<b>F-Measure</b>
Hospital	50	44	39	0.89	0.78	0.84

Table 4.4: Sentiment Classification: Expriment1 (Amharic Language) based on data collected for two months.

## 4.6 Discussion

### 4.6.1 Amharic Review discussion

Aspect detection on hospital data set was carried out under experiment1 and the result with 87.5 % of Precision and 61 % of recall was registered. The precision was better when compared to recall, the result of recall registered less as compared to precision this is due to some aspect other than noun and noun phrase presented as noun by part of speech tagger. Therefore, selecting best part of speech tagger can increase the performance registered by the system.

Aspect detection the precision was better when compared to recall, the result of recall registered less as compared to precision this is due to some aspect other than noun and noun phrase presented as noun by part of speech tagger. But when we compare experiment1 that registered 0.875, 0.7 and 0.78 respectively precision, recall and F-measure with experiment2 that registered 0.89, 0.78 and 0.84 the different score was registered this was due to an algorithm ability to learn from an experience or once machine made the mistake it has an ability to learn from an experience.

For sentiment classification experiment1: the highest score was registered by Tekle Haimanot General Hospital showed as 0.69, 0.65 and 0.67 of precision, recall and F-measure respectively. And the lowest score was registered by St.Pauls Hospital with 0.35, 0.3 and 0.32 of precision, recall and F-measure respectively.

<b>Hospital</b>	<b>Manually identified opinion</b>	<b>Identified by the system</b>	<b>Identified by the system correctly</b>	<b>Precision</b>	<b>Recall</b>	<b>F-Measure</b>
Black Lion Hospital	150	139	92	0.66	0.61	0.63
Tekle Haimanot General Hospital	150	142	98	0.69	0.65	0.67
ALERT Hospital	150	134	87	0.69	0.58	0.63
Zewditu Memorial Hospital	150	137	66	0.48	0.44	0.46
St.Pauls Hospital	150	128	45	0.35	0.3	0.32
<b>AVERAGE</b>				<b>0.574</b>	<b>0.516</b>	<b>0.542</b>

Table 4.5: Sentiment Classification: Exprimment2 (Amharic Language) based on data collected for two months.)

<b>Hospital</b>	<b>Manually identified Aspect</b>	<b>Identified by the system</b>	<b>Identified by the system correctly</b>	<b>Precision</b>	<b>Recall</b>	<b>F-Measure</b>
Black Lion Hospital	150	145	97	0.67	0.64	0.66
Tekle Haimanot General Hospital	150	142	98	0.69	0.65	0.67
ALERT Hospital	150	136	89	0.65	0.59	0.63
Zewditu Memorial Hospital	150	140	72	0.51	0.48	0.49
St.Pauls Hospital	150	130	48	0.36	0.32	0.34
<b>AVERAGE</b>				<b>0.576</b>	<b>0.536</b>	<b>0.56</b>

Table 4.6: Aspect Detection: Exprimment1 (Afaan Oromo Language Reviews)

For sentiment classification exprimment2: the highest score was registered by Tekle Haimanot General Hospital showed as 0.69, 0.65 and 0.67 of precision, recall and F-measure respectively.

<b>Hospital</b>	<b>Manually identified Aspect</b>	<b>Identified by the system</b>	<b>Identified by the system correctly</b>	<b>Precision</b>	<b>Recall</b>	<b>F-Measure</b>
Hospital	30	18	15	0.83	0.5	0.62

Table 4.7: Aspect Detection: Exprimet2 (Afaan Oromo Language Reviews)

<b>Hospital</b>	<b>Manually identified Aspect</b>	<b>Identified by the system</b>	<b>Identified by the system correctly</b>	<b>Precision</b>	<b>Recall</b>	<b>F-Measure</b>
Hospital	30	21	21	1	0.7	0.82

Table 4.8: Sentiment Classification: Exprimet1 (Afaan Oromo Language Reviews) based on data collected for two months.)

<b>Hospital</b>	<b>Manually identified opinion</b>	<b>Identified by the system</b>	<b>Identified by the system correctly</b>	<b>Precision</b>	<b>Recall</b>	<b>F-Measure</b>
Black Lion Hospital	100	85	73	0.86	0.73	0.79
Tekle Haimanot General Hospital	100	67	62	0.93	0.62	0.74
ALERT Hospital	100	78	74	0.95	0.74	0.83
Zewditu Memorial Hospital	100	75	71	0.95	0.71	0.81
St.Pauls Hospital	100	67	45	0.67	0.45	0.54
<b>AVERAGE</b>				<b>0.872</b>	<b>0.65</b>	<b>0.742</b>

Table 4.9: Sentiment Classification: Exprimet2 (Afaan Oromo Language Reviews))

And the lowest score was registered by St.Pauls Hospital with 0.36, 0.32 and 0.34 of precision, recall and F-measure respectively. This doesnt mean always the Tekle Haimanot General Hospital better than St.Pauls Hospital, since the registered score also depends on the language ability of the opinion holders (patient) or opinion complexity of text. Patient in St.Pauls may provide the reviews that are full of semantic knowledge. Additionally, Tekle Haimanot Hospital result was remains the same for both experiment this is due to all opinion registered were less than 75 %, since algorithm can drop sentiment polarity score less than 75 %. But the scored result of both experiment show good result, this is due to the language nature of Amharic,

<b>Hospital</b>	<b>Manually identified Aspect</b>	<b>Identified by the system</b>	<b>Identified by the system correctly</b>	<b>Precision</b>	<b>Recall</b>	<b>F-Measure</b>
Black Lion Hospital	100	87	80	0.91	0.80	0.85
Tekle Haimanot General Hospital	100	74	72	0.97	0.72	0.83
ALERT Hospital	100	78	69	0.85	0.69	0.76
Zewditu Memorial Hospital	100	80	71	0.89	0.71	0.79
St.Pauls Hospital	100	68	51	0.75	0.51	0.6
<b>AVERAGE</b>				<b>0.87</b>	<b>0.69</b>	<b>0.77</b>

Table 4.10: Aspect Detection: Expriment1 (Afaan Oromo Language Reviews))

<b>Hospital</b>	<b>Manually identified Aspect</b>	<b>Identified by the system</b>	<b>Identified by the system correctly</b>	<b>Precision</b>	<b>Recall</b>	<b>F-Measure</b>
Hospital	30	18	15	0.83	0.5	0.62

Table 4.11: Aspect Detection: Expriment2 (Afaan Oromo Language Reviews))

<b>Hospital</b>	<b>Manually identified Aspect</b>	<b>Identified by the system</b>	<b>Identified by the system correctly</b>	<b>Precision</b>	<b>Recall</b>	<b>F-Measure</b>
Hospital	30	18	15	0.83	0.5	0.62

Table 4.12: Aspect Detection: Expriment2 (Afaan Oromo Language Reviews))

most of monolingual Amharic speaker do not provide simple machine understandable reviews. This is due to nature of Amharic language is more of pragmatics, idiomatic that were very difficult for any machine. additionally, the difference of language words from place to place in the country.

Generally, For sentiment classification expriment1 has registered average score of 0.574, 0.516, 0.52 precision, recall and F-measure respectively. Then after expriment1 has carried out expriment2 followed and registered average score of 0.576, 0.536, 0.56 precision, recall and F-measure respectively. From this result we can understand that the developed algorithm can

Language	Summarizer	Number of Comments	Man Power	Time
Afaan Oromo	Human Summarizer	50 comments	2 person	1:15 minutes
	Machine	50 comments	1 person	20 sec
Amharic	Human Summarizer	50 comments	2 person	1:35 minutes
	Machine	50 comments	1 person	24 sec

Table 4.13: Human and Machine Summarizer )

Evaluation	Cosine Similarity with N-gram(VSM)	Cosine Similarity with TF-IDF(VSM)	Naïve Bays	HMM(Hidden Markov Model)	Cosine Similarity with N-gram(VSM) +Rule Based
<b>Precision</b>	<b>67%</b>	<b>51%</b>	<b>93%</b>	<b>52%</b>	<b>84.2%</b>
<b>Recall</b>	<b>64%</b>	<b>48%</b>	<b>62%</b>	<b>48%</b>	<b>74.2%</b>
<b>F-Measure</b>	<b>66%</b>	<b>49%</b>	<b>74%</b>	<b>50%</b>	<b>79.1%</b>

Table 4.14: Average Sentiment Classification Evaluation Comparison between various algorithms )

learn from an experience. If various experiment carried out then an effectiveness and efficiency of the system will increase.

#### 4.6.2 Afaan Oromo Review discussion

Aspect detection on hospital data set was carried out under experiment 1 and the result with 83 % of Precision, 50 % of recall and 62 % of F-measure was registered. The precision was better when compared to recall, the result of recall registered less as compared to precision this is due to some aspect other than noun and noun phrase presented as noun by part of speech tagger. Since, Afaan Oromo language in nature have more noun than others according to linguistic expert.

Therefore, selecting best part of speech tagger can increase the performance registered by the system. Aspect detection the precision was better when compared to recall, the result of recall registered less as compared to precision this is due to some aspect other than noun and

noun phrase presented as noun by part of speech tagger. But when we compare experiment1 that registered 87.5, 70% and 78% respectively precision, recall and F-measure with experiment1 that registered 100%, 70% and 82% the different score was registered this was due to an algorithm ability to learn from an experience or once machine made the mistake it has an ability to learn from an experience.

For sentiment classification experiment1: the highest score was registered by ALERT Hospital showed as 0.95, 0.74 and 0.83 of precision, recall and F-measure respectively. And the lowest score was registered by St.Pauls Hospital with 0.67, 0.45 and 0.54 of precision, recall and F-measure respectively.

For sentiment classification experiment2: the highest score was registered by Black Lion Hospital showed as 0.91, 0.80 and 0.85 of precision, recall and F-measure respectively. And the lowest score was registered by St.Pauls Hospital with 0.75, 0.51 and 0.6 of precision, recall and F-measure respectively. This doesn't mean always the Tekle Haimanot General Hospital better than St.Pauls Hospital, since the registered score also depends on the language ability of the opinion holders (patient) or opinion complexity of text. Patient in St.Pauls may provide the reviews that are full of semantic knowledge.

Additionally, Tekle Haimanot Hospital result was remains the same for both experiment this is due to all opinion registered were less than 75 %, since algorithm can drop sentiment polarity score less than 75 %. But the scored result of both experiment show good result, this is due to the language nature of Amharic, most of mono lingual Afaan Oromo do not provide simple machine understandable reviews. Generally, for sentiment classification experiment1 has registered average score of 0.95, 0.74 and 0.83 precision, recall and F-measure respectively.

Then after experiment1 has carried out experiment2 followed and registered average score of 0.91, 0.80 and 0.85 precision, recall and F-measure respectively. From this result we can understand that the developed algorithm can learn from an experience. If various experiment carried out then an effectiveness and efficiency of the system will increase.

#### **4.6.3 the overall sentiment classification algorithm discussion**

According to evaluation carried out on the figure above, Even though, precision of naive bays algorithm is high. Hybrid Cosine Similarity with N-gram (VSM) +Rule Based algorithm beats the other algorithm Therefore; an algorithm is selected to build a sentiment classification model.

### **4.7 Fact Based Information**

Fact based information is an information about hospital that one can't get through opinion[13]. The fact based information was required to for recommendation due, to inclusiveness boundary problem occurred during recommendation using only opinion. Data that was collected from hospital was only collected from follow-up patient. But fact based information can able to handle mortality of the patient. For this reason the need of combining both fact based information and opinion summary were required. For fact based information vector space model was used, sine it scores high result after evaluation of various algorithm carried out on data. It can be used the semantic similarity between old patients and new patients who are seeking for better hospital.

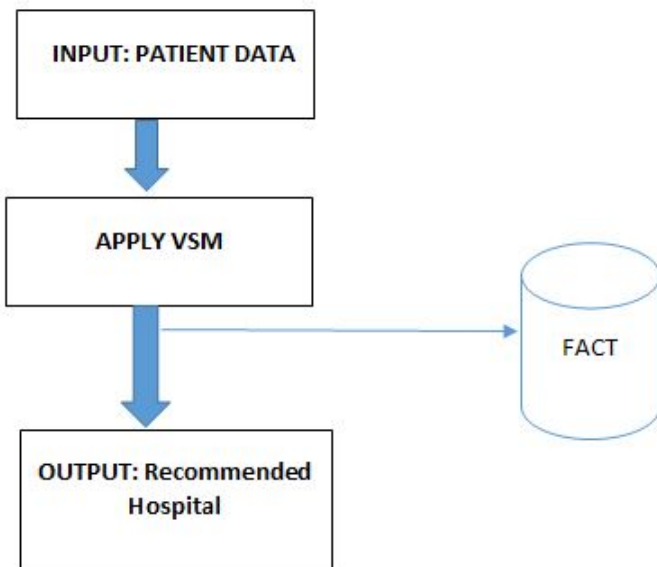


Figure 4.20: Model for fact based information

Evaluation	Cosine Similarity with N-gram(VSM)	Cosine Similarity with TF-IDF(VSM)	Naïve Bays	HMM(Hidden Markov Model)	Cosine Similarity with N-gram(VSM) +Rule Based
Precision	87%	73%	60%	52%	72%
Recall	65%	49%	62%	48%	62%
F-Measure	76%	61%	61%	50%	67%

Table 4.15: Fact Based Evaluation Comparison between various algorithms.

#### Algorithm 4.4. Fact and Opinion Joint

1. For every predicted result coming from knowledge based system
2. If the result predicted is Normal
  - 2.1. Then display normal and do not recommend hospital
3. Else pass the predicted result to opinion summarizer
  - 3.1. accept the returned opinion summary with the weighted result of each hospital
  - 3.2. accept the user parameter input from interface
  - 3.3. pass the value to fact based recommender
  - 3.3. assign the returned weighted scored fact result
  - 3.4. hospital ranked based on result returned from 3.1 and 3.3
  - 3.5. then the result will be displayed to user interface
4. End

Experiment and Result Discussion The common standard of information retrieval (IR) effectiveness evaluation techniques are used in order to evaluate the developed prototype of the

system. Precision, recall and F-measure metrics are used in order to measure the retrieval effectiveness. Precision is a metric used to provide ratio of the number of relevant retrieved document to total number of document retrieved.

$$\text{Precision} = (\text{Number of relevant documents retrieved}) / (\text{Total Number of retrieved documents}) \dots \text{equation1}$$

Recalls is a metric used to provide ratio of the number of relevant retrieved document to total number of relevant document.

$$\text{Recall} = (\text{Number of relevant documents retrieved}) / (\text{Total Number of relevant documents in database}) \dots \text{equation2}$$

F-measure is another metric that considered both precision and recall this this metric often performed by harmonizing both precision and recall together. F-measure is defined as:

$$\text{F-measure} = (2 * \text{percision} * \text{recall}) / (\text{Percision} + \text{Recall}) \dots \text{equation3}$$

## Results

For sentiment classification evaluation experiment were conducted in two different experiment were performed on five hospitals that are found in Addis Ababa and currently serving public such as ALERT, Black Lion Hospital, Teklehaimanot General Hospital, Zewditu Memorial Hospital, St.Pauls Hospital .

Every system is developed to solve a given problem and to know the extent to which the system attain the required goal is going to be evaluated toward the real situation through effectiveness and efficiency of the system. Efficiency can be the amount of time taken by the system to produce a given output and effectiveness related to the extent to which the problem is solved.

## Discussion

Based on the experiment above Tekle Haimanot General Hospital has registered highest score of 0.95,0.93,0.94 precision, recall and F-measure respectively. in contrary to this less score was registered as 0.95,0.93,0.94 precision, recall and F-measure respectively by St.Pauls Hospital. The highest score doesnt mean the hospital is better than other in all aspect but, its about proportionality of ratio of people treated well divided by total number of treated patients.



<b>Hospital</b>	<b>Manually identified recommended</b>	<b>Identified by the system</b>	<b>Identified by the system correctly</b>	<b>Precision</b>	<b>Recall</b>	<b>F-Measure</b>
Black Lion Hospital	100	92	87	0.95	0.92	0.934
Tekle Haimanot General Hospital	100	93	88	0.95	0.93	0.94
ALERT Hospital	100	94	86	0.91	0.94	0.92
Zewditu Memorial Hospital	100	95	85	0.89	0.95	0.92
St.Pauls Hospital	100	93	78	0.84	0.93	0.89
<b>AVERAGE</b>				<b>0.9</b>	<b>0.93</b>	<b>0.92</b>

Table 4.16: Fact Based Recommendation: Exprimtent1 based on data collected for two months.

## Chapter 5

# INTEGRATING THE PREDICTIVE MODEL INTO KNOWLEDGE BASED SYSTEM

One of the objective of this study was integrating data mining of pre-cervical cancer result with knowledge based system for diagnosing whether a particular patient was affected by the disease or not. As described under the model below the development of system integration mainly composed of three parts: data mining parts using model that was selected after an experiment was carried out for the use of prediction of the cervical cancer disease, knowledge based system that was used to provide an inference engine services that will accept the user question and provide prediction and explanation and an integration part that was used to bridge the communication that was carried out between both data mining and knowledge based system; this is done by the help of JPL(Java Prolog Language) library, which is Bi-Directional.

Not only limited to this an Integration parts has the integrated rules that is common between both data mining and knowledge based system. Under these we have sub-parts that was carried out to collaborate with each other and acts as a single system for the user of the system. Generally, the detailed discussion on integrating data mining results with knowledge based system would be discussed through an architecture on figure 5.1 shown below.

### 5.0.1 Predictive Model

This phase, works for predicting the disease as cervical cancer positive or negative. Before, predicting the class there are many task that has to be performed. As we have already discussed in chapter three, the task of prediction starts from problem understanding and ends with prediction model that is ready for integration. Therefore, the following task were under taken to produce predictive model.

### 5.0.2 Problem Understanding

Problem understanding was seen as an important step were feasibility of particular problems in a given domain has carried out. As we tried to discuss in previous chapters, this is the first step when hybrid model is followed, understanding organizational/institutional (Hospital) problem is not a simple task. If the problem is not exposed correctly it is difficult to achieve.

Therefore, assessing the problem are considered as an important issues in data mining research. The task of problem understanding were carried out manually. At least for understanding particular business the prerequisite is smooth-communication between the researchers and an expert. Identifying, defining, understanding were a common trend that could be followed in order to formulate business domain problems. Additionally, various techniques were utilized in attaining the goal such as, interview, observation, document analysis and closed-ended questionnaires that were distributed across different workers in hospital horizontally and vertically based on the organizational structure.

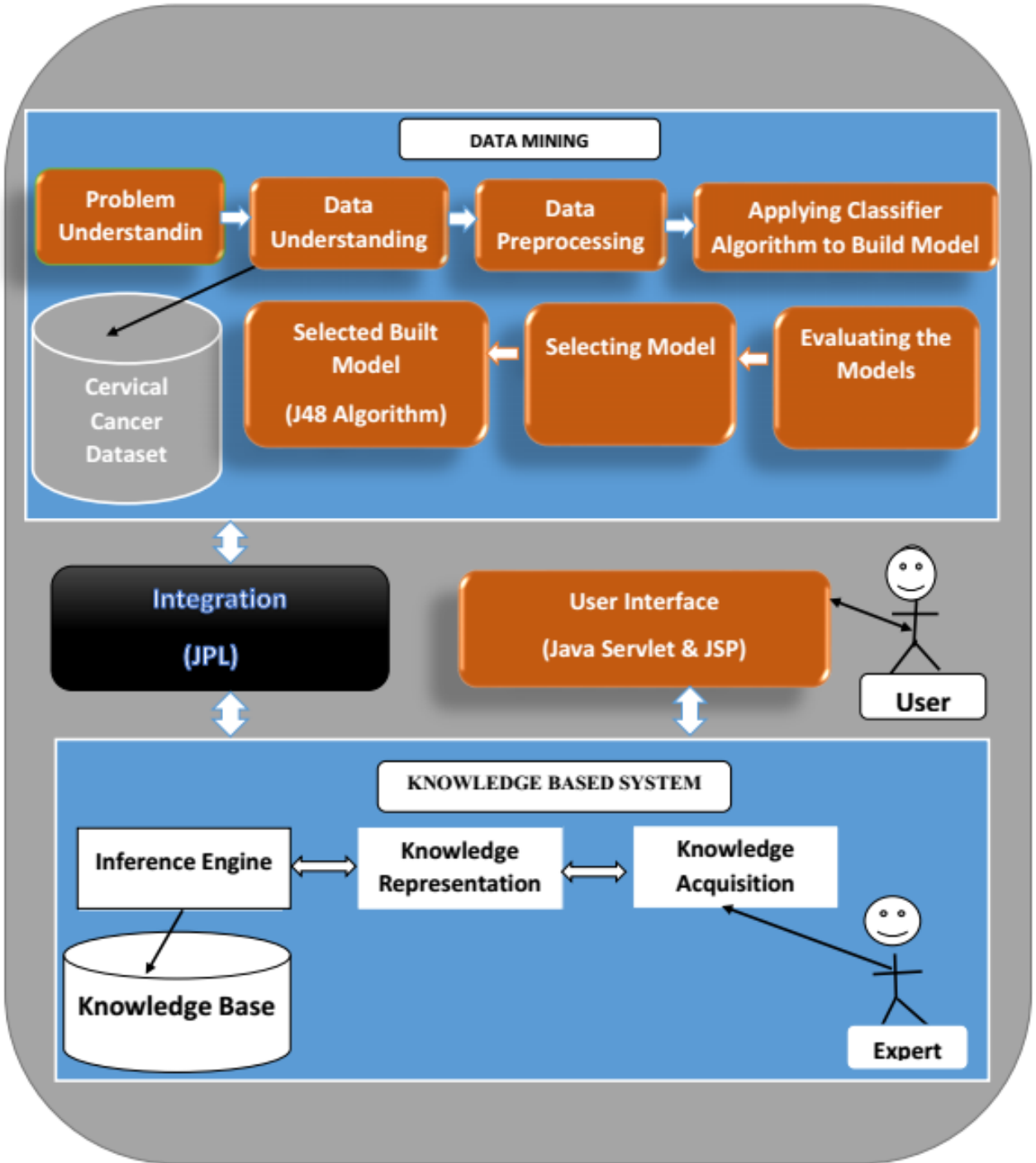


Figure 5.1: Architecture for Integration Data Mining Result with Knowledge based system

During this process many consult from domain expert was obtained regarding especially an organizational problems and legal issues with cancer and hospital recommendation that was placed by the Federal ministry of health(FMOH). Therefore, most of the operation carried out under this step was carried out manually.

### **5.0.3 DataSet**

Often, data-set could be in a two-dimensional spreadsheet format or a database table. The data set is used for predicting cervical cancer positive or negative. Since there is no currently established cancer center in Ethiopia data of cervical cancer were collected from five recognized sites: Black Lion Hospital, Teklehaimanot General Hospital, Zewditu Memorial Hospital, St.Paul's Hospital and ALERT hospital.

In these hospitals, some of the hospitals kept the patient data in excel format but other data was registered and placed in hard copy. The data that were registered in hard copy format was copied into excel format for further preparation and getting the data in form that is suitable for data mining tool like WEKA 3.6 and 3.7. After this phase has taken place the data that were copied from hard copy and already in excel format are mingled to one consistent excel format. And finally these initial data set are defined and viewed in Excel 2013 format to observe the properties of the dataset in gross. Some hospital register their data in software and others were not. Therefore, in this step data was obtained manually and automatically.

### **5.0.4 Data Understanding**

According to [19] model, the following step after business understanding is data understanding. The first task that has to be followed during data mining field of study was knowing of data itself. Therefore, the task of understanding data was made both manually under this step.

### **5.0.5 Data Preprocessing**

After data has understood correctly as much as possible the next step is data preprocessing. Data that were found in real-world might be less in quality. Using this data that is less in quality and feeding it to machine learning algorithm and expecting quality data is very difficult.

Therefore the work of getting quality with maximum effort was carried out for this study as discussed in previous section of chapter three. Detailed explanation on how preprocessing step was carried out was discussed there. These step taken place in manual and automatic manner.

### **5.0.6 Applying Classifier Algorithm to Build Model**

After the data were preprocessed and kept in such a way that machine learning algorithm can use the classifier algorithm were selected and applied to build a model.

### **5.0.7 Evaluating the Models**

After the model were built, an evaluation of a model has been carried out to select an algorithm that are capable of providing maximum output. Two algorithm that were commonly utilized under Decision Tree Classification Algorithm and Two rule based classifier algorithm. In order to obtain better result four classifier algorithms were applied and comparison has made using different evaluation parameters.

### **5.0.8 Selecting the Model**

After the model was evaluated using two algorithm that were commonly utilized under Decision Tree Classification Algorithm and two rule based classifier algorithm. The best fitting model (J48) was selected manually through inspection of the obtained result after an experiment has made.

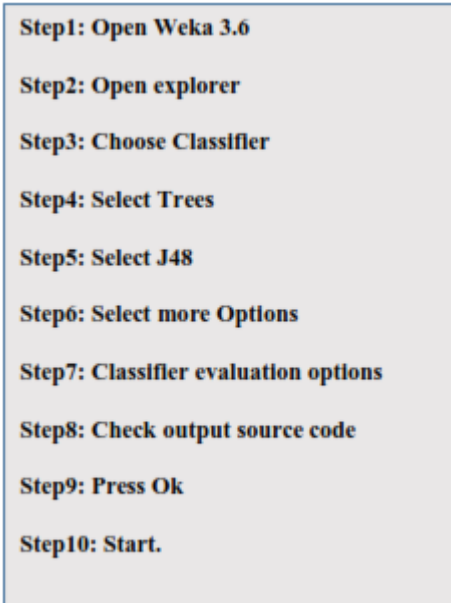
- 
- Step1: Open Weka 3.6**
  - Step2: Open explorer**
  - Step3: Choose Classifier**
  - Step4: Select Trees**
  - Step5: Select J48**
  - Step6: Select more Options**
  - Step7: Classifier evaluation options**
  - Step8: Check output source code**
  - Step9: Press Ok**
  - Step10: Start.**

Figure 5.2: Step to Extract selected model for Integration

### 5.0.9 Selected Built Model

After the model was selected then the selected model were integrated with knowledge based system through the use of JPL. Therefore, the communication coming from an interface passed through this built predictive model and interact with knowledge based system or vice verse. This was due support of JPL of Bi-directional communication.

### 5.0.10 Selected Built Model Extracted

After building the model it has to be extracted and added to java as jar file. The step for generating model source code disused as below with nine steps.

### 5.0.11 User Interface

In this study, for implementation of user interface, some programming and scripting language such as Java with UTF-8, JSP (Java Server Page), Servlet, Html, CSS, and JavaScript are employed. We choose this languages like java for two reason: the first one is based on their popularity and platform Independence. The second reason is we are very familiar with this types of languages than others. Creating interactive web Application GUI (Graphical User Interface) are very important for the user of the system. Moreover, tools such as, (Apache Tomcat 8.0.27.0 and Glass Fish Server 4.1.1), Netbeans IDE 8.1 with JDK 1.8.0\_45 are used for disease prediction and advice interface.

## 5.1 Knowledge Based System

Knowledge based system (KBS) is a discipline that has emerged from field of study known as Artificial Intelligence (AI). Knowledge based system was accomplished through the following steps listed below.

### 5.1.1 knowledge acquisition

For this study Knowledge acquisition techniques was carried out manually through interview (structured and unstructured), tracking methods and observation. The open discussion was carried out between researchers and domain expert; to obtain domain knowledge under cervical

```

static double N168ddf112(Object []i) {
    double p = Double.NaN;
    if (i[2] == null) {
        p = 1;
    } else if (i[2].equals("elementary")) {
        p = 1;
    } else if (i[2].equals("preparatory")) {
        p = 0;
    } else if (i[2].equals("diploma")) {
        p = 1;
    } else if (i[2].equals("ue")) {
        p = 1;
    } else if (i[2].equals("highschool")) {

```

Figure 5.3: Sample code to Extract selected model for Integration

cancer. An interview was also carried out to get tacit knowledge or to understand domain knowledge to some extent.

### 5.1.2 Knowledge Representation

Knowledge of cervical cancer management and treatment was represented from the decision tree after the diagnosing was predicted by data mining techniques after this process was done. The knowledge on treatment was developed based on the knowledge represented from automatically generated rules.

**Knowledge Base** It is represented knowledge in IF-Then format, while representing this, it was represented in manner it was consistent with the rule generated by J48 Algorithm of data mining.

**Inference Engine** Forward chaining technique was applied to determine the goal based on the data coming from predicted model with data mining technique (J48).

### 5.1.3 Integration

The main aim of integrating the system was to collaborate the system and works to act as a single system for the user who are using the system. This can be attained by calling the WEKA Jar file from Java then run the Java servlet code.

The selected J48 algorithm analysis result was revealed by generating the corresponding rule in IF-Then format. After this process can be carried out, the rules are converted into prolog understandable form after the data was obtained from GUI (Java Server Page).

```

public static void main(String args[]) throws Exception
{
    CSVLoader loader = new CSVLoader();
    loader.setSource(new File(sourcepath));
    Instances data = loader.getDataSet();
    ArffSaver saver = new ArffSaver();
    saver.setInstances(data);
    saver.setDestination(new File(destpath));
    saver.writeBatch();
}
Convert("E:\\cervical.csv", "E:\\cervical.arff");
}

```

Figure 5.4: Java code to transform CSV file to Arff file format

```

Output - JavaApplication27 (run)
adult,married,elementary,3,yes,no,no,no,yes,no,negative
young,married,degree,0,no,no,no,no,yes,yes,negative
young,married,diploma,2,yes,no,no,no,yes,yes,negative
adult,widow,elementary,2,no,no,no,no,yes,yes,negative
young,married,highschool,3,yes,no,no,no,yes,yes,negative
young,married,preparatory,3,no,no,no,no,yes,yes,negative
young,married,highschool,4,yes,no,no,no,yes,yes,negative
adult,widow,preparatory,2,yes,no,no,no,yes,yes,negative
young,divorced,preparatory,1,no,no,no,no,yes,yes,negative
young,widow,ue,2,yes,no,no,no,yes,yes,negative
adult,married,elementary,3,no,no,no,no,yes,yes,negative
young,single,preparatory,0,no,no,no,no,yes,yes,negative
adult,married,diploma,3,yes,no,yes,no,no,yes,postive
adult,married,elementary,3,yes,no,no,no,yes,yes,negative
young,married,diploma,1,no,no,no,no,yes,yes,negative
adult,single,ue,0,no,no,no,no,yes,yes,negative
young,married,preparatory,3,no,no,no,no,yes,yes,negative
young,married,preparatory,1,no,no,no,no,yes,no,negative
young,married,elementary,4,no,no,no,no,yes,yes,negative
young,single,ue,2,no,no,no,no,yes,yes,negative
young,married,highschool,0,no,no,no,no,yes,yes,negative
young,married,highschool,2,yes,no,no,no,yes,yes,negative
young,married,diploma,2,yes,no,no,no,yes,yes,negative
young,married,elementary,1,no,no,no,no,yes,yes,negative
young,single,elementary,2,yes,no,no,no,yes,yes,negative
BUILD SUCCESSFUL (total time: 7 seconds)

```

Figure 5.5: sample code read by Java

```

WekaClassifier Mod=new WekaClassifier ();
jpl_call( Mod, Predict_Disease, [age,maritalstatus,estatus,hsti,nob,sexstdate,VIC,sVIA],result)

```

Figure 5.6: .. CALLING PROLOG FROM JAVA

## 5.2 TESTING AND EVALUATION

One of the main objective of user acceptance testing is to make sure how well system prototype is performing on the eyes of users to assure that the system is acceptable and usable by them. To accomplish this testing was carried out by preparing questionnaires from SUS that have become an industry standard with references in over 600 publications [73], were used with small modifications. It had nine close ended questions. The weight value as suggested by Solomon [74] has been given as follows:

Excellent = 5, Very Good =4, Good =3, Fair =2 and Poor =1. Excellent=5, Very good =4, Good=3, Fair= 2 and Poor =1

No.	Evaluation Criteria	Poor	Fair	Good	Very Good	Excellent	Average	
1	The understandability of the system				2	3	4.6	
2	Attractiveness of the system				2	3	4.6	
3	The efficiency of the system				1	4	4.8	
4	Adequacy of the system in predicting disease					5	5	
5	Adequacy of the system in providing Advice			2	3		2.6	
6	The ability of the system in Recommending Health care					5	5	
7	The significance of the system for the patient					5	5	
8	The significance of the system for health care administrator					5	5	
9	The system accessibility				4	1	4.53	
		Average						4.6

Table 5.1: Summary of User Acceptance Testing.

As shown in the table above on the understandability of the system 60 % of an examiner scored as Excellent and 40 % as very good. on Attractiveness of the system an examiner put 60 % as Excellent and 40 % as very good. on the efficiency of the system an examiner put 80 % as Excellent and 20 % as very good. on the Adequacy of the system, the ability of the system in Recommending Health care in predicting disease, the significance of the system for the patient, the significance of the system for health care administrator an examiner put 100 % as Excellent. on the efficiency of the system an examiner put 80 % as excellent and 20 % as very good. on the Adequacy of the system in providing Advice an examiner put 75 % as Very good and 66 % as good. at the last, on the system accessibility an examiner put 75 % as Very good and 66 % as good.

Finally, the average performance of the prototype system according to the evaluation results filled by the domain experts is 4.53 out of 5 or 90.6 % which is a very good achievement.



## Chapter 6

# INTEGRATING OPINION SUMMARY AND FACT BASED INFORMATION WITH KNOWLEDGE BASED SYSTEM AND DATA MINING

### 6.1 General Architecture of Integration

The general architecture of the system that helps for disease prediction and hospital recommendation is illustrated on the Figure 6.1 below. Disease prediction and hospital recommendation system was developed by integrating fact based information and opinion summary with Data mining and knowledge based system. The architecture revealed from end to end various steps that were used for disease prediction and hospital recommendation. These architecture can be discussed in various steps as below.

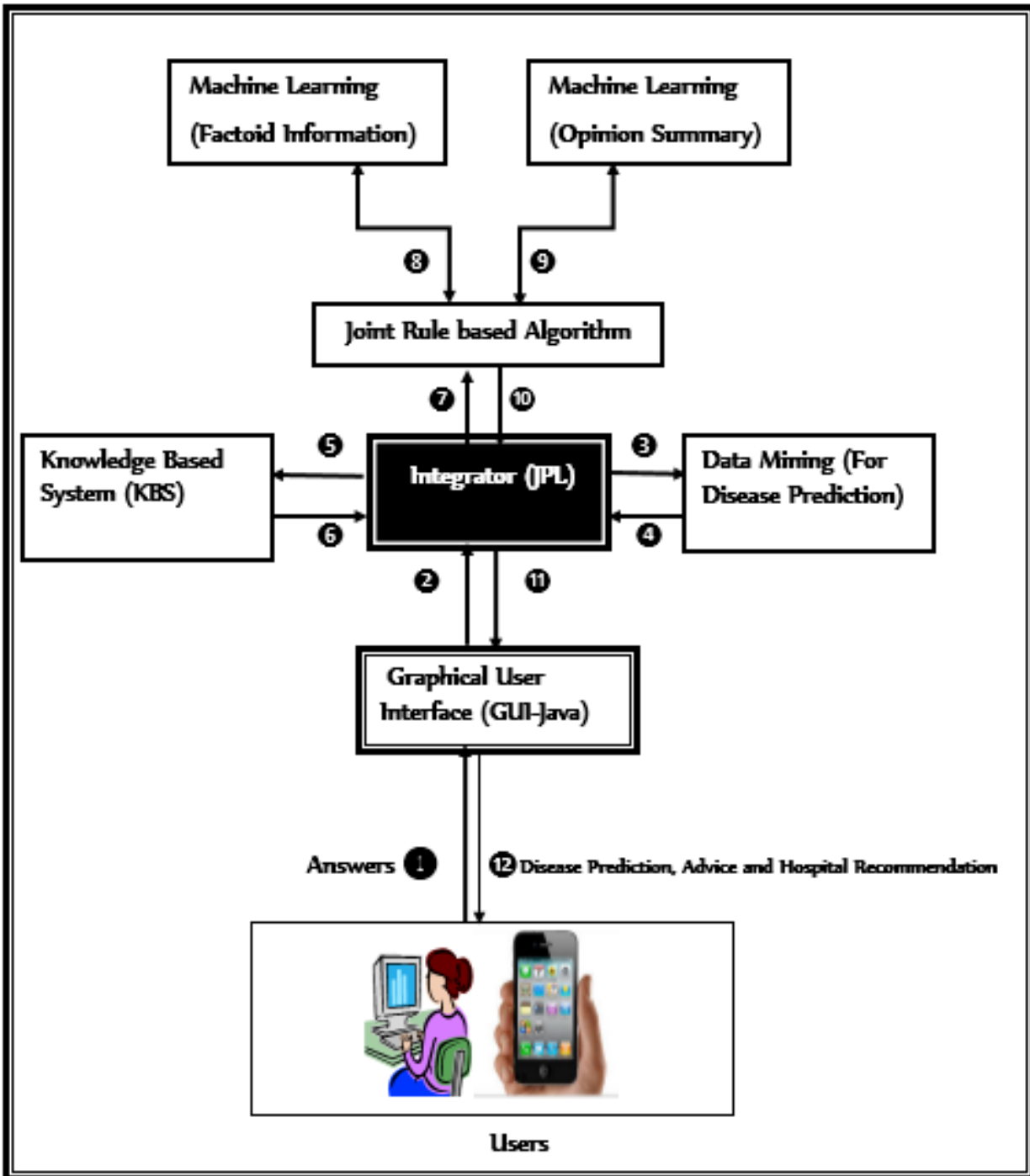


Figure 6.1 General Architecture for Integration of Fact Based Information and Opinion Summary model with Knowledge Based system and Data Mining

At first step, the user begins to interact with the system by browsing the web with domain name, then the Java Server Page (JSP) of the requested page would be displayed for the user to select the language that is familiar with the user. The second step, the user will select for the language that is familiar with them, after selecting the language the question for cervical cancer diagnosis will be displayed, the user can interact with the system by filling the question that would be asked.

After filling the questions, the user would press the predict button then an answer passed to java servlet then the java servlet will pass an answer to data mining models that were built using J48 algorithm. These, algorithm will predict the result by displaying as 1.0 or 0.0. Since, the user do not understand 1.0 and 0.0 it should be displayed in a way that is understandable by user. Therefore, these format could be converted into user understandable format to make

the system usable by passing the predicted result that was placed as 1.0 or 0.0 to knowledge based system (KBS), therefore it can convert not user understandable format of result into user understandable format of “Cancer” or “Normal”. But, the communication between data mining and knowledge based system model was carried out by the help of bidirectional JPL.

JPL is a library using the SWI-Prolog foreign interface and the Java jni interface providing a bidirectional interface between Java and Prolog that can be used to embed Prolog in Java as well as for embedding Java in Prolog[31][32]. After, the result was converted through JPL by the prolog programming it will be provided for Joint Rule Based Algorithm, for this case the Joint Rule based algorithms were developed for the task of accepting and passing the Cancer result produced by the knowledge based system using prolog for recommendation of the “Cancer” result.

Recommendation of hospital do not works for user who were normal but, only prediction will be displayed for them. This model can also works if the data of the real time can be given or feed to the system; its scalable.

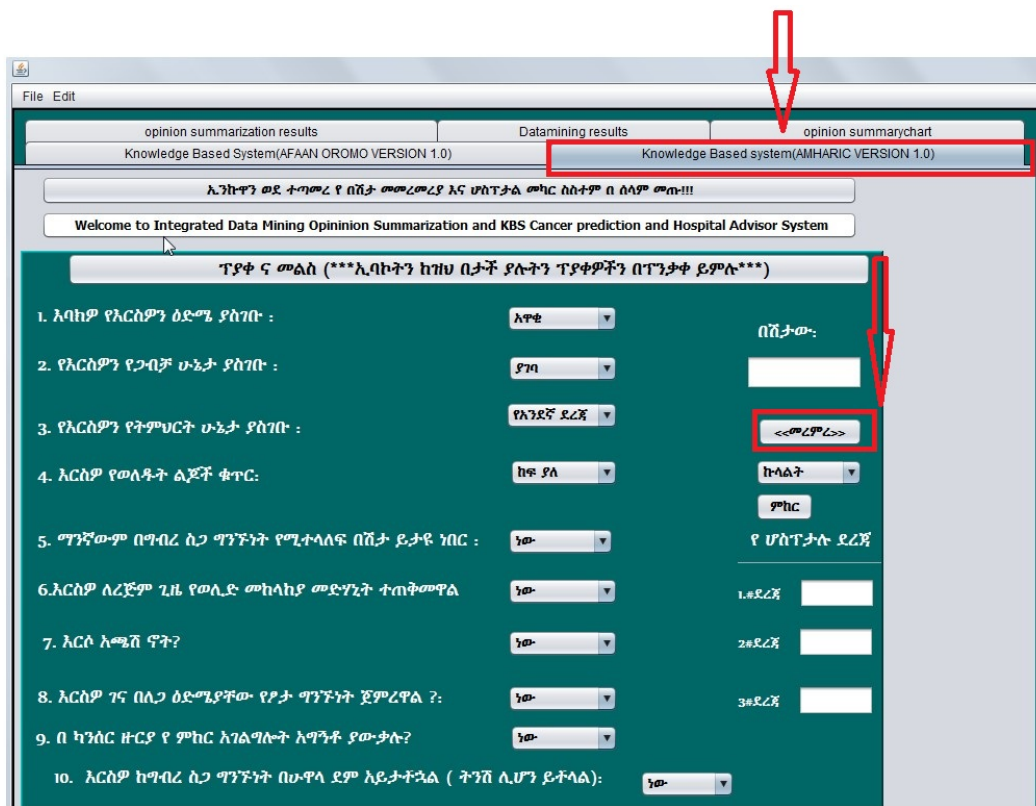


Figure 6.2 Prototype of Integrated of Fact Based Information and Opinion Summary model with Knowledge Based system and Data Mining

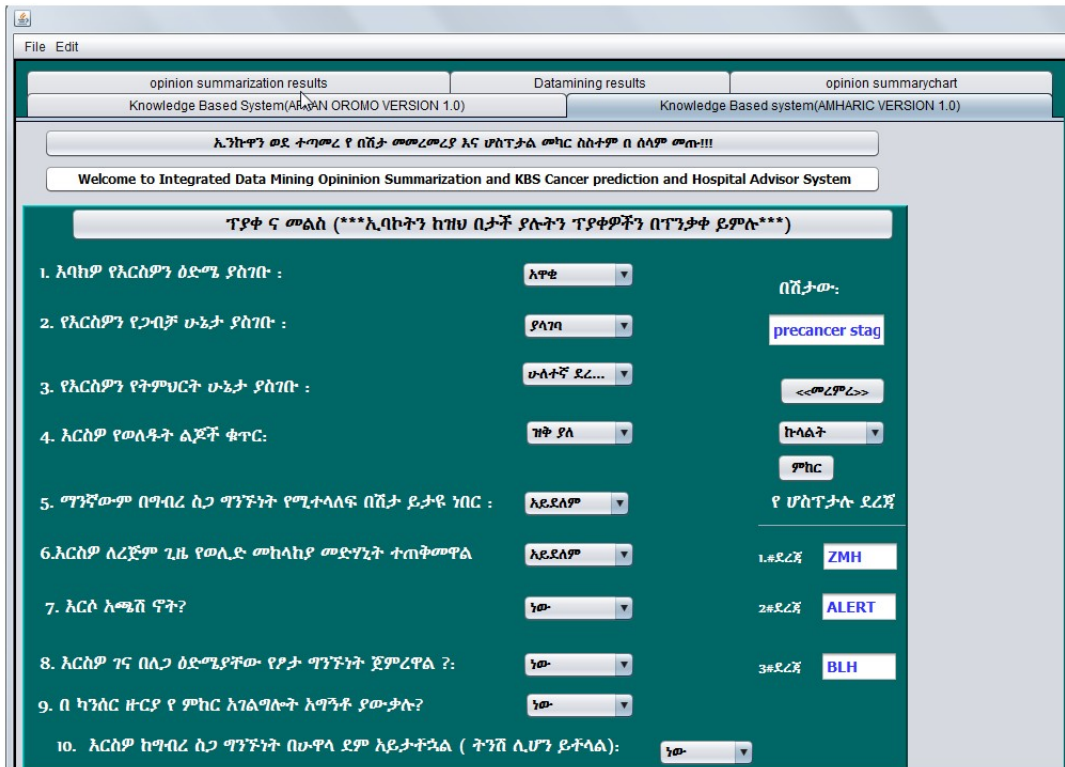


Figure 6.3 Prototype of Integrated of Fact Based Information and Opinion Summary model with Knowledge Based system and Data Mining(After Processing)

## Chapter 7

# RESULT AND DISCUSSION

### 7.1 Discussion

As already discussed, the result registered by the system showed less performance for both fact based recommendation and opinion based recommendation. When the comparison has made between the Fact based recommendation and opinion based recommendation for a given hospital the result were almost similar. An evaluation has been carried out manually through test plan the results that were obtained from fact based information were almost similar with opinion based information to recommend particular hospitals.

This indicates if 10 hospitals recommended by both recommender system they are different in only three hospitals and they are the same for six particular hospital. This result showed that the result obtained by both fact based recommendation and opinion based recommendation are almost similar. This indicate that most of the opinion that were provided by the customers (patients) was based on the fact that were available on the ground.

result was decreased was less because of its collected time different; since opinion were collected in current years while fact based data was collected starting from 2010. If an opinion is available starting from 2010, it will almost provide similar result. Finally, we conclude that this variation shows, there were some fact that was changed. The fact can be a number of specialist, the service delivery system, an experience of specialist. During selecting better algorithm for disease prediction and fact based recommendation we have selected the same algorithm since, the result were the same. This happens due to a few column variation that appears between disease prediction dataset and fact based information dataset.

The developed system has achieved the overall performance of 80.5 %, even though the system had performed better results there are some challenges that disallow the system to register best performance than the current performance. Those challenges under opinion summarization were discussed as follows: The first reason was since Afaan Oromo and Amharic language were among the less resourced language with less IR (Information Retrieval) and NLP (Natural Language Processing) tools that are found online compared to high resourced English language. These tools can be a Parser, Pos-Tagger, Word sense Disambiguation tool, Semantic Analyzer, Grammar checker, sentiment classifier, SentiWordnet, Amharic and Afaan Oromo Wikipedia and statistical machine translator. Therefore, in the absence of the tools discussed above the development of opinion summarization system is very difficult.

The second reason was the language nature of Afaan Oromo and Amharic language. Since, most of the opinion that were provided by the users (patient) are very local to the people; it differs from people to people. Most people express opinion through proverbs, pragmatics and idiomatic.

The third reason was the absence of Amharic and Afaan Oromo opinionated corpus that were collected, preprocessed and ready for machine learning algorithm, especially for hospital domain. The fourth reason was aspect extraction was very difficult, since patients do not provide opinion by touching aspect like an expert. Occasionally, opinion forwarded from patients were not written in such a way that its extracted by machine learning algorithm. Nave users were not in a position to provide an opinion in such manner.

## Chapter 8

# CONCLUSION AND RECOMMENDATIONS

### 8.1 CONCLUSION

Health care is one of the business organization that use customer satisfaction as vital tool of evaluating to what extent the service is smoothly being delivered to the customers (patients)[1][2][3]. It can be utilized as an instrument that indicate the weak and strong side of the organizational aspect . Getting people who is enough to tell you the right hospital for the right disease is very difficult in this many alternatives. One is not enough to say hospital 'X' is better than hospital 'Y' due to alternative available and lack of skills of knowing all the hospitals aspect. This is due to privacy issues that was regulated in the health care circle. Moreover, it's due to the dynamic nature of the health information that was available previously and right now.The challenge is, no one has real time information about every alternatives aspects of dynamic health care service available for the users (patients). The healthcare professionals and customers (patients) of the health care domain are not proportional in Ethiopia[19][16].Solving the above told problem requires the collaboration of various elements. Currently, diagnosing the patient and telling the right hospital, requires a collaboration of various system elements together.

Hence in this study an effort has been made to develop disease prediction and hospital recommendation system by using an integrated fact based information and opinion mining with knowledge based system and data mining.Cancer is one of the disease that requires early detection for treatment to prevent. To predict cervical cancer at its early stage data mining was applied, to make the system usable knowledge based system was integrated with data mining. For recommendation of health care service,an integrated approach of both fact and opinion based information was used. Then, the developed model can able to recommend suitable health care services based on predicted disease. The developed model was developed and applied on five hospitals data set. The recommender system is scalable and works for more than one disease if the EDPHR Model are supplied with adequate and proper data. The developed EDPHR Model has strong side of helping the health care organization by telling the weak side and strong side of the various health care aspects.Moreover, it can also help both patient who already know the disease and the new patients .for the patient who already know the disease it will recommend the best choice hospitals. for new patient it will detect the disease at its early stage and recommend the best choice among the hospitals. The developed system can minimize the disproportion between the patients and professionals.

Generally, efficiency of opinion summarization depend on the frequency of summarization. The more input is summarized, the more efficient and effective the algorithm will be. Once the vector space model has been made between particular opinion terms and equivalent opinion, no need of doing the same thing again another time unseen words occurs, instead an opinion term will be added to the opinion corpus and then vector space model can be made between unseen opinion terms and opinion corpus. Therefore, for the second time, time to calculate vector

space model between opinion terms and equivalent opinion or opinion corpus can be reduced, since the word had been added to the opinion database already at the first summary. hence, the system can annotate the corpus by itself.

The effectiveness of the system was done using adjective/verb equivalent of part of speech tagged sentences. The use of vector space model with n-gram has also played a great role for word sense disambiguation. Since, the opinion terms has different opinion in different context for example, the term “long” has different meaning in different context. “The employees took long time to serve a patient” and “the doctors took long time for justification”.

The developed models was less efficient for Afaan Oromo language and Amharic language as compared to resourced English language, this is due to time it takes for translation of these local language to English Language if the word is not found in equivalent opinion corpus. Although, the developed model was less efficient, the effectiveness of the system was high. The effectiveness of the system is high, since, the model developed can use the equivalent opinion corpus and English resource (i.e. English POS-Tagger).

aspect detection was carried out using annotator developed using naive bayes classifier and good result was performed. but, the detection of aspect don't work if more than one aspect is available in a given review sentence.

Lastly, as we understand from an experiment carried out in Amharic and English version St.Paul's hospitals registered less score in every experiment. Therefore, such types of output cannot be considered as patient language complexity problem but there is some fundamental problem that should be addressed by the health care.

Generally, the system achieves good performance and meets the objectives of the study. However, in order to make the system applicable in health care domain collaboration of various health professionals with quality data is very important.



## 8.2 Recommendation

Even though, the system has achieved good results in this study, the following recommendations can enhance the performance of the system for the future study:

- Though the system has made an attempt in order to predict and recommend hospital at its early stage, the development of full-fledged system, which is efficient and effective in predicting and recommending hospital at its early stage requires the involvement and coordination of different domain expert from different place. For prediction and suggestion of cancer disease, strong coordination are expected from the medical expert and to fill the gap that were seen by experts researchers are also expected more. For recommendation of hospital that includes the task of recommendation using facts the domain experts and other hospital personnel coordination requires. And for the task of recommendation using opinion, we need coordination of linguistic expert in the language domain especially, Amharic and Afaan Oromo linguistic experts. Furthermore, we need patient comments regarding the developed system to help the system more efficient and effective.
- The system developed can make early diagnosis for the patient infected with only one disease (i.e. cervical cancer), but in the future we planned to enhance model that support early diagnosis for the patient infected with more than one disease (for example, patient with cancer and hypertension).
- The developed system can works only for medical domain, but in the future we planned to extend system to develop model for other domain such as university, hotels and banks. For university, we planned to develop system that can helps student by recommending best university that provides quality education in particular field, by predicting the field of study that can best fit the student using the profiles of the student and senior students who have already joined the same department. The future study can also help those student who needs early recommendation of the University for Particular Field of study.
- The system developed can make early diagnosis for only one disease cervical cancer. But, the future works will include other disease that requires early diagnosis.
- The developed system can works only for five hospitals that were currently providing service and treatment for cervical cancer disease type, but further study will include other hospital that provides treatment for the same disease.
- The system developed are designed to support the patients that speaks at least one of three languages such as English, Afaan Oromo and Amharic. Including all other Ethiopian local languages has seen as future research direction.
- The system developed were not integrated with real time patient records, therefore, in the future the researcher has also aimed to integrate the developed system with real time patient records.
- The developed system doesn't work for visually impaired users, therefore, the researcher aimed to integrate the system with speech based factoid question answering system for the future study.
- Integrating fact based information and opinion summary of hospital directly with Weka tool than already built disease model can increase the efficiency of the system.
- Since, the system developed doesn't consider illiteracy level of patient, including human computer interaction concept to the system will increase usability by the end-user increase the effectiveness of the system.
- In this study, the system developed can obtain only knowledge from the machine learning outcome. However, if other mechanisms like domain expert knowledge are incorporated, the system accuracy and adequacy could increase.

- The developed system can provide advice based on incorporated static tacit knowledge, however, context based advice that consider the current situation of patient were not included. Therefore, including this can create new dimension for the study.
- The developed system can provide future like early disease prediction and hospital recommendation. But, it doesnt include drug recommendation. Therefore, including drug recommendation can enhance the system capability.
- The developed system can recommend hospital only by using the predicted disease of the patient, therefore, recommendation that enhance the patient background information such as patient economy level, patient education level will be more helpful to make the system effective.
- The system can able to show better result for opinion summarization if it use statistical method of translation using decoder such as GIZA++ ,PHARAOH, for both Amharic and Afaan Oromo language.
- The developed model perform an accuracy of 90.6 %,this shows that it doesnt work for almost 10 % of the patient or it provides wrong result for the patient who seek decision. Therefore, in the upcoming we planned to rise the accuracy of the result.
- The developed system cant handle fake input data, for example, if the hospitals enter false data the model developed can recommend hospital with false data, therefore in the future we planned to add a mechanism to handle such types of problem that leads to biased decision.
- Real time recommendation of hospital if the system is directly connected with real time data of each hospitals.
- Treatment material recommender system based on the patient screening result has also seen as one of the future research direction for these study.
- The study can tries to predict cervical cancer as generalized positive or negative and Including other detail available status and advice were intended for the future. Therefore, Further studies would be carried out to apply all available status and advice should be provided especially for cancer positive patients.

# Bibliography

- [1] T. R. WEBSTER et al., “A brief questionnaire for assessing patient healthcare experiences in low-income settings”,*International Journal for Quality in Health Care* , vol. 23, no. 3, pp.1-5, April. 2011.
- [2] B.Prakash, “Patient Satisfaction,” *journal of cutaneous and aesthetic surgery*, vol. 3, no. 3, pp.1-2, Sep-Dec 2010.
- [3] N.Shiferaw et al., “The Single-Visit Approach as a Cervical Cancer Prevention Strategy among Women with HIV in Ethiopia: Successes and Lessons Learned,” *Glob Health Sci Proj* ,vol. 4, no. 1, pp.1-4, Mar 2016.
- [4] P.Ahmad et al., “Techniques of Data Mining In Healthcare: A Review,” *International Journal of Computer Applications*, vol. 120, no. 15, pp.1-6, June, 2015.
- [5] M. Kantardzic, “Data Mining Concepts, Model, Methods and Algorithms”, *New Jersey, USA:John Wiley and Sons Publication Inc*, 2011.
- [6] B. Liu, “Sentiment Analysis and Opinion Mining”, *Morgan and Claypool Publishers*, April 22, 2012.
- [7] L.Zhang and B.Liu.,Aspect and Entity Extraction for Opinion Mining,” pp. 14.
- [8] M. Ghosh and A. Kar, “Unsupervised Linguistic Approach for Sentiment Classification from Online Reviews Using Sentiwordnet 3.0,” vol. 2, no. 9, pp. 5560, 2013.
- [9] M. Elarnaoty, S. Abdelrahman, and A. Fahmy, “A MACHINE LEARNING APPROACH FOR OPINION,” vol. 3, no. 2, pp. 14, 2012.
- [10] B. Pang and L. Lee, “Opinion Mining and Sentiment Analysis,” *Foundations and Trends in Information Retrieval*, vol. 2, no. 12. pp. 15, 2008.
- [11] B.Liu and L.Zhang, “A SURVEY OF OPINION MINING AND SENTIMENT ANALYSIS ”,MINING TEXT DATA pp. 240.
- [12] K. Erk, “Vector Space Models of Word Meaning and Phrase Meaning?: A Survey,” *Language and Linguistics Compass*,vol.6 , no.10, pp. 15, 2012.
- [13] N.Kurian and S.Asokan, “Summarizing User Opinions?: A Method for Labeled-Data Scarce Product Domains,” *International Conference on Information and Communication Technologies (ICICT 2014)*, vol. 46, pp. 35, 2015.
- [14] S. P. and Z. c. Kim D.,Ganesan K., “Comprehensive Review of Opinion Summarization,” pp. 130.
- [15] L. Raut V., “Survey on Opinion Mining and Summarization of User Reviews on Web,” *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 2, pp. 35, 2014.
- [16] Mulugeta T. and Beshah, “International Journal of Advanced Research in Computer Science and Software Engineering Integrating Data Mining Results with the Knowledge Based System for Diagnosis and Treatment of Visceral Leishmaniasis,” vol. 5, no. 5, pp. 14, 2015.
- [17] A.Mohammed, Towards Integrating Data Mining with Knowledge Based System: The Case of Network Intrusion detection,” M.Sc thesis ADDIS ABABA UNIVERSITY, ADDIS ABABA, June , 2013.
- [18] Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy: *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press 1996
- [19] S. GEBREMARIAM, A SELF-LEARNING KNOWLEDGE BASED SYSTEM FOR DIAGNOSIS AND TREATMENT OF DIABETES,” M.Sc thesis ADDIS ABABA UNIVERSITY, ADDIS ABABA, JANUARY 2013
- [20] G. Violetta, O. Mieczyslaw and Richard F., “On Features of Decision Trees as a Technique

- of Knowledge Modeling,” in Proceedings of the Workshop on Computer Science and [21] S. Russell and P. Norvig, “Artificial Intelligence: A Modern Approach, Prentice-Hall”, 1995.
- [22] R. Akerkar and P.S. Sajja, “Knowledge-Based Systems”, USA: *Jones and Bartlett Publishers*,2010.
- [23] Migbaru S., “Cervical Cancer?: Open Access Trends of Cervical Cancer in Ethiopia,” vol. 1, no. 1, pp. 14, 2016.
- [24] M. Taboada et al., Lexicon-Based Methods for Sentiment Analysis, Association for Computational Linguistics, pp. 15, 2011.
- [25] Bashetha. A and G.U.Srikanth et al., “Effective Cancer Detection Using Soft Computing Technique”, *Journal of Computer Engineering (IOSR-JCE)*, vol. 17, no. 1, pp.3-5, Jan Feb. 2015.
- [26] T. Yesuf, “Survival and associated factors among cervical cancer patients in BlackLion Hospital, Addis Ababa, Ethiopia, 2008-2012, a retrospective longitudinal study,” Msc. thesis, Addis Ababa University, COLLEGE OF HEALTH SCIENCES, Addis Ababa, February, 2014.
- [28] B. Pang and L. Lee, Opinion mining and sentiment analysis, Jan. 2008, vol.2, no.5, pp. 19, IEEE
- [29] F.Assefa et al., “Assessment of Clients’ Satisfaction with Health Service Deliveries at Jimma University Specialized Hospital”, *Ethiop J Health Sci* ,2011, vol. 21, no. 2, pp.1-5, Jul. 2011.
- [31] J. Wielemaker, SWI-Prolog 5.6 Reference Manual, Human-Computer Studies (HCS, formerly SWI), Eds. Kruislaan 419, 1098 VA Amsterdam the Netherlands, 2008, pp-1-6.
- [32] T.Ali et al., JPL : IMPLEMENTATION OF A PROLOG SYSTEM SUPPORTING INCREMENTAL TABULATION, Computer Science and Information Technology (CS IT), vol. 1, no. 1, pp.1-9, 2016.
- [33] G. Nigussie, “Developing an Integrated Knowledge-Based Crop Irrigation System with Soil Moisture Content Prediction,” M.Sc thesis, Debre Birhan University, Debre Birhan, June 2014.
- [30] D.Ankitkumar et al., “A Survey on Sentiment Analysis and Opinion,” *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 2, no. 11, pp. 24, 2014.
- [35] A.Dhocart et al., “Review on Techniques and Tools used for Opinion Mining,” International Journal of Computer Applications Technology and Research(IJCATR), vol. 4, no. 6, pp. 419424, 2015.
- [38] P.S. Sajja and R. Akerkar, “Advanced Knowledge Based Systems: Model, Applications and Research,” 2010.
- [47] F.Narducci et al., A Recommender System for Connecting Patients to the Right Doctors in the HealthNet Social Network,” *WWW 2015 companion* , pp.1-2, Florence Italy, May 2015.
- [27] X. Wan, “Using Bilingual Knowledge and Ensemble Techniques for Unsupervised Chinese Sentiment Analysis,” pp. 56, 2008.
- [28] B. Pang and L. Lee, Opinion mining and sentiment analysis Jan. 2008, vol.2, no.5, pp. 19, IEEE
- [29] F.Assefa et al., “Assessment of Clients’ Satisfaction with Health Service Deliveries at Jimma University Specialized Hospital”, *Ethiop J Health Sci* ,2011, vol. 21, no. 2, pp.1-5, Jul. 2011.
- [30] D.Ankitkumar et al., “A Survey on Sentiment Analysis and Opinion,” *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 2, no. 11, pp. 24, 2014.
- [31] J. Wielemaker, SWI-Prolog 5.6 Reference Manual, Human-Computer Studies (HCS, formerly SWI), Eds. Kruislaan 419, 1098 VA Amsterdam the Netherlands, 2008, pp-1-6.
- [32] T.Ali et al., JPL : IMPLEMENTATION OF A PROLOG SYSTEM SUPPORTING INCREMENTAL TABULATION, Computer Science and Information Technology (CS IT), vol. 1, no. 1, pp.1-9, 2016. [33] G. Nigussie, “Developing an Integrated Knowledge-Based Crop Irrigation System with Soil Moisture Content Prediction,” M.Sc thesis, Debre Birhan University, Debre Birhan, June 2014.
- [34] L.Rokach and O. Maimon, “Data Mining With Decision Trees Theory and Applications”, *Tel Aviv, Israel: World Scientific Publishing*, 2015.
- [36] A. Buche, M. B. Chandak, and et al., “OPINION MINING AND ANALYSIS?: A SURVEY,” International Journal on Natural Language Computing (IJNLC), vol. 2, no. 3, pp. 58, 2013.
- [37] Sajja P. and Akerkar R., Knowledge-Based Systems for Development, vol. 1. 2010.

- [38] P.S. Sajja and R. Akerkar, "Advanced Knowledge Based Systems: Model, Applications and Research," 2010.
- [39] A.Hans, Schreiber A., Anjewierden R., deHoog N., Shadbolt W. V., deVelde, "Knowledge Engineering and Management: The Common KADS Methodology," in MIT Press, Massachusetts, 1999.
- [40] Priti Sajja Rajendra Akerkar, "Knowledge-Based Systems", (2013, Feb 18). [online]
- [41] Polettini N., "The Vector Space Model in Information Retrieval - Term Weighting Problem Local Term-Weighting," pp. 14, 2004.
- [42] B. Liu, "Sentiment Analysis," 5th Text Analytics Summit Boston, June 1-2 2009.
- [43] S. Poobana and S. Rekha, "Opinion Mining From Text Reviews Using Machine Learning Algorithm," vol. 3, no. 3, pp. 17, 2015.
- [44] Lenjisa D., "The Status of Afan Oromo as Indigenous Language Program in Ethiopian Higher Education Institutions?: An Exploratory Study of Three Selected Universities," vol. 4, no. 4, pp. 24, 2015.
- [45] X. Xie, "Classification Rule Induction with Ant Colony Optimization Algorithm," M.Sc thesis, University of Texas, Texas, USA, 2004.
- [46] L. Augustyniak et al., "Simpler is Better?? Lexicon-based Ensemble Sentiment Classification Beats Supervised Methods," International Workshop on Curbing Collusive Cyber-gossips in Social Networks (C3-2014), pp. 25, August 2014.
- [47] Narducci et al., "A Recommender System for Connecting Patients to the Right Doctors in the HealthNet Social Network," pp. 20132014, 2015.
- [48] E. Marrese-Taylor, J. D. Vel'squez, and F. Bravo-Marquez, "A novel deterministic approach for aspect-based opinion mining in tourism products reviews," *Expert Syst. Appl.*, vol. 41, no. 17, pp. 47, 2014.
- [50] J. Zhao, K. Liu, and G. Wang, "Adding Redundant Features for CRFs-based Sentence Sentiment Classification," pp. 16, 2008.
- [51] W. Philemon, "A Machine Learning Approach to Multi-Scale Sentiment Analysis of Amharic Online Posts," vol. 2, no. 2, pp. 14.
- [52] P. Palanisamy, V. Yadav, and H. Elchuri, "Serendio?: Simple and Practical lexicon based approach to Sentiment Analysis," *Second Joint Conference on Lexical and Computational Semantics (\*SEM)*, vol. 2.7, pp. 15, 2012.
- [53] F. Wogenstein, J. Drescher, D. Reinel, and D. Reinel, "Evaluation of an Algorithm for Aspect-Based Opinion Mining Using a Lexicon-Based Approach," 2013.
- [54] A. S. Kamale, "A Survey on Feature-Sentiment Classification Techniques," vol. 2, no. 12, pp. 17, 2014.
- [55] A. Mohan, R. Manisha, B. Vijayaa, and J. Naren, "An Approach to Perform Aspect level Sentiment Analysis on Customer Reviews using Sentiscore Algorithm and Priority Based Classification," vol. 5, no. 3, pp. 48, 2014.
- [56] C. Sivagami and S. C. Punitha, "Ontology Based Sentiment Clustering Of Movie Review," vol. 2, no. 4, pp. 38, 2013. [57] N. M. Shelke, "Survey of Techniques for Opinion Mining," vol. 57, no. 13, pp. 35, 2012.
- [58] L. A. De Freitas, P. Alegre, and R. Vieira, "Ontology-based Feature Level Opinion Mining for Portuguese Reviews," pp. 13, 2013.
- [59] V. S. Jagtap and K. Pawar, "Analysis of different approaches to Sentence-Level Sentiment Classification," vol. 2, no. 3, pp. 16, 2013.
- [60] Akkermans, Hans G., Schreiber A., Anjewierden R., deHoog N., Shadbolt W. V., deVelde and B. Wielinga, "Knowledge Engineering and Management: The Common KADS Methodology," in MIT Press, Massachusetts, 1999.
- [61] G. Violetta, O. Mieczyslaw and Richard F., "On Features of Decision Trees as a Technique of Knowledge Modeling," in Proceedings of the Workshop on Computer Science and Information Technologies CSIT99, Moscow, Russia, 1999
- [62] TILAHUN T., "OPINION MINING FROM AMHARIC BLOG," 2013.
- [63] Migbaru S., "Cervical Cancer?: Open Access Trends of Cervical Cancer in Ethiopia," vol. 1, no. 1, pp. 14, 2016.

- [64] Lenjisa D., “The Status of Afan Oromo as Indigenous Language Program in Ethiopian Higher Education Institutions?: An Exploratory Study of Three Selected Universities,” vol. 4, no. 4, pp. 24, 2015.
- [65] Polettini N., “The Vector Space Model in Information Retrieval - Term Weighting Problem Local Term-Weighting,” pp. 14, 2004.
- [66] Hiemstra D., Using language models for information retrieval. 2001.
- [67] Ahmad K., “Review Mining for Feature Based Opinion Summarization and Visualization,” pp. 16.
- [68] Thangavel et al., “Data Mining Approach to Cervical Cancer Patients Analysis Using Clustering Techniques,” Asian J. Inf. Technol., vol. 5, no. 4, 2006.
- [69] Abdosh B., “The quality of hospital services in eastern Ethiopia?: Patients perspective.”
- [70] S.K.Tadesse, “Preventive Mechanisms and Treatment of Cervical Cancer in Ethiopia,” *Cervical Cancer: Open Access*, vol. 1, no. 1, pp.1-6, 2016.
- [71] T.Legessee, “Factors Affecting the Practices of Cervical Cancer Screening among Female Nurses at Public Health Institutions in Mekelle Town,” *Journal of Cancer Research*, vol. 1, no. 1, pp.1-6, Feb, 2016.
- [72] J. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publish, 2001

## 8.3 APPENDICES

### 8.3.1 Appendix I: Interview Questions to Obtain Tacit Knowledge from Expert

The main objectives of this interview is to obtain tacit knowledge from experts engaged in cancer specialization. This interview helps to develop knowledge based system, which is one part of the study. After introducing the objective of the study, respondents were asked to take part by responding the interview questions. The answers of the respondents were recorded by using pen, paper, and mobile phone for the following interview questions. The following interview questions are the main area to cover how specialists diagnose a patient. Thank you for your support and useful information.

1. What are the common cancer in Ethiopia?
2. What is the mortality and morbidity rate of cancer in Ethiopia?
3. What do you think about the major cause of cancer in Ethiopia?
4. How can we differentiate various types of cancer?
5. What are the main challenges that encounter during diagnosis of cancer?
6. What are the mechanism to control different types of cancer?

### 8.3.2 Appendix II: Policy Related Interview Questions:

The main objectives of this interview is to obtain tacit knowledge from experts engaged in cancer specialization. This interview helps to develop knowledge based system, which is one part of the study. After introducing the objective of the study, respondents were asked to take part by responding the interview questions. The answers of the respondents were recorded by using pen, paper, and mobile phone for the following interview questions. The following interview questions are the main area to cover how specialists diagnose a patient. Thank you for your support and useful information.

1. What are the control and prevention effectiveness of Ethiopian health policy towards chronic disease including cancer?
  - Is there any special treatment procedures to cervical cancer?
  - Is there any special cancer related health centers in the country?
2. To what extent resources (i.e. material, human and financial) is available in the country to overcome cancer related problems?
3. How far community based cancer related awareness and knowledge transfer issues are worked?
4. What are the criteria MOH (Ethiopian Ministry of Health considers) has considered to rank Ethiopian hospitals at the national level?

### 8.3.3 Appendix III: Sample Cervical cancer Dataset

	A	B	C	D	E	F	G	H	I	J	K	L
1	<b>age</b>	<b>mstatus</b>	<b>estatus</b>	<b>nob</b>	<b>hsti</b>	<b>ESS</b>	<b>CS</b>	<b>CM</b>	<b>VIAC</b>	<b>BAS</b>	<b>BP</b>	<b>Scresult</b>
2	adult	married	elementary	high	no	no	no	no	yes	no	no	postive
3	young	single	preparatory	low	no	no	no	no	yes	no	no	postive
4	adult	married	diploma	high	yes	no	no	no	no	no	no	postive
5	adult	married	elementary	high	yes	no	no	no	yes	no	no	postive
6	young	married	diploma	low	no	no	no	no	yes	no	no	postive
7	adult	single	ue	low	no	no	no	no	yes	no	no	postive
8	young	married	preparatory	high	no	no	no	no	yes	no	no	postive
9	young	married	preparatory	low	no	no	no	no	yes	no	no	postive
10	young	married	elementary	high	no	no	no	no	yes	no	no	postive
11	young	single	ue	low	no	no	no	no	yes	no	no	postive
12	young	married	highschool	low	no	no	no	no	yes	no	no	postive
13	young	married	highschool	low	yes	no	no	no	yes	no	no	postive
14	young	married	diploma	low	yes	no	no	no	yes	no	no	postive
15	young	married	elementary	low	no	no	no	no	yes	no	no	postive
16	young	single	elementary	low	yes	no	no	no	yes	no	no	postive
17	young	divorced	diploma	low	no	no	no	no	yes	no	no	postive
18	young	married	elementary	low	yes	no	no	no	yes	no	no	postive
19	young	married	preparatory	low	yes	no	no	no	yes	no	no	postive

### 8.3.4 Appendix IV: Sample cervical cancer hospital Dataset

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	<b>age</b>	<b>mstatus</b>	<b>estatus</b>	<b>nob</b>	<b>hsti</b>	<b>ESS</b>	<b>CS</b>	<b>CM</b>	<b>VIAC</b>	<b>BAS</b>	<b>BP</b>	<b>Scresult</b>	<b>Ras</b>	<b>hospital</b>
2	adult	married	elementary	high	no	no	no	no	yes	no	no	postive	negative	BLH
3	young	single	preparatory	low	no	no	no	no	yes	no	no	postive	negative	BLH
4	adult	married	diploma	high	yes	no	no	no	no	no	no	postive	negative	BLH
5	adult	married	elementary	high	yes	no	no	no	yes	no	no	postive	negative	BLH
6	young	married	diploma	low	no	no	no	no	yes	no	no	postive	negative	BLH
7	adult	single	ue	low	no	no	no	no	yes	no	no	postive	negative	BLH
8	young	married	preparatory	high	no	no	no	no	yes	no	no	postive	negative	BLH
9	young	married	preparatory	low	no	no	no	no	yes	no	no	postive	negative	BLH
10	young	married	elementary	high	no	no	no	no	yes	no	no	postive	negative	BLH
11	young	single	ue	low	no	no	no	no	yes	no	no	postive	negative	BLH
12	young	married	highschool	low	no	no	no	no	yes	no	no	postive	negative	BLH
13	young	married	highschool	low	yes	no	no	no	yes	no	no	postive	negative	BLH
14	young	married	diploma	low	yes	no	no	no	yes	no	no	postive	negative	BLH
15	young	married	elementary	low	no	no	no	no	yes	no	no	postive	negative	BLH
16	young	single	elementary	low	yes	no	no	no	yes	no	no	postive	negative	BLH
17	young	divorced	diploma	low	no	no	no	no	yes	no	no	postive	negative	BLH
18	young	married	elementary	low	yes	no	no	no	yes	no	no	postive	negative	BLH
19	young	married	preparatory	low	yes	no	no	no	yes	no	no	postive	negative	BLH



## Appendix V: - Amharic Language Corpus for Machine Translation

```
amharic.xml | index.html | webcontent.xml | TagText.class | new 1.txt | web.xml | new 1 | amharic.t
31 <lang id="1004">
32
33 <amharic>ዋው ሆስፒታል የማጎጸን ካንሰር ሕክምና ላይ አስገራጫ ነው.</amharic>
34
35 <english>wow the hospital is smart at cervical cancer treatment.</english>
36
37 </lang>
38
39
40 <lang id="1005">
41
42 <amharic>ዋው ሆስፒታል የማጎጸን ካንሰር ሕክምና ላይ ዘመናዊ ነው.</amharic>
43
44 <english>wow the hospital is smart at cervical cancer treatment.</english>
45
46 </lang>
47
48
49 <lang id="1006">
50
51 <amharic>ዋው ሆስፒታል የማጎጸን ካንሰር ሕክምና ላይ አረፍ ነው.</amharic>
52
53 <english>wow the hospital is cool at cervical cancer treatment.</english>
54
55 </lang>
56
57
58 <lang id="1008">
59
60 <amharic>ሆስፒታል የማጎጸን ካንሰር ሕክምና ላይ በጣም ጥሩ ነው.</amharic>
61
62 <english>the hospital is very nice at cervical cancer treatment.</english>
63
64 </lang>
```

8.3.5 Appendix V: Questioner to collect Afaan Oromo opinion from different hospital patient

Universiti Jimma  
Kolleji computing  
Departimenti Tekinoloji Odefaano  
Faayadamtota hospitaloota adda adda irrati dhufanif  
Gaafiwan Guutamu qaban

Ani Abdulkadir Ahmed Barata dipartimenti computing Univarsiti Jimma yoo ta'u, Qu'ana mata dure Opinion Summarization for Afaan Oromo Language jedhamu irrati hospitaloota finfine irrati hojjacha jira waan ta'ee fi degarsa barbachisa akka na gootanu barbada. Kanafuu yaada tajajila hospitalicha irrati qabdan baka duwaa arman gaditi keename irrati gababsa nu ibsa.

---

---

---

---

---

---

---

---

---

---

Qophesan: Abdulkadir Ahmed

"Waan yaada keesan nu keenitanif galatooma."

8.3.6 Appendix VI: Questioner to collect Amharic opinion from different hospital patient

በጅማዩኒቨርሲቲ

በኮምፒዩተር ሌጅ፣ የቦሌንፎርምሲን ቲክኖሎጂ ትምህርት ክፍል

የተለያዩ ሆስፒታሎች ተጠቃሚ

የሚገኝ ማጠቃለያ

የሚጠይቁ አላማ

በጅማዩኒቨርሲቲ በቦሌንፎርምሲን ቲክኖሎጂ ትምህርት ክፍል ለማስተርስ ዲግሪ ምርምር ማዘጋጀት የሚገኝ በአሚኛ ቋንቋ ላይ ማረጋገጥ ያደረገ የሆስፒታሎች አገልግሎት ለማሻሻል እና ማረጋገጥ ለማጠናከር ይረዳዎታል። ይህ የሆስፒታልና ክሌኒካል ተጠቃሚ አስተያየት ማለያ ሲስተም ለማራት ግብዓት የሚገኝ አስተያየቶችን ለማሰባሰብ ነው።

በቅድሚያ ለሚደረጉ ጉዳዮች ትብብር ከልብ እና ማስገናኘት፡፡

ሊደረግ/ሊበረታታ እና ሊስተካከል የሚችል አጠር ባለ ማህተም ይገለጻል፡፡

ሙሉ ስም \_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

የሚጠይቁ አዘጋጅ፡ አብዱልቃዳር አህመድ በቦሌንፎርምሲን ቲክኖሎጂ ትምህርት ክፍል  
የማስተርስ ዲግሪ ተማሪ፡፡

### 8.3.7 Appendix VII: System Prototype Evaluation Form

Dear Evaluator, your evaluation has value!! The following evaluation form is used to evaluate to what extent does the developed system are usable and acceptable by the end-users of the system. Therefore, you are kindly requested to evaluate the system by putting ( ) symbol on the space provided for the respective evaluation criteria mentioned below. I would like to thanks your volunteerism in wasting your time and in providing the valuable information that helps us to evaluate the developed prototype. Note: - the values for all evaluation criteria in the table can be ranked as: Excellent=5, Very good =4, Good=3, Fair= 2 and Poor =1

No.	Evaluation Criteria	Poor	Fair	Good	Very Good	Excellent
1	The understandability of the system					
2	Attractiveness of the system					
3	The efficiency of the system					
4	Adequacy of the system in predicting disease					
5	Adequacy of the system in providing Advice					
6	The ability of the system in Recommending Health care					
7	The significance of the system for the patient					
8	The significance of the system for health care administrator					
9	The system accessibility					

### 8.3.8 Appendix VIII :sample Integration code

#### APendix:Sample Code

```
package javaapplication3;
import org.jpl7.Query;
import javax.swing.*;
import java.awt.*;
import java.awt.event.ActionEvent;
import java.awt.event.ActionListener;
import java.io.BufferedReader;
import java.io.File;
import java.io.FileInputStream;
import java.io.IOException;
import java.io.InputStreamReader;
import java.io.UnsupportedEncodingException;
import java.util.ArrayList;
import java.util.HashMap;
import java.util.Hashtable;
import java.util.Iterator;
import java.util.LinkedList;
import java.util.Map;
import java.util.Scanner;
```

```

import java.util.Set;
import java.util.logging.Level;
import java.util.logging.Logger;

private void finishActionPerformed(java.awt.event.ActionEvent evt)
{IV
try {
Thesis b=new Thesis();

String conexion="consult('Finall_mainmain.pl)";

Query conn=new Query(conexion);

System.out.println(conexion+""+(conn.hasMoreSolutions()?"Integrated"
:"failed"));

prediction ***");

double t=
b.calling((String)q1.getSelectedItem(),(String)q2.getSelectedItem(),(String)
q3.getSelectedItem(),(String) q4.getSelectedItem(),(String)
q5.getSelectedItem(),(String)
q6.getSelectedItem(),(String) q7.getSelectedItem(),(String)
q8.getSelectedItem(),(String)
q9.getSelectedItem(),(String) q10.getSelectedItem(),(String)
q11.getSelectedItem(),(String)

```

8.3.9 Appendix IX:sample medical data request letter

ደብረብርሃን ዩኒቨርሲቲ  
የኮምፒውተርና ኮሌጅ



DEBRE BERHAN UNIVERSITY  
College of Computing

ቁጥር/Ref.No:MB/የኮም/2017/479/2/17  
ቀን/Date: 9/05/2017

- ሰጤና ምክትል (MOE)
- ሰጥቁር አንበሳ ስፔሻላይዝ ሆስፒታል
- ሰዘውዲቱ መታሰቢያ ሆስፒታል
- ለቅዱስ ፓውሎስ ሆስፒታል
- የካቲት 12 ሆስፒታል
- ተክለሃይማኖት ጠቅላላ ሆስፒታል
- ለአለርት ሆስፒታል
- ሚሊኒክ ፲፯ ሆስፒታል
- አዲስ አበባ
- ለደብረብርሃን ሆስፒታል
- ደብረብርሃን

**ጉዳዩ:- ትብብር እንዲደረግላቸው ስለመጠየቅ**

ከላይ ለመግለፅ እንደተጠነቀቀው መ/ር አብዱልቃደር አህመድ የ2ኛ ዲግሪ (MSc) ተግባራዊ ስነ-ምግባር ስር "Cancer Disease Prediction and Hospital Recommendation: The use of Integrated opinion summary and fact Based Information with Data mining and Knowledge Based System" በሚል ርዕስ ጥናት እያጠኑ ስለሆነ ለሚፈልጉት መረጃ ህመምተኛውን መለየት የሚያስችሉ መረጃዎች ወይም "identifiable data" (i.e Name, Family name, Specific Kebele, Kebele Id) የማያስፈልጉ መሆኑን እየገለፁን መረጃ በመስጠት አስፈላጊውን ትብብር እንድታደርጉላቸው ስንል በአክብሮት እንጠይቃለን።



ከሰላምታ ጋር  
 ገብረ ሰጋሪ አዳም ወርቅነህ  
 Dr. Girma Adem Workneh  
 Dean, College of Computing

በመልስ ለጽፎች ስያን ደብረብርሃን ቁጥር ያጥቁ።  
 In replying, please quote our ref. no.  
 ☎ +251-6118961014  
 📠 +251-116812665

## DECLARATION

This thesis is my original work, has not been presented for a partial fulfillment of the requirement of a degree in any university and that all sources of material used for the thesis have been duly acknowledged.

---

Abdulkadir Ahmed

15, April, 2017

This thesis has been submitted for examination with my approval as a university advisor.

---

Debela Tesfaye