

**ADDIS ABABA UNIVERSITY**  
**SCHOOL OF GRADUATE STUDIES**  
**INFORMATICS FACULTY**  
**DEPARTEMENT OF INFORMATION SCIENCE**

**AUTOMATIC CLASSIFICATION OF AFAAN OROMO NEWS TEXT:  
THE CASE OF RADIO FANA**

**BY**

**ABERA DIRIBA GEMECHU**

**A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENT  
FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATION SCIENCE**

**MARCH, 2009**

---

**ADDIS ABABA UNIVERSITY  
SCHOOL OF GRADUATE STUDIES  
INFORMATICS FACULTY**

**DEPARTEMENT OF INFORMATION SCIENCE**

**AUTOMATIC CLASSIFICATION OF AFAAN OROMO NEWS TEXT:**

**THE CASE OF RADIO FANA**

**BY**

**ABERA DIRIBA GEMECHU**

**Name and Signature of Members of the Examining Board**

**Dr. Gashaw Kebede, Chairman, Examining Board** \_\_\_\_\_

**Dr. Dejene Ejigu, Advisor** \_\_\_\_\_

**Dr. Dida Midekso, External Examiner** \_\_\_\_\_

\_\_\_\_\_  
**Chairman, Faculty**

\_\_\_\_\_  
**Signature**

\_\_\_\_\_  
**Date**

\_\_\_\_\_  
**Chairman, Graduate Council**

\_\_\_\_\_  
**Signature**

\_\_\_\_\_  
**Date**

---

## **ACKNOWLEDGEMENT**

I am deeply grateful for my advisor Dr. Dejene Ejigu for his concern, constructive comments and encouragements on my work. I am also grateful to all staff of the faculty of Informatics for their support during my stay at the faculty.

This research couldn't come into being without the help of the Radio Fana Share Company, letting me use their news data. In particular thanks goes to the IT Department, Ato Muluken Berhanu, Ato Mulugeta Mihretu and the news editors particularly to Ato Admasu Damtew and Ato Mesfin Abreha for their support during data acquisition and data cleaning.

I would like to thank Oromiya Culture & Tourism Bureau for providing me Afaan Oromo Language materials and suggestions.

I am also grateful to my employer the General Education Quality Assurance and Examination Agency (GEQAEA) and particularly Ato Yohannes Afework for his understanding of the problems and encouragement, and the GEQAEA Computer Department employees for their unlimited cooperation.

Finally, I thank all my families and friends for their continuous motivation and encouragement during my stay in the university.

# TABLE OF CONTENTS

ACKNOWLEDGEMENT .....	I
LIST OF TABLES .....	VI
LIST OF FIGURES .....	VII
LIST OF APPENDICES .....	VIII
ABSTRACT .....	IX
CHAPTER ONE	
INTRODUCTION	
1.1 BACKGROUND .....	1
1.2 RADIO FANA SHARE COMPANY .....	4
1.3 AFAAN OROMO LANGUAGE .....	4
1.4 THE CURRENT NEWS MANAGEMENT SYSTEM .....	5
1.5 STATEMENT OF THE PROBLEM .....	6
1.6 JUSTIFICATION OF THE STUDY .....	8
1.7 OBJECTIVES OF THE STUDY .....	10
1.7.1 GENERAL OBJECTIVE .....	10
1.7.2 SPECIFIC OBJECTIVES .....	10
1.8 SCOPE AND LIMITATION OF THE STUDY .....	11
1.9 RESEARCH METHODOLOGY .....	11
1.9.1 LITERATURE REVIEW .....	12
1.9.2 DOCUMENT ANALYSIS .....	12
1.9.3 INTERVIEWS .....	12
1.9.4 DATA COLLECTION .....	13
1.9.5 DATA PROCESSING AND EXPERIMENTATION .....	13
1.10 ORGANIZATION OF THE THESIS .....	14
CHAPTER TWO	
LITERATURE REVIEW ON AUTOMATIC CLASSIFICATION	
2.1 INTRODUCTION .....	15
2.2 VIEWS REGARDING THE MEANING OF TEXT CLASSIFICATION AND TEXT CATEGORIZATION .....	16

<b>2.3</b>	<b>APPROACHES TO AUTOMATIC CLASSIFICATION .....</b>	<b>16</b>
2.3.1	MANUAL CLASSIFICATION .....	17
2.3.2	AUTOMATIC CLASSIFICATION .....	18
2.3.2.1	KNOWLEDGE ENGINEERING AUTOMATIC CLASSIFICATION .....	18
2.3.2.2	MACHINE LEARNING (ML) AUTOMATIC CLASSIFICATION .....	19
<b>2.4</b>	<b>BASIC CONCEPTS IN AUTOMATIC CLASSIFICATION .....</b>	<b>21</b>
2.4.1	DOCUMENT SIMILARITY MEASURES .....	22
2.4.2	SINGLE-LABEL AND MULTI-LABEL TEXT CATEGORIZATION.....	23
2.4.3	“HARD” OR “SOFT” TEXT CATEGORIZATION.....	24
2.4.4	CATEGORY-PIVOTED AND DOCUMENT-PIVOTED TEXT CATEGORIZATION .....	25
<b>2.5</b>	<b>APPLICATIONS OF TEXT CLASSIFICATIONS.....</b>	<b>25</b>
2.5.1	AUTOMATIC INDEXING FOR BOOLEAN INFORMATION RETRIEVAL SYSTEMS.....	27
2.5.2	DOCUMENT ORGANIZATION.....	28
2.5.3	TEXT FILTERING .....	29
2.5.4	WORD SENSE DISAMBIGUATION .....	29
2.5.5	HIERARCHICAL CATEGORIZATION OF WEB PAGES .....	30

**CHAPTER THREE**

**LITERATURE REVIEW ON MACHINE LEARNING (ML) APPROACH TO TEXT CLASSIFICATION (TC)**

<b>3.1</b>	<b>INTRODUCTION .....</b>	<b>32</b>
<b>3.2</b>	<b>THE IMPORTANCE OF MACHINE LEARNING.....</b>	<b>33</b>
<b>3.3</b>	<b>PROCEDURES IN AUTOMATIC DOCUMENT CLASSIFICATIONS .....</b>	<b>34</b>
3.3.1	DOCUMENT INDEXING .....	35
3.3.1.1	LEXICAL ANALYSIS OF TEXT AND DOCUMENT REPRESENTATION.....	36
3.3.1.2	ELIMINATION OF THE STOP WORDS.....	37
3.3.1.3	STEMMING .....	37
3.3.1.4	SELECTION OF THE INDEX TERMS .....	38
3.3.1.5	TERM WEIGHTING .....	39
3.3.2	CLASSIFIER LEARNING .....	41
3.3.2.1	INDUCTIVE LEARNING AND ALGORITHMS.....	43
3.3.2.2	SUPPORT VECTOR MACHINES(SVMs) .....	44
3.3.2.3	BAYESIAN CLASSIFICATION .....	50
3.3.2.4	K-NEAREST NEIGHBOR CLASSIFIERS(KNN).....	53
3.3.2.5	CLASSIFICATION BY DECISION TREE INDUCTION .....	54
3.3.2.6	MAJOR CHALLENGES OF TEXT CLASSIFICATIONS.....	58

3.3.3 CLASSIFIER EVALUATION .....	61
3.3.3.1 CLASSIFIER PERFORMANCE MEASURES .....	61
3.3.3.2 ESTIMATING CLASSIFIER ACCURACY .....	66
<b>CHAPTER FOUR</b>	
<b>THE AFAAN OROMOO LANGUAGE</b>	
4.1 INTRODUCTION .....	68
4.2 THE OROMO ALPHABET .....	69
4.2.1 THE AFAAN OROMO VOWELS .....	69
4.2.2 THE AFAAN OROMO CONSONANTS .....	72
4.3 MORPHOLOGY .....	73
4.3.1 INFLECTIONAL MORPHOLOGY .....	75
4.3.2 DERIVATIONAL MORPHOLOGY .....	75
4.4 THE AFAAN OROMO MORPHOLOGY.....	76
<b>CHAPTER FIVE</b>	
<b>AUTOMATIC CLASSIFICATION OF AFAAN OROMOO TEXTS</b>	
5.1 INTRODUCTION.....	78
5.2 DATA SOURCE .....	80
5.3 CREATING THE AFAAN OROMO NEWS ITEMS DATABASE.....	81
5.4 DATA PREPARATION .....	83
5.4.1 THE HEADER OF THE NEWS ITEMS .....	84
5.4.2 THE SUMMARY OF THE NEWS ITEMS.....	84
5.4.3 THE BODY OF THE NEWS ITEMS .....	84
5.5 NEWS TEXT PRE-PROCESSING FOR CLASSIFICATION .....	85
5.5.1 TOKENIZATION .....	85
5.5.2 REMOVAL OF THE STOP WORDS .....	86
5.5.3 WORD STEMMING.....	87
5.5.4 DATA TRANSFORMATION AND SCALING.....	88
5.6 THE NEWS ITEM CLASSIFICATION PROCESS.....	89

<b>5.7</b>	<b>TESTING CLASSIFIER ALGORITHMS.....</b>	<b>89</b>
5.7.1	CLASSIFICATION USING SEQUENTIAL MINIMAL OPTIMIZATION (SMO).....	90
5.7.1.1	EXPERIMENT ON FOUR CATEGORIES.....	90
5.7.1.2	EXPERIMENT ON SEVEN CATEGORIES .....	93
5.7.1.3	EXPERIMENT ON ELEVEN CATEGORIES .....	94
5.7.2	CLASSIFICATION USING NAIVEBAYESMULTINOMINAL.....	96
5.7.2.1	EXPERIMENT ON FOUR CATEGORIES.....	96
5.7.2.2	EXPERIMENT ON SEVEN CATEGORIES .....	97
5.7.2.3	EXPERIMENT ON ELEVEN CATEGORIES .....	99
<b>5.8</b>	<b>AUTOMATIC TEXT CLASSIFIERS AND AFAAN OROMO TEXTS.....</b>	<b>100</b>
 <b>CHAPTER SIX</b>		
<b>CONCLUSIONS AND RECOMMANDATIONS</b>		
<b>6.1</b>	<b>CONCLUSIONS.....</b>	<b>104</b>
<b>6.2</b>	<b>RECOMMANDATIONS.....</b>	<b>107</b>
<b>BIBLIOGRAPHY.....</b>		<b>109</b>
<b>APPENDICES.....</b>		<b>115</b>
<b>DECLARATION.....</b>		<b>120</b>

## LIST OF TABLES

Table 3.1	Two Class Confusion Matrix.....	62
Table 4.1	The Afaan Oromoo Vowels.....	70
Table 4.2	Major Places of articulations.....	71
Table 4.3	The Afaan Oromoo Consonants.....	72
Table 5.1	Main News Categories.....	81
Table 5.2	Number of News Items by Categories.....	83
Table 5.3	SMO 4 Category Confusion Matrix.....	91
Table 5.4	SMO 4 Category Detailed Accuracy.....	92
Table 5.5	SMO 7 Category Confusion Matrix.....	93
Table 5.6	SMO 7 Category Detailed Accuracy.....	94
Table 5.7	SMO 11 Category Confusion Matrix.....	95
Table 5.8	SMO 11 Category Detailed Accuracy.....	95
Table 5.9	NBM 4 Category Confusion Matrix.....	96
Table 5.10	NBM 4 Category Detailed Accuracy.....	97
Table 5.11	NBM 7 Category Confusion Matrix.....	98
Table 5.12	NBM 7 Category Detailed Accuracy.....	98
Table 5.13	NBM 11 Category Confusion Matrix.....	99
Table 5.14	NBM 11 Category Detailed Accuracy.....	100
Table 5.15	Classifiers Average Accuracy.....	102



## LIST OF FIGURES

Figure 3.1	SVM Small Margin.....	46
Figure 3.2	SVM Large Margin.....	47
Figure 3.3	Non Linear Region.....	49
Figure 3.4	Kernel Function Transform data.....	49
Figure 5.1	Machine Learning Approach to Text Classification.....	79
Figure 5.1	Average Accuracy Graph.....	103

## **LIST OF APPENDICES**

Appendix 1	Samples of Records removed during Manual Scanning.....	115
Appendix 2	Samples of the News Head Line from the Database.....	116
Appendix 3	List of Afaan Oromo Stop Words used in the Research.....	117

## **ABSTRACT**

The vast growth of information and communication technology resulted in a huge volume of information very large bulk of which is stored as unstructured text. The presence of so much text in electronic form is a challenge to natural language processing. As the volume of electronic information increases, there is growing interest in developing tools to help people better find, filter, and manage these resources. Arguably, the only way for humans to cope with the information explosion is to exploit computational techniques that can sift through huge bodies of text.

Currently news agencies in Ethiopia in which large amount of news from all the available sources are processed every day is implementing a manual classification system to categorize news items in their daily activities despite the fact, they are using computerized system to store and edit news items. Radio Fana is the one among these agencies.

The objective of this research is to develop and adopt processing tools for Afaan Oromo text classification and investigate the application of machine learning techniques for automatic classification of Afaan Oromo news items.

The data source for this research is the Afaan Oromo news items obtained from Radio Fana Share Company.

In this research, tools for pre-processing Afaan Oromo news items such as tokenization, removal of extraneous characters, removal of stop-words and removal of affixes from the words are prepared to facilitate the experimentation process for the automatic classifiers.

Among the automatic classifiers which are applicable on high dimensional data, four of them; Sequential Minimal Optimization (SMO) algorithm from Support Vector Machines, NaiveBayesMultiNominal (NBM) from Bayesian Classifiers, J48 algorithm from the Decision trees and K-Nearest Neighbor (KNN) from the Lazy Learners have been experimented on the final data. The data, the pre-processed Afaan Oromo news items, is organized in to categories of four classes, seven classes and all (eleven) classes for the experimentation purpose and the experimentation uses 10-fold stratified cross validation for training and test data.

For the SMO and NBM classifiers, which have best accuracy over the others, the detailed accuracy by class together with the confusion matrix of the experimentation is shown, whereas for J48 and KNN classifiers the average accuracy on each category is presented in this thesis.

The result of the experimentation is encouraging, the best result (accuracy) from both the SMO and BayesMultiNominal classifiers, 95.82% and 96.58% respectively, is obtained when the number of instance documents is approximately equal in the classes, and it was for the four categories of news items. The lower accuracy seen is for J48 on category of 7 classes, 79.69% and on category of 11 classes, 82.05%. SMO tends to have better accuracy over the other classifiers for the Afaan Oromo news items classification. In all the classifiers, unevenly distributions of instances of documents in classes tend to decrease the accuracy of the classifiers when taken together, i.e, experimentation on all of the eleven categories taken together; while an increase in number of instances in a given class tends to increase the accuracy for the class.

Accordingly, from the result of this research, it was observed that Machine Learning approach can be applied to Afaan Oromo news items classification task, nevertheless, additional works are recommended in order to come up with best result.

**Key Words:**

Natural language processing, machine learning, text classification, document indexing, Classifier algorithms, Afaan Oromo news

# CHAPTER ONE

## INTRODUCTION

### 1.1 BACKGROUND

The vast growth of information and communication technology enforced organizations, professionals and people from all walks of life to be confronted by unprecedented volume of information very large bulk of which is stored as unstructured text. The presence of so much text in electronic form is a huge challenge to natural language processing. Arguably, the only way for humans to cope with the information explosion is to exploit computational techniques that can sift through huge bodies of text.

Information overload (in which individuals are faced with an oversupply of content) could become the road rage in this new millennium. In order for this content to become useful information that can empower users, rather than frustrate or confuse them, we need novel ways of delivering only what is needed, at the right time and in the right format [Rafael, et al., 2004].

News agencies and mass media enterprises including radio and television agencies are among those in which information overload increase rapidly. News stories are the most important source of up-to-date information about the world around us, and represent some of the most often updated and highest quality content in these agencies. Therefore, it is most important to develop ways to processes news efficiently.

News is the reporting of current information on television, radio, and in newspapers and magazines. In its border sense News is any new information or information on current events which is presented by print, broadcast, Internet, or word of mouth to a third party or mass audience.

With this explosive growth in the number of electronic documents available on the Internet, digital libraries and news, it is increasingly difficult for news agencies to categorize (classify) the electronic documents using just a manual approach. To improve the effectiveness and efficiency of document categorization at the news agencies, more in-depth studies of using automatic document classification methods to categorize news items are required. Automatic classification techniques use algorithms that learn from human classifications techniques. Its goal is the classification of documents into a fixed number of predefined categories.

Classification is the process of finding a set of models (or functions) that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data, objects whose class label is known [Han and Kamber, 2004].

In a manual text categorization (classification) organizations define subjective topic trees or categories, based on certain preferences and assign documents they write or receive to these categories according to this subjective definition of categories or classes.

Manual categorization therefore, is based on human judgments. Organizations and users search for information and save it in categories meaningful to them selves. Due to the subjective nature of manual document classification, one may not follow the categorization made by the other.

In a situation where there is a large collection of electronic document, it is difficult to manage and classify documents manually; this is especially more difficult in a situation where time and speed is a big constraint as is in news agencies. Therefore, development of the language processing tools for the information retrieval system and automatic classification of the documents becomes important for the situations mentioned.

Document classification is a time consuming task that has historically been performed by humans. The advent of the World Wide Web, combined with today's information-driven economy, has generated substantial interest in the automatic classification of these documents in order to save time and money [Mark W. Pope, 2007].

Automatic text Classification, or the task of automatically assigning semantic categories to natural language text, has become one of the key methods for organizing online information. Since hand-coding classification rules is costly or even impractical, most modern approaches employ machine learning techniques to automatically learn text classifiers from examples [Joachims T,2002].

## **1.2 Radio Fana Share Company**

Radio Fana is a private share company registered under the Federal law of Ethiopia and licensed by the Broadcast Authority of Ethiopia. It is the nation's first Commercial, National Broadcaster and multi-lingual Radio Station established in November 21, 1994. At establishment, Radio Fana started its broadcasting in Amharic and Afaan Oromo languages for 38 hours a week.

Currently the station possess a high-quality audibility and a nation-wide audience reach market edge; for which it operate a powerful and state-of-the-art SW,MW and FM transmission facilities. It has launched the FM transmission on March 10, 2007. Radio Fana currently broadcast programs in Amharic, Afaan Oromo, Somali and Afar languages for a total of 252 hours service per week both on the national and FM 98.1 service. Radio Fana is a News-Talk-Variety (Information and entertainment) format station.

## **1.3 Afaan Oromo Language**

Afaan Oromo is a widely spoken language in Ethiopia. It is a working language of the regional state of Oromiya and is a medium of instruction for primary schools in the region. Afaan Oromo language besides being given in all primary and secondary schools in Oromiya, currently many governmental and private colleges and universities give training in Afaan Oromo as a major field of study.

There are varieties of publications in hard copies and vast amount of information in electronic formats written in Afaan Oromo.



The expansion of telecommunication and computers with the Internet through out the country as well as in Oromiya in governmental offices, higher learning institutions and business enterprises etc, made all sorts of information to become available in Afaan Oromo electronically. The radio and television broadcasts in Afaan Oromo by both the regional state of Oromiya and the Federal government of Ethiopia are among those in which large quantities of electronic materials in Afaan Oromo is processed.

Radio Fana Share Company is among the major organizations in which vast inflow of daily Afaan Oromo news items is processed and broadcasted.

#### **1.4 The Current News Management System**

The currently existing computerized system at Radio Fana was emerged in 2006 as the solution for the previously existing system for news management. However, now the current system itself needs change to cope up with this day's information explosion.

Before 2006 news items from all corners of the reporters was hand-written at the office and is delivered to the typists to be typed so that it can be used for documentation and broadcast. The main problems in this system was the time news items take to be typed by the limited number of typists, timeliness, freshness, and the documentation error of items in their proper location for retrieval purpose.

The currently computerized existing system allows each news editor to write and save news items electronically in one of the 11 main categories and different number of sub categories under the main categories. The category assignment to the news items by the editor is done according to the news items content, for example sport, economy, social affairs e.t.c. The system used is a bi-lingual – Amahric and Afaan Oromo.

The system saves each news items prepared from both Amharic and Afaan Oromo desks to a single RadioFana database in a classes/categories<sup>1</sup> supplied by the editors. It is a client-server architecture which help the editors working from their office and save on a centralized server. The system provides different report generation mechanism on the past news item; for example news items by the name of editors, by date, or by its classification. The system is a database application whose front-end is made using Microsoft Visual Basic Programming Language. The database is implemented on Microsoft SQL Server Database management system. There are 26 tables designed for data storage and different administrative purposes in a database. The news story table has information (attributes) about each news items, such as ID, Head Line, dates created, key words, slug, story owner, classification code, full story etc. The ID field is an auto number which increases for every news item created for Amharic or Afaan Oromo. At any time the news items previously saved can be called and edited or deleted by authorized personnel or different kinds of reports can be generated from the database.

### **1.5 Statement of the problem**

In news industries large amount of news from all the available sources are processed every day. Reporters write news story articles. News agencies collect the news articles, edit it and prepare it in a way suitable for the users.

Radio Fana being one of the main Radio station that serves for many hours a week, established a system, which enables news items editors to electronically process each news items that makes news processing and delivering effective and efficient.

---

<sup>1</sup> Categorization and classification are used interchangeably in this thesis.

The editors prepare the content of each news items and categorize it by it's content-subject and saves it by the production date of the news item, so that it can be retrieved later by its date and category when needed. There are several practical reasons why previously saved news items are needed now and then, for example, the past news may be used for background of the current news items, the number of news items per each category performed in the past half a year may be evaluated, several reports on the past news items are needed to be generated for current executions and future plans.

However, now it has been practically seen, that News items are normally classified to the categories according to the taxonomies that are considered relevant by the editors subjectively. As a consequence of this, it is some times difficult to find news items in their assumed categories. This caused problem during search for previous news items and when retrieval of certain sub-categories of news items are required. For example, news items which would have been categorized, as a culture would be found in an economy or sport category etc.

Further more, this classification scheme becomes very difficult because it requires human expertise to spot relationships between the taxonomy and the documents. It also depends on how much the editors are responsible during the classification of the news items. Even the experts some times do not agree on what should go where and this resulted in considerable classification variation.

In addition to this, manual classification is time consuming which is in a contrary to the fact that speed is a major factor in news agencies, other wise a flood of information occurs and decreases its usability. All classifications, even the most general are carried out for some more or less explicit special purpose or set of purposes which should

influence the choice of classification method and the result obtained. Therefore, there should be a way to get rid of this manual categorization method.

One of the main requirements in news agencies is the freshness of news that is delivered to the target audience. With this requirement in mind, the usage of complex time-consuming processes should be replaced by systems that come with the idea of using automatic classification techniques.

A promising way to deal with automatic classification is to learn a categorization scheme automatically from training examples. Once the categorization scheme is learned, it can be used for classifying future documents [Lam W., et al., 1999].

## **1.6 Justification of the study**

There are a number of news agencies preparing news in Afaan Oromo language in Ethiopia and abroad. For example, in Ethiopia, the Ethiopian Radio and Television Agency, The Radio and Television programs prepared by the Oromiya Regional State and the Radio Fana Share Company can be cited.

Currently, Radio Fana also has Web-based news both in text and voice over the Internet in Afaan Oromo language.

The volume of electronic materials available in Afaan Oromo language continues to increase; therefore there is a growing interest in automatic classification techniques for the language. The assignment of natural language texts to a predefined category based on their contents is an important component in information organization and management tasks. Its most widespread application to date has been for assigning subject categories to documents to support text retrieval, routing and filtering.

Automatic text classification can play an important role in a wide variety of more flexible, dynamic information management tasks as well. For example real-time sorting of files into folder hierarchies, topic identifications to support topic-specific processing operations; structured search, or finding documents that match long-term standing interests or more dynamic task-based interests.

The widespread and increasing availability of text documents in electronic form increases the importance of using automatic methods to analyze the content of text documents, because the method using domain experts to identify new text documents and allocate them to well-defined categories is time-consuming and expensive, has limits, and does not provide continuous measure of the degree of confidence with which the allocation was made [Olivier, 2000]. As a result, the identification and classification of text documents based on their contents are becoming imperative [Zhang et al, 2005].

This research focuses on the problem of automatic categorization of texts in general and that of Afaan Oromo language in particular. Text categorization is the task of deciding to what category a document belongs among a set of pre-specified classes of documents. Automatic classification schemes can greatly facilitate the process of categorization.

Automatic classification requires preparation of the source data before the classifiers can use it. Text categorization starts by transforming the documents, which typically are strings of characters, into a representation suitable for the learning algorithm and the classification task. This includes data collection and cleaning, performing lexical analysis

on the data, feature selections, data transformations and employ classifiers on the processed data.

Therefore, this research contributes to the development of processing tools and towards finding optimal classifier for Afaan Oromo texts.

Furthermore, it can be said that in computational aspects of the Afaan Oromo language, there are only few works being done so far. Development of stemming algorithm for Afaan Oromoo text [Wakshum, 2000]; Text to Speech for Afaan Oromoo [Morka, 2001]; Sentence Parser for Afaan Oromoo texts [Diriba, 2002] and Development of Morphological Analyzer for Afaan Oromoo Text [Assefa, 2005], but no one considered automatic classification of Afaan Oromoo texts to the best of the researcher's knowledge. Hence, it is possible to say only a little has been done on Natural Language Processing for Afaan Oromoo language. Therefore, this study contributes to the Natural Language Processing of Afaan Oromoo Language.

## **1.7 Objectives of the Study**

### **1.7.1 General Objective**

The major objective of this thesis is to develop and adopt processing tools for Afaan Oromo text classification and evaluate the performance of selected automatic classifiers for Afaan Oromo text classification activities based on contents of the documents.

### **1.7.2 Specific Objectives**

- Review literature on the concept of classification and available techniques for classifying documents automatically
- Study existing tools for Afaan Oromo text processing
- Build stop word list for Afaan Oromo news text

- Develop and adopt tools for pre-processing Afaan Oromo documents for classification purpose.
- Apply selected classifiers on processed Afaan Oromo text.
- Evaluate performance of the applied classifiers.

## 1.8 Scope and Limitation of the Study

Radio Fana is a multi-language service station. Currently it prepares electronic news items in Amharic, Afaan Oromo, Somali, and Afar languages. This research work is about automatic classification of Afaan Oromo news items based on data from the Radio Fana.

The researcher has seen that there is a stemmer for Afaan Oromo text developed by [Wakshum, 2000]. However, despite the effort made to get the developed stemmer, the researcher could not get the stemmer to create fully stemmed terms for document representations. Instead of a full stemmer, simple suffix removal tool is developed by the researcher, which does not serve as a full stemmer.

Due to time constraint to experimentation and analyze the results; not all applicable algorithms which are available for automatic classification purposes in the tool (WEKA<sup>2</sup>) are tested for automatic categorization of the Afaan Oromo news items.

## 1.9 Research Methodology

In machine learning approach to automatic document classifications, three phases can be roughly distinguished in the life cycle of automatic classification systems. In the first phase the document texts need pre-processing tasks such as lexical analysis, stop

---

<sup>2</sup> WEKA : Waikato Environment for Knowledge Analysis. Developed at the University of Waikato in New Zeland.

words removal, features-selection, stemming, and documents representation which is referred to as the document indexing phase. The second phase is a building of a text classifier from predefined classes which is referred to as classifier learning phase. And the third is the classifier evaluation phase. Depending on each phases, the following methods are employed in conducting this research.

### **1.9.1 Literature Review**

Literatures were reviewed on:

- Automatic classifications of documents from articles in journals, books, thesis, Internet, previous related works etc.
- Classifier algorithms and their applications.
- Developed tools and techniques for Afaan Oromo data and information processing.
- Current and past practices of news classification systems.

### **1.9.2 Document Analysis**

In order to enrich the understanding of the manual Afaan Oromo news items classification system at Radio Fana, the following were observed.

- The current classification scheme at Radio Fana
- The manually classified Afaan Oromo news items
- The currently used software

### **1.9.3 Interviews**

To have clear understanding on the current manual classification scheme and the problem area, interviews with the Radio Fana employees, domain-experts working in the area and the management were conducted.



#### **1.9.4 Data Collection**

The source of data was the news items that were obtained from the Radio Fana Share Company. The data is in SQL server database, which comprises both the Amaharic and Afaan Oromo news items in a single database. The SQL server database was imported into Ms-Access in order to make it more suitable for preprocessing. After, filtering for the Afaan Oromo news items from the former database, 3,710 news items in Afaan Oromo were prepared for further preprocessing. The Afaan Oromo news items were manually classified by the experts into 11 classes. The preprocessing activities on the news items are discussed in details in chapter 5.

#### **1.9.5 Data Processing and Experimentation**

The source data, the Afaan Oromo news items, cleaned using both preliminary manual inspection and automatic methods before it is supplied to the classifier. The Visual Basic Programming Language is used in the process. The choice for the programming language is its ease of use in database manipulation and the researcher's familiarity in use of the language.

The data pre-processing carried out in this research, involved developing and adopting tools for

- Tokenization of Afaan Oromo text i.e. word identification, extraneous characters, digits and punctuation mark removal
- Group news items in to categories from a database
- Removal of the stop words and word affixes etc...

The processed Afaan Oromo documents were changed to Arff (attribute relation file format) file format, which is suitable for the tool called (Weka), used in this research. Weka application package tool is used because of its free availability and is widely used software that performs statistical text classification.

Selected automatic classification algorithms were applied on the experimental data and their performance were evaluated. The details of text classification life cycle are presented in chapter 3.

### **1.10 Organization of the Thesis**

The thesis is organized into six chapters. The first chapter was Introduction and overview to the research. It has presented background, problem statement, objectives and methodology. In chapter two concepts in automatic text classification and its application areas are discussed. Chapter three reviews machine learning concepts and the phases of document classification together with text classifier algorithms. Chapter four is a review of the Afaan Oromo language writing system and word formation. Chapter five presents the automatic classification of Afaan Oromo texts. Finally the research findings and recommendations are presented in chapter six.

## **CHAPTER TWO**

### **LITERATURE REIVIEW ON**

### **AUTOMATIC CLASSIFICATION**

#### **2.1 INTRODUCTION**

Classification is the process of assigning elements or units to classes (or types or groups) according to some criteria. Often it is said that the criteria of classification are likeliness; a classification unites like things and it scatters unlike things. Things may; however be like in many different ways, while a classification should unite things from a functional or a pragmatic point of view based on the purpose of the classification.

Humans are used to perform different kinds of classifications and inferences nearly every second of their existence and often without even realizing it. These tasks range from looking at an object and classifying it as a house or a horse; over listening to a new piece of music and correctly assigning its composer.

Human beings try to understand things in their taxonomic relations. Taxonomy is the classification of items into groups, for example dogs, cats and humans are part of the group of mammals.

The automated categorization (or classification) of texts into pre-specified categories, although dating back to the early 1960s, has witnessed a booming interest in the last ten years, due to the increased availability of documents in digital form and the ensuing need to organize them [Sebastiani, 2002]

In this chapter, the basics of automatic classification and a review of different approaches towards automatic classification and the applications of text classification are discussed.

## **2.2 Views regarding the meaning of text classification and text categorization**

Many researchers used the terms “Text classification” and “Text categorization” interchangeably.

Text categorization (also known as text classification, or topic spotting), is the activity of labeling natural language texts with thematic categories from a predefined Set [Sebastiani, 2002].

Text categorization (also known as text classification) is the task of assigning predefined categories to free-text documents [Yang Y. et.al., 2008].

Text categorization is the classification of texts with respect to a set of predefined categories [Tokunaga, 1994]

Automatic classification and automatic categorization is used interchangeably in this study.

## **2.3 Approaches to Automatic Classification**

Automatic classification is the process of constructing a procedure that will be applied to continuous sequences of cases, in which each new case must be assigned to one or more predefined classes/categories. These categories may or may not be exclusive, that is for example in a document classification a document may or may not belong to more than one category. Indeed, a document may not belong to any categories defined at all.

The goal of text categorization is the classification of documents in to a fixed number of predefined categories [Joachims T., 1998].

Historically, the first attempts at automated classification involved the manual generation of deterministic rules that assigned a document to a category. But as the number of document to be classified increased exponentially the use of automatic classification which require less human expert involvement become a must.

A text classification system can be used in indexing documents to assist information retrieval tasks as well as in classifying e-mails, memos or web pages in a different manner. Needless to say, automatic text classification is essential.

### **2.3.1 Manual classification**

Manual classification involves human experts manually classify documents based on classification rules and subjective judgment. This approach has been used in library practice for many years to organize, index and retrieve documents. Human experts typically assign each book a code representing a category according to a set of classification scheme.

The main challenge of manual classification is its demand on resources. Manual classification is a time consuming process that relies heavily on domain knowledge. It requires significant investment of time from many human experts with knowledge of different domains. Another associated problem is that subjective judgments can generate inconsistent classification results. Because of these limitations, manual classification works best for relatively small document collection.

### **2.3.2 Automatic classification**

To address the problems of manual classification, automatic text classification is explored as an alternative approach using different techniques. In contrast to manual classification, automatic classification offers the advantages of automation, efficiency, and consistency.

Automatic classification employed rule-based or knowledge engineering and machine learning (ML) automatic classification techniques.

#### **2.3.2.1 Knowledge engineering automatic classification**

In knowledge engineering also called rule-based classification, you group your documents together, decide on categories, and formulate the rules that define those categories; these rules are actually query phrases. You then index the rules and use these rules to classify documents.

Human-engineering, rule-based models for assigning subject codes, while relatively effective, are also very expensive in time and effort for their development and continued support. Defining rules can be tedious for large document sets with many categories. As the document set grows, it may needed to write correspondingly more rules.

This approach was fraught with problems, amongst them the chronic lack of scalability of the rule set and maintenance issues required dealing with generalizing to new documents and categories. In short, the rule approach required too much work for too little gain. As such, a move to a less maintenance heavy formalism was required.

### **2.3.2.2 Machine learning (ML) automatic classification**

Instead of manually classifying documents or hand-crafting automatic classification rules, statistical text categorization uses machine learning methods to learn automatic classification rules based on human-labeled training documents.

Machine learning (ML) can be defined as a way of how machines can generalize from experiences made (by presenting training data or interaction with their surrounding). The problem to be solved is generally perceived as a function that maps some input to a reaction or classification ( $f: \mathbf{X} \rightarrow \mathbf{Y}$ ). The goal of training is to approximate the real function  $f$ . That means a learning system that was presented some examples of houses, should learn the patterns, which all (or most) houses have in common. The function here could be the mapping of a pixel matrix into the domain {true, false}. These patterns should enable them to classify a new house, they have not “seen” before, as a house, by verifying the learned patterns.

Modern classification approaches now employ Machine Learning techniques. Machine learning is a fast growing field of computer science which is concerned with the creation of computer programs that learn from experience (supervised learning). In the case of text classification, the task to be learnt is an objective function from a set of functions called the hypothesis space, which maps candidate documents onto one or more categories. A set of pre-classified documents provides the “experience” necessary for a classifier to learn the objective function. This set is typically marked up by a human, and is the only human input required to operate a classifier: the learning and subsequent classification can be done automatically. It is possible to learn from a document set that has not already been classified (this is unsupervised learning in contrast to supervised

learning), but the vast majority of approaches do not consider this possibility. Classification performance is typically significantly higher with supervised learning (Robinson H., 2003).

Using machine learning techniques, automatic text classification assigns documents to a set of pre-defined categories. This approach has been applied in many areas, such as patent classification, news delivery, and email spam filtering.

Due to bulk increase of textual data within the past decade, the interest in automated text categorization has steadily increased. The applications range from automatic document indexing for information retrieval systems, document organization, text filtering, word sense disambiguation, categorization of web pages and, most recently, spam filters. One particularly interesting application area is the news industry.

In general, automated text categorization may be regarded as two separate tasks.

First, document indexing is needed in order to transform the natural language text into a numerical representation suitable for further processing.

The second task is the actual classification. Document indexing consists of choosing the appropriate set of keywords (feature selection) based on the whole corpus of documents, and assigning weights to those keywords for each particular document, thus transforming each document into a vector of keyword weights. This chapter discusses the basic steps of automatic classification.



## **2.4 BASIC CONCEPTS IN AUTOMATIC CLASSIFICATION**

Automatic classification is analyzing the content of the given document, and then assigns it to the set of documents or class which has similar content characteristics to the current document.

The properties or the attributes of an object can be any thing that characterizes that object in its identification. For example the height of a people can be used to classify them as short, medium or long.

Considering the document objects, automatic classification attempts to select the correct predefined class based on the characteristics of a document to be classified and the characteristics of the documents previously assigned to each class. In other words, the process involves training systems to recognize characteristics of documents belonging to a particular classification group [Zelalem, 2001].

Text categorization is the procedure of labeling a textual document with one or more predefined categories [Sebastiani, 2002].

The goal of classification process is to group similar documents together. The central notion of text classification, which is a membership of a document  $D_j$  in a class  $C_i$  is based on the content-meaning of  $D_j$  and  $C_i$ . All the members of a given class have more or less the same properties than the members of the other different class.

The similarity between two objects is normally computed as a function of the number of properties that are assigned to both objects; in addition the number of properties that are jointly absent from both objects may be taken into account (Salton G., 1983)

This implies, the procedure of classifying documents into groups require a quantitative measure of the “likeness” of the document for a given class, and the separation of unlike ones using their contents that describe their properties. Therefore, the first thing in the process of automatic classification is to identify the attributes of documents and classes so that the matching of the two becomes possible.

The properties of the documents which are used to identify their behavior, their content, are content bearing words in a given document.

The classic models in information retrieval consider that each document is described by a set of representative keywords called index terms. An index term is simply a (document) word whose semantics helps in remembering the document’s main themes [Salton G, 1983].

In fact, representing documents for the purpose of classification is the first and important step in the process of automatic classification. In other words, documents must be represented through features and these features will be used to determine the similarity of documents with different classes. Similarly, the classes should be represented in order to be able to evaluate their similarity with documents. Once the classes and documents are represented, then a matching technique can be applied to determine the similarity between documents and classes to identify the most similar class to a given document.[Zelalem, 2001]

#### **2.4.1 Document similarity measures**

There are a number of functions which measures the similarity or likeness of documents. As mentioned previously the similarity between two objects is normally computed as a function of the number of properties that are assigned to both objects.

In the case of documents the properties (attributes) are the index words of the documents. Therefore, the similarity measure between documents is based on the likeness of their index terms. Documents are represented as a vector using their index terms weight. Consider two documents  $i$  and  $j$

$$\text{Doc}_i = (\text{term}_{i1}, \text{term}_{i2}, \text{term}_{i3}, \dots, \text{term}_{it})$$

$$\text{Doc}_j = (\text{term}_{j1}, \text{term}_{j2}, \text{term}_{j3}, \dots, \text{term}_{jt})$$

Any of the similarity functions for instance Cosine coefficient, Dice coefficient or other can be used on the vector of the documents to find the similarity between the two documents.

For example, in Cosine similarity function, we measure the angle between the vectors of the documents. As the angle between the vectors shortens, meaning that the two vectors are getting closer, similarity of the documents represented by the vectors increases.

#### **2.4.2 Single-label and Multi-label text categorization**

In text classification different constraints may be enforced on the text classification task according to its application. For the set of classes  $\mathbf{C}$  and for the set of documents  $\mathbf{D}$  where :

$$\mathbf{C} = \{c_1, c_2, c_3, c_4, c_5, c_6, \dots, c_m\}$$

$$\mathbf{D} = \{d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8, d_9, d_{10}, \dots, d_n\}$$

We can consider the following, for a given integer  $\mathbf{k}$

Exactly  $\mathbf{k}$  elements of  $\mathbf{C}$  are assigned to each  $\mathbf{d}_j \in \mathbf{D}$ . The case in which exactly one category ( $\mathbf{k} = 1$ ) must be assigned to each  $\mathbf{d}_j \in \mathbf{D}$  is often called the Single- labeled

(also called non overlapping categories) case, while the case in which any number of categories from  $0$  to  $m$  may be assigned to the same  $d_j \in \mathbf{D}$  is called multi label ( Overlapping categories) case [Sebastiani, 2002].

This study employs single-label categorization as the selected data for the experimentation allows such classification task.

### **2.4.3 “Hard” or “Soft” text categorization**

Hard categorization is when the classifier algorithm decides, yes or no, as to whether a document  $d_j \in \mathbf{D}$  belongs to a category  $c_i \in \mathbf{C}$  . This is the kind of decisions that are taken by autonomous text classifiers, that is, software systems that need to decide and act accordingly without human supervision.

A different type of decision, sometimes referred to as a “soft” categorization decision is the one which consists of attributing a numeric score (e.g. between 0 and 1) to the pair  $(d_j, c_i)$  reflecting the degree of confidence of the classifier in the fact that  $d_j$  belongs to  $c_i$  . This allows, for instance, ranking a set of documents in terms of their estimated appropriateness for category  $c_i$ , or ranking a set of categories in terms of their estimated appropriateness for  $d_j$ . Such rankings are often useful for non-autonomous, interactive classifiers, i.e. systems whose goal is to recommend a categorization decision to a human expert, who is responsible of taking the final decision. For instance, in a single-label text categorization task a human expert in charge of the final classification decision may profit from a system which pre-ranks the categories in terms of their estimated appropriateness to a given document  $d_j$ . [Sebastiani, 2002].

#### **2.4.4 Category-Pivoted and Document-Pivoted text categorization**

There are two different ways in document class arrangements for using it. Given a document, we might want to find all the categories under which it should be filed (document-pivoted categorization – DPC); alternatively, given a category, we might want to find all the documents that should be filed under it (category-pivoted categorization – CPC). Quite obviously this distinction is more pragmatic than conceptual, but is important in the sense that the sets **C** of categories and **D** of documents might not always be available in their entirety right from the start. DPC is thus suitable when documents become available one at a time over a long span of time, e.g. in filtering e-mail. CPC is instead suitable if it is possible that

- i. A new category  $c_{n+1}$  is added to an existing set  $\mathbf{C} = \{c_1, c_2, c_3 \dots , c_n\}$  after a number of documents have already been classified under **C**, and
- ii. These documents need to be reconsidered for classification under  $c_{n+1}$

DPC is more commonly used than CPC, as the former situation is somehow more common than the latter [Sebastiani, 2002].

### **2.5 Applications of text Classifications**

Text classification is the task of assigning predefined categories to free-text documents. It can provide conceptual views of document collections and has important applications in the real world. For example, news stories are typically organized by subject categories (topics) or geographical codes; academic papers are often classified by technical domains and sub-domains; patient reports in health-care organizations are often indexed from multiple aspects, using taxonomies of disease categories, types of

surgical procedures, insurance reimbursement codes and so on. Another widespread application of text categorization is spam filtering, where email messages are classified into the two categories of spam and non-spam, respectively [Yang Y. et.al., 2008].

Generally speaking, the applications of text classification are manifold. Common traits among all of them are:

- The need to handle and organize documents in which the textual component is either unique, or dominant, or simplest to interpret, component.
- The need to handle and organize large quantities of such documents that is large enough that their manual organization into classes is either too expensive or not feasible within the time constraints imposed by the application.
- The fact that the set of categories is known in advance, and its variation over time is small.

The difference between applications of text classification mainly varies across the wide dimensions of its applicability.

[Sebastiani, 2005] Referring to [chakrabarti et. al., 1998] and [Slattery et. al., 2000] mentioned that the application of text classification vary along the following various dimensions.

- **The nature of the document** ; documents may be structured texts ( such as scientific articles), newswire stories, classified advertisements, image captions, e-mail messages, transcripts of spoken texts, hypertexts, or other. If the documents are hyper textual, rather than textual, very different techniques may be used, since links provide a rich source of information on which classifier learning activity can leverage.

- **The structure of the classification scheme**, i.e. whether this is flat or hierarchical. Hierarchical classification schemes may in turn be tree-shaped, or allow for multiple inheritances. Again, the hierarchical structure of the classification scheme may allow radically more efficient, and also more effective, Classification algorithms, which can take advantage of early sub tree pruning improved selection of negative examples, or improved estimation of word occurrence statistics in leaf nodes.
- The nature of the task, i.e. whether the task is single-label or multi-label.

In the following sections, a brief review on most common applications of the automatic text classification will be seen.

### **2.5.1 Automatic indexing for Boolean Information Retrieval Systems**

Organizing text documents based on their contents, called indexing, is an important process in an information retrieval system. Basically, indexing is performed by assigning each document with keywords or descriptive terms representing the document. The assigned terms must reflect the content of the document to allow effective keyword searching. In the past, indexing has been done manually by trained persons who are familiar with the topics of the texts [Adriani et.al., 1997].

These key words or descriptive terms belong to a finite set called controlled dictionary, often consisting of a thematic hierarchical thesaurus (e.g. the NASA<sup>3</sup> thesaurus for the

---

<sup>3</sup> NASA: National Aeronautics and Space Administration: the US government organization that is responsible for space travel and the scientific study of space.

aerospace discipline, or the MeSH<sup>4</sup> thesaurus for medicine). In the past, Indexing has been done manually by trained persons who are familiar with the topics of the text.

Now if the entries in the controlled vocabulary are viewed as categories, text indexing is an instance of text classification, and may thus be addressed by the automatic classification techniques.

The issue of automatic indexing with controlled dictionaries is closely related to the topic of automated metadata generation. In digital libraries one is usually interested in tagging documents by metadata that describe them under a variety of aspects (e.g. creation date, document type or format, availability, etc.). Usually, some of these metadata are thematic, i.e. their role is to describe the semantics of Machine Learning in Automated Text Categorization the document by means of bibliographic codes, keywords or key phrases. The generation of these metadata may thus be viewed as a problem of document indexing with controlled dictionary, and thus tackled by means of text classification techniques [Sebastiani, 2002].

### **2.5.2 Document organization**

Indexing with a controlled vocabulary is one instance of the general problem of document base organization. In general, many other issues pertaining to document organization and filing, be it for purposes of personal organization or structuring of a corporate document base, may be addressed by text classification techniques. For instance, at the offices of a newspaper incoming “classified” advertisements must be, prior to publication, categorized under the categories used in the scheme adopted by the newspaper; typical categories might be Personals, Cars for Sale, Real Estate, etc.

---

<sup>4</sup> MeSH: Medical Subject Heading, US National Library of Medicine



While most newspapers would handle this application manually, those dealing with a high volume of classified advertisements might prefer an automatic system to choose the most suitable category for a given ad. In this case a typical constraint is that exactly one category is assigned to each document. Similar applications are the organization of patents into categories for making their search easier the automatic filing of newspaper articles under the appropriate sections (e.g. Politics, Home News, Lifestyles, etc.), or the automatic grouping of conference papers into sessions.

### **2.5.3 Text filtering**

Text filtering is the activity of classifying a dynamic collection of texts, i.e. a stream of incoming documents dispatched in an asynchronous way by an information producer to an information consumer. A typical case is a news feed, where the producer is a news agency (e.g. Reuters or Associated Press) and the consumer is a newspaper. In this case the filtering system should block the delivery to the consumer of the documents the consumer is likely not interested in (e.g. all news not concerning sports, in the case of a sports newspaper). Filtering can be seen as a case of single-label categorization, i.e. the classification of incoming documents in two disjoint categories, the relevant and the irrelevant. Similarly, an e-mail filter might be trained to discard “junk” mail and further classify non-junk mail into topical categories of interest to the user.

### **2.5.4 Word sense disambiguation**

Word Sense Disambiguation (WSD) is the process of identifying which sense of a word is used in any given sentence, when the word has a number of distinct senses. For instance, the English word ‘bank’ may have (at least) two different senses, as in the

Bank of England (a financial institution) or the bank of river Thames (a hydraulic engineering artifact). It is thus a WSD task to decide to which of the above senses the occurrence of 'bank' in last week, "I borrowed some money from the bank", refers to. WSD is very important for a number of applications, including natural language understanding, or indexing documents by word senses rather than by words for IR purposes.

Quite obviously this is a single-label categorization case and one in which document-pivoted categorization is most likely to be the right choice. WSD is just an example of the more general issue of resolving natural language ambiguities, one of the most important problems in computational linguistics [Sebastiani, 2002].

### **2.5.5 Hierarchical categorization of web pages**

Automatic document categorization has recently aroused a lot of interest also for its possible Internet applications. One of these is automatically classifying Web pages, or sites, into one or several of the categories that make up the commercial hierarchical catalogues hosted by popular Internet portals. When Web documents are catalogued in this way, rather than addressing a generic query to a general purpose Web search engine a searcher may find it easier to first navigate in the hierarchy of categories and then issue the search from (i.e. restrict the search to) a particular category of interest. Automatically classifying Web pages has obvious advantages, since the manual categorization of a large enough subset of the Web is infeasible. Web page text categorization applications, the automatic categorization of Web pages has two essential peculiarities:

(1) The hyper textual nature of the documents: hyperlinks constitute a rich source of information, as they may be understood as statements of relevance of the linked page to the linking page.

(2) The hierarchical structure of the category set: this may be used e.g. by decomposing the classification problem into a series of smaller classification problems corresponding each to a branching decision at an internal node.

## CHAPTER THREE

### LITERATURE REVIEW ON

#### MACHINE LEARNING (ML) APPROACH TO TEXT CLASSIFICATION (TC)

##### 3.1 INTRODUCTION

Machine learning refers to a system capable of the autonomous acquisition and integration of knowledge. This capacity to learn from experience, analytical observation, and other means, results in a system that can continuously self-improve and thereby offer increased efficiency and effectiveness.

Historically, the first attempts at automated classification involved the manual generation of deterministic rules that assigned a document to a category. This approach was fraught with problems, amongst them the chronic lack of scalability of the rule set and maintenance issues required to deal with generalizing to new documents and categories. In short, the rule approach required too much work for too little gain. As such, a move to a less maintenance heavy formalism was required. Modern classification approaches now employ Machine Learning techniques [Robinson H, 2003].

Learning is acquiring knowledge of a subject or skill in an art as a result of study, experience or teaching. Machine learning is a subfield of artificial intelligence that is concerned with the design and development of algorithms and techniques that allow computers to "learn". In general, there are two types of learning: **inductive**, and **deductive**. Inductive machine learning methods extract rules and patterns out of massive data sets, while deductive learning involves presenting a generalization and then seeking or providing examples.

The major focus of machine learning research is to extract information from data automatically, by computational and statistical methods. Hence, machine learning is closely related not only to data mining and statistics, but also to theoretical computer science.

It is a very familiar fact that there are a plethora of documents available on-line.

This has motivated fruitful research into many areas like data mining and the semantic web. One such area of research falling under natural language processing (NLP) is that of automated text categorization. Here, one uses a machine to categorize a given document according to one or more of its properties. In this chapter the basic concepts of machine learning, the need for machine learning, text representation and feature selection in machine learning are discussed.

### **3.2 The Importance of machine learning**

The importance of machine learning can be understood from the following facts:

- Relationships and correlations can be hidden within large amounts of data. Machine Learning/Data Mining may be able to find these relationships.
- Human designers often produce machines that do not work as well as desired in the environments in which they are used, because certain characteristic of the environment may not be known at start, ML methods can be used for on-job improvement.
- The amount of knowledge available about certain tasks might be too large for explicit encoding by humans (e.g., medical diagnostics).

- Environments change over time. Machine that can adapt to the changing environment may reduce for the need for constant re-design.
- New knowledge about tasks is constantly being discovered by humans. It may be difficult to continuously re-design systems “by hand”.

Until the late 1980s, the most effective approach to TC seemed to be that of manually building automatic TC systems by means of knowledge-engineering techniques that is manually defining a set of logical rules that encode expert knowledge on how to classify documents under the given set of categories. In the 1990s this perspective has been overturn, and the machine learning paradigm to automated TC has emerged and definitely superseded the knowledge-engineering approach. Within the machine learning paradigm, a general inductive process automatically builds an automatic text classifier by “learning”, from a set of previously classified documents, the characteristics of the categories of interest. The advantages of this approach are accuracy comparable to human performance and a considerable savings in terms of expert manpower, since no intervention from either knowledge engineers or domain experts is needed. Current day TC may thus be seen as the meeting point of machine learning and information retrieval (IR), the “mother” of all disciplines concerned with automated content-based document management [Sebastiani,1999].

### **3.3 Procedures in automatic document classifications**

Almost every popular classifier accepts as input a feature vector that characterizes the document to be classified, as well as those that were the classifier is trained on. The construction of this feature vector is very important to the successful operation of the

classifier. This step in automatic document classification is known as document preprocessing.

Indexing is an information retrieval operation concerned in assigning appropriate terms and identifiers capable of representing the content of the collection items. This task normally performed manually by trained experts. In relatively modern environments the indexing task is performed automatically [Salton G., 1983].

Not all words are equally significant for representing the semantics of a document. Therefore, it is usually considered worthwhile to preprocess the text of the document collection to determine the terms to be used as index terms. During this preprocessing phase other useful text operations can be performed such as elimination of stop words, stemming (reduction of a word to its grammatical root), the building of a thesaurus, and compression [Beza-Yates, 1999]

Text classification system can be roughly distinguished in three different phases in the life cycle of a text classification system, which have traditionally tackled in isolation of each other (i.e. a solution to one problem not being influenced by the solutions given to the other two): document indexing, classifier learning, and classifier evaluation [Sebastiani, 2005].

### **3.3.1 Document indexing**

Document indexing denotes the activity of mapping a document  $d_j$  in to a compact representation of its content that can be directly interpreted by a classifier building algorithm and then by the classifier itself once it has been built.

The document indexing methods usually employed in text classification are borrowed from Information Retrieval (IR), where a text  $d_j$  is typically represented as a vector of term weights

$$d_j \text{ vector} = (w_{1j}, w_{2j}, w_{3j}, w_{4j}, \dots, w_{|t|j}) \quad (3.1)$$

Document preprocessing is a set of procedures in document indexing phase of the classification system which can be divided mainly into five text operations (or transformations), these are seen briefly in following sub sections:

### **3.3.1.1 Lexical Analysis of Text and Document Representation**

Lexical analysis is the process of converting a stream of characters (the text of the documents) into a stream of words (the candidate words to be adopted as index terms). Thus, one of the major objectives of the lexical analysis phase is the identification of the words in the text.

Lexical analysis of the text deals with the objective of treating digits, hyphens, punctuation marks, and the case of letters. All markup tags and special formatting are removed from the document. Thus, for an HTML document all tags and text inside these are removed. This normally would include all element attributes, scripts, comment lines and text placed into these documents [Beza-yates, 1999].

Documents can be considered as a stream of characters. However, for the problem of automatic classification these streams should be transformed into representatives, which are suitable for the process of classification.

In fact, word-based representation has been found very effective in information retrieval and text classification. They are the basis for most work in text classification [JoachimsT., 2002]



A substantial advantage of word-based representation is their simplicity. It is relatively easy to design algorithms that efficiently decompose text into words. A simple algorithm that splits a string into words by white space characters will usually produce satisfactory results for example, for English language.

Ignoring logical structure and layout, using words as document representation transforms a document from a string of characters into a sequence of words. In addition it is usually assumed that the ordering of the words is irrelevant (or at least of minor importance) for the classification task. So, a word sequence can be projected onto a bag of words. Only the frequency of a word in a document is recorded, while all structure of the document ignored. This representation is called the bag-of-words approach [Joachims T., 2002].

#### **3.3.1.2 Elimination of the stop words**

Words which are too frequent among the document in the collection are not good discriminators. In fact, a word which occurs in 80% of the documents in the collection is useless for purpose of index term. Articles, prepositions, and conjunctions are natural candidates for a list of stop words.

Elimination of stop words is the objectives of filtering out words with very low discrimination values or the high-frequency function words which comprises 40 to 50 percent of the text words. These words are poor discriminator and can not be used by themselves to identify document content. Some times called stop list or negative dictionary [Beza-Yates, 1999]

#### **3.3.1.3 Stemming**

Stemming of words is the removing of affixes and allowing the retrieval of documents containing syntactic variations of terms.

It refers to the process of reducing terms to their stems or root variant. (eg. Connect for connects, connecting, connected connection connector etc...). A stem is portion of a word which is left after the removal of its affixes [Beza-Yates, 1999]

Referring to Frakes, [Beza-Yates, 1999] explains that there are four types of stemming strategies: Table lookup, successor variety, n-grams and affix removal;

**Table lookup** consists of simply of looking for the stem of a word in a table. It is a simple procedure but one which is dependent on data on stems for the whole language. Since such data is not readily available and might require considerable storage space, this type of stemming algorithm might not be practical.

**Successor variety** stemming is based on the determination of morpheme boundaries, uses knowledge from structural linguistics, and is more complex than affix removal stemming algorithms.

**N-grams stemming** is based on the identification of digrams and trigrams and is more a term clustering procedure than a stemming one.

**Affix removal** is the most important part is suffix removal because most variant of a word are generated by the introduction of suffixes instead of prefixes.

#### **3.3.1.4 Selection of the index terms**

If a full text representation of the text is adopted then all words in the text are used as index terms. The alternative is to adopt more abstract view in which not all words are used as index terms. This implies that the set of terms used as indices must be selected.

Usually, the decision on whether a particular word will be used as an index term is related to the syntactic nature of the word. In fact, noun words frequently carry more semantics than adjectives, adverbs, and verbs. [Beza- Yates, 1999]

### 3.3.1.5 Term Weighting

After the word stems are generated, it becomes necessary to recognize equivalent stems occurring in the texts and to choose stems to be used as index terms.

The bag-of-words representation makes text accessible to most machine learning algorithms. These algorithms require that each example (document) be described by a vector of fixed dimensionality. Each component of a vector represents the value of one attribute of the example. Commonly, each word  $\omega$  is treated as one such attribute. The value of an attribute for a particular document  $\mathbf{d}$  can be, for example, the number of times it occurs in the document. This quantity is called the term frequency,  $\mathbf{tf}$ , of a word  $\omega$  in document  $\mathbf{d}$ .

The idea of term frequency is that terms occurring more often in a document are more relevant than infrequent terms. However, a frequent term may occur in almost every document of a collection. Such terms will not discriminate between classes. The collection component is designed to assign lower weight to such terms. Its basic statistic is document frequency  $\mathbf{df}$ , it is the number of document in which the word  $\omega$  occurs at least once. If the document frequency is high, the weight of the term is reduced, and also, document can be of different length. Therefore, a normalizing component is supposed to adjust the weights so that small and large documents can be compared on the same scale. A high standard of performance at modest cost is obtainable by using the inverse document frequency function  $\mathbf{idf}$ , which is the total number of documents in a collection by the number of documents in which the word occur.

Weighting is the final stage in most IR indexing applications. Terms are weighted according to a given weighting model which may include local weighting global weighting or both. If local weights are used, then term weights are normally expressed as term frequencies,  $tf$ . If global weights are used, the weight of a term is given by **idf** values. The most common (and basic) weighting scheme is one in which local and global weights are used (weight of a term =  $tf * idf$ ). This is commonly referred to as **tf \* idf** weighting.

[Beza-Yates, 1999]

In order to represent a document by a numeric vector, we need a function that assigns a weight factor to each of the keywords chosen in the preprocessing step. The weight factor should represent the importance of the keyword for the classification of the document. Thus all keywords that do not appear in the document should have a weight of 0. The simplest method is Boolean weighting, where the keywords that do not appear in the document are assigned the weight of 0, and all keywords that appear in the document are assigned the weight 1. Obviously, this method discards valuable information contained in the frequency of appearance of the keywords – a keyword appearing once has the same weight as the one appearing ten times in the document. A better method should use this information and assign a higher weight to the keywords that appear more. Another factor that can be used is the frequency of keyword appearance in other documents. If a keyword appears in a large number of documents, it is likely to be a more general word that is less useful for classification. A word appearing in a smaller number of documents is likely to be more typical and representative, and should therefore have a higher weight.

These considerations gave rise to the most popular weight determination method, the term frequency – inverse document frequency (**tf-idf**) function as discussed above.

Accordingly, the formula to compute the weight of a word  $\omega$  in a document  $d$  is given by:

$$\omega_{ik} = f_{ik} * \log (N/n_i) \quad (3.2)$$

Where:

- $\omega_{ik}$  is the weight of term  $i$  in the  $k^{\text{th}}$  document
- $f_{ik}$  is the frequency of the  $i^{\text{th}}$  term in the  $k^{\text{th}}$  document
- $N$  is the number of documents in the collection
- $n_i$  is the number of documents in which the  $i^{\text{th}}$  term occurs

### 3.3.2 Classifier Learning

As it has been said so far, automatic classification of documents using machine learning approach requires the learning process to be initiated by supplying the examples labeled with their class-category from which the systems starts to learn.

The essential idea is to infer a classifier (i.e. a rule that decides whether or not a document should be assigned to a category) from a set of labeled documents (i.e. documents with known category assignments). Standard statistical classification tools such as Naive Bayes, logistic regression, and decision trees are immediately relevant and have been used with some success [Sebastiani,2002].

These examples which are analyzed to build a model are collectively referred to as an example documents or train sets.

The classification system then analyzes the statistical occurrences of each concept in the example documents and constructs a model or classifier for each category that is used to classify the subsequent documents automatically. The system refines its model, in a sense “Learning” the categories as new documents are processed.

After a classifier has been built, it is desirable to evaluate its effectiveness. A set on which the built classifier effectiveness is observed is referred to as test set.

Referring to different researchers [Surafel , 2003] mentioned that the size of training set and test is not necessarily of equal size. Accordingly the following researchers use 20% [McCallum & Nigam,1998], 30%[Koller & Sahami,1997] or 33%[Joachims, n.d] of data for test set and the remaining for training set respectively.

According to [Sebastiani, 2002], the training set is inductively built by observing the characteristics of the documents. In most research settings, once a classifier has been built it is desirable to evaluate its effectiveness. A test set is used for testing the effectiveness of the classifier. Each document from the test set is fed to the classifier, and the classifier decisions are compared with the expert decisions. The documents in test set can not participate in any way in the inductive construction of the classifiers; otherwise, the experiment results obtained would likely be unrealistically good, and the evaluation is considered not scientific.

The goal of classification is to build a set of models that can correctly predict the class of the different objects. The input to these methods is a set of objects (i.e., training data), the classes which these objects belong to (i.e., dependent variables), and a set of variables describing different characteristics of the objects (i.e., independent variables). Once such a predictive model is built, it can be used to predict the class of the objects for which class information is not known a priori. The key advantage of supervised learning methods over unsupervised methods (for example, clustering) is that by having an explicit knowledge of the classes the different objects belong to, these algorithms can perform an effective feature selection if that leads to better prediction accuracy.

Machine learning algorithms are organized into taxonomy, based on the desired outcome of the algorithm. Common algorithm types include:

- Supervised learning — in which the algorithm generates a function that maps inputs to desired outputs. One standard formulation of the supervised learning task is the classification problem: the learner is required to learn (to approximate) the behavior of a function which maps a vector  $(x_1, x_2, x_3, \dots, X_n)$  into one of several classes by looking at several input-output examples of the function.
- Unsupervised learning (or clustering) — an agent which models a set of inputs: labeled examples are not available and the number of set of classes to be learned may not be known in advance.

### **3.3.2.1 Inductive Learning and algorithms**

Induction learning is reasoning from observed training cases to general rules, which are then applied to the test cases. In contrast to transduction or transductive inference which is reasoning from observed specific (training) cases to specific (test) cases.

The goal in inductive learning is to infer a general classification rule from a sample of labeled training documents. This classification rule should classify new examples with high accuracy. Inductive text classification is a two step process. In the first step the learner uses the training data to induce a classification rule. This step is commonly called the learning phase. The second step is the classification phase; the classification rule is repeatedly used to classify new examples [Joachims T., 2002].

In inductive learning the learner tries to induce a decision function which has a low error rate on the whole distribution of examples for the particular learning task. It wants to

learn an unknown function  $f(x) = y$ , where  $x$  is an input example and  $y$  is the desired output. A classifier is a function that maps an input attribute vector to a given class with a certain confidence. The problem to be solved is generally perceived as a function that maps some input to a reaction or classification ( $f: \mathbf{x} \rightarrow \mathbf{y}$ ). The goal of training is to approximate the real Function  $f$ , That means a learning system that was presented some examples of houses, should learn the patterns, which all (or most) houses have in common.

Inductive learning methods require a certain number of training examples to achieve a given level of generalization accuracy. Decision tree, k-nearest neighbors, neural networks and Bayesian learning algorithms are examples of inductive learning algorithms that require large data [Mitchell S., 1997]

### **3.3.2.2 Support vector machines(SVMs)**

Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. Viewing input data as two sets of vectors in an  $n$ -dimensional space, an SVM will construct a separating hyper planes in that space, one which maximizes the margin between the two data sets. To calculate the margin, two parallel hyper planes are constructed, one on each side of the separating hyper plane, which is “pushed up against” the two data sets. Intuitively, a good separation is achieved by the hyper plane that has the largest distance to the neighboring data points of both classes, since in general the larger the margin the better the generalization error of the classifier.



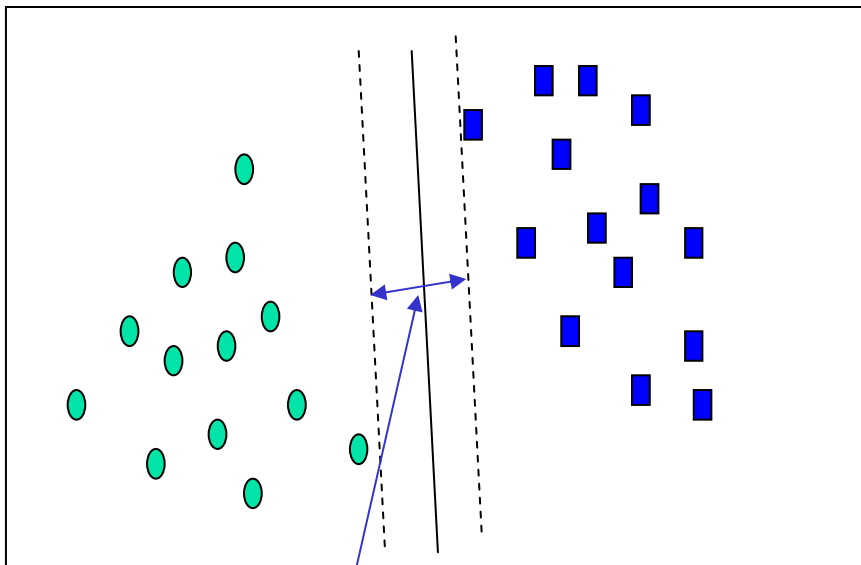
SVMs are very universal learners. In their basic form, SVMs learn linear threshold function. Nevertheless, by a simple "plug-in" of an appropriate kernel function, they can be used to learn polynomial classifiers, radial basic function (RBF) networks, and three-layer sigmoid neural nets. One remarkable property of SVMs is that their ability to learn can be independent of the dimensionality of the feature space. SVMs measure the complexity of hypotheses based on the margin with which they separate the data, not the number of features. This means that we can generalize even in the presence of very many features, if our data is separable with a wide margin using functions from the hypothesis space [Joachims T., 1997].

In Order to understand the general philosophy of the SVM, leaving out the mathematical<sup>5</sup> details, consider only 2-dimensional example, suppose some given data points each belong to one of two classes (Figure 3.1). One category of the target variable is represented by rectangles while the other category is represented by ovals, and the goal is to decide which class a new data point will be in. In the case of support vector machines, a data point is viewed as a  $p$ -dimensional vector, where  $p$  is the number of attributes. In this example, as 2-dimensional vector, and we want to know whether we can separate such points with a  $(p - 1)$ -dimensional hyperplane (in this example since  $p$  is taken to be 2, 1-dimensional hyperplane is a straight line). This is called a linear classifier. There are many hyperplanes (lines in this case for 2 dimensions) that might classify the data. However, we are additionally interested in finding out if we can achieve maximum separation (margin) between the two classes.

---

<sup>5</sup> Mathematical details for SVMs can be found in [Han and Kamber, 2006] pages 337 - 344

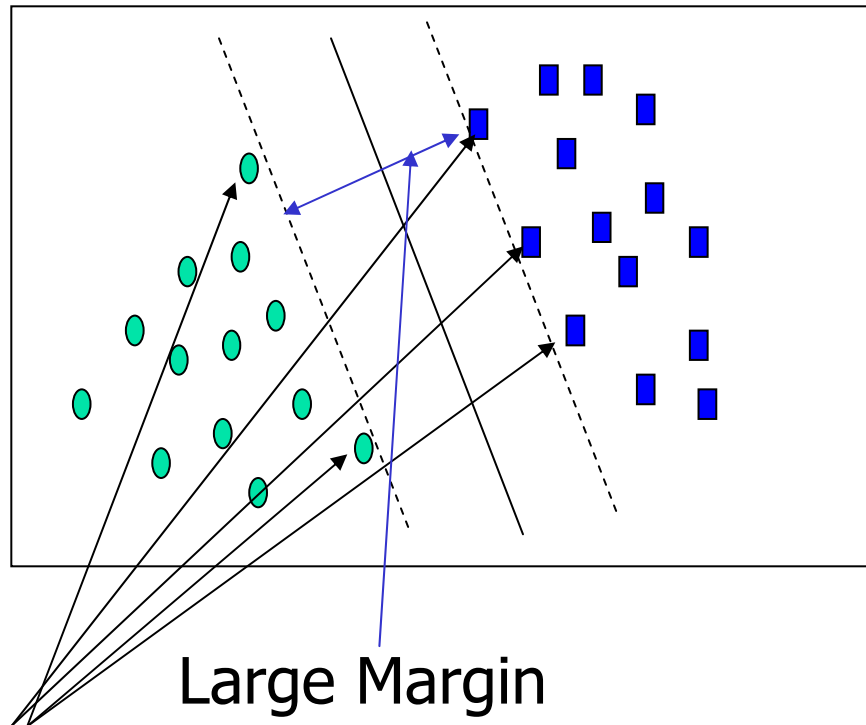
By this we mean that we pick the hyperplane so that the distance from the hyperplane to the nearest data point is maximized. That is to say that the nearest distance between a point in one separated hyperplane and a point in the other separated hyperplane is maximized. Now, if such a hyperplane exists, it is clearly of interest and is known as the **maximum-margin hyperplane** and such a linear classifier is known as a maximum margin classifier.



Small Margin

*Figure 3.1 SVM Small Margin*

The dashed lines drawn parallel to the separating line mark the distance between the dividing line and the closest vectors to the line. The distance between the dashed lines is called the margin. The vectors (points) that constrain the width of the margin are the support vectors.



## Support Vectors

*Figure 3.2 SVM Large Margin*

In this idealized example, the cases with one category are in the lower left corner (oval) and the cases with the other category are in the upper right corner (squares); the cases are completely separated. The SVM analysis attempts to find a 1-dimensional hyperplane (i.e. a line) that separates the cases based on their target categories. There are an infinite number of possible lines; two candidate lines are shown (Figure 3.1 and Figure 3.2) above.

An SVM analysis finds the line (or, in general, hyperplane) that is oriented so that the margin between the support vectors is maximized. In the figures above, the line in figure 3.2 is superior to the line in the figure 3.1

If all analyses consisted of two-category target variables with two predictor variables, and the cluster of points could be divided by a straight line, life would be easy. Unfortunately, this is not generally the case, so SVM must deal with:

- (a) More than two predictor variables,
- (b) Separating the points with non-linear curves,
- (c) Handling the cases where clusters cannot be completely separated,
- (d) Handling classifications with more than two categories.

If we add a third predictor variable, then we can use its value for a third dimension and plot the points in a 3-dimensional cube. Points on a 2-dimensional plane can be separated by a 1-dimensional line as seen above. Similarly, points in a 3-dimensional cube can be separated by a 2-dimensional plane. As we add additional predictor variables (attributes), the data points can be represented in N-dimensional space, and a (N-1)-dimensional hyperplane can separate them.

The simplest way to divide two groups is with a straight line, flat plane or an N-dimensional hyperplane. But in reality, there are cases when clusters cannot be completely separated linearly, or need the consideration of non-linear regions. Look at figure 3.3 below, in which case we need a nonlinear dividing line.

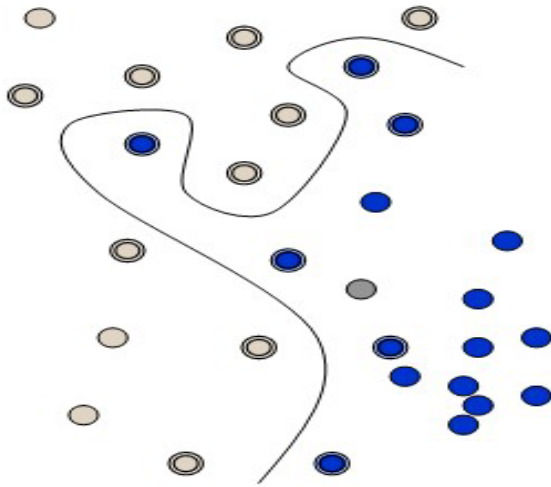


Figure 3.3 Non-Linear region

In these cases, rather than fitting nonlinear curves to the data, SVM handles this by using a kernel function to map the data into a different space where a hyperplane can be used to do the separation.

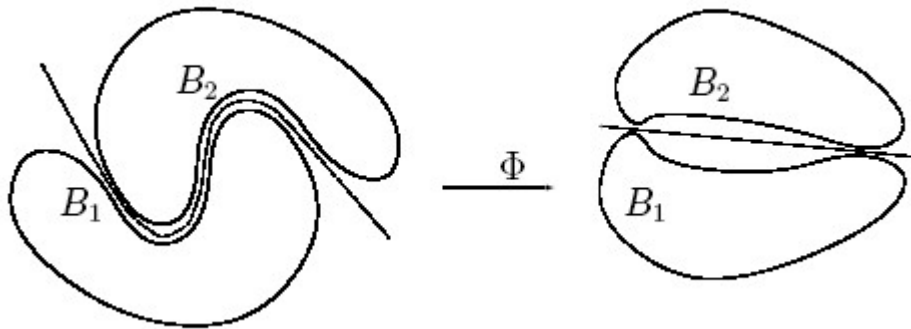


Figure 3.4 Kernel function transform data

The kernel function may transform the data into a higher dimensional space to make it possible to perform the separation.

The approach described for linear SVM can be extended to create nonlinear SVMs for the classification of linearly inseparable data (also called non-linearly separable data, or nonlinear data, for short). Such SVMs are capable of finding nonlinear decision boundaries (i.e. nonlinear hypersurfaces) in input space [Han and Kamber,2006 ]

### 3.3.2.3 Bayesian Classification

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given sample belongs to a particular class [Han and Kamber,2004]

The underlying rule for Bayesian text classification is the Bayes' theorem and its classifier models include naïve Bayesian classifier and Bayesian belief networks.

**Bayes Theorem** states:

For a given data sample  $X$  whose class label unknown, let  $H$  be some hypothesis such that the data sample  $X$  belongs to a specified class  $C$ , for classification problem, we want to determine  $P(H | X)$ , the probability that the hypothesis  $H$  holds given the observed data sample  $X$ .

$$P(H/X) = \frac{P(X/H) * P(H)}{P(X)} \quad (3.3)$$

Naïve Bayesian (NB) classifier assumes that the effect of an attribute value of a given class is independent of the values of the other attributes.

Naïve Bayesian algorithm has been widely used for document classification, and shown to produce very good performance.

The basic idea is to use the joint probabilities of words and categories to estimate the probabilities of categories given a document. Naïve Bayesian algorithm computes the posterior probability that the document belongs to different classes and assigns it to the class with the highest posterior probability. The posterior probability of class is computed using Bayes rule and the testing sample is assigned to the class with the highest posterior probability.

The naive part of naïve Bayesian algorithm is the assumption of word independence that the conditional probability of a word given a category is assumed to be independent from the conditional probabilities of other words given that category.

There are two generative models of Naïve Bayesian algorithm. One is the multi-variate Bernoulli event model that only takes into account the presence or absence of a particular term, so it doesn't capture the number of occurrence of each word. A document is represented by a vector of binary attributes indicating which words occur and do not occur in the document. The number of times a word occurs in a document is not captured. The other model is the multinomial model that captures the word frequency information in documents. A document is represented by the set of word occurrences from the document in which the number of occurrences of each word in the document is captured [McCallum et.al., 1998].

Bayesian belief networks are graphical models, which unlike naïve Bayesian classifiers allow the representation of dependencies among subset of attributes.

Among the advantages of the Bayesian classifiers is that simple technique results in high accuracy, especially when combined with other methods. The limitations when using naïve bayes is the independent and equally important assumption which may cause skewed results, especially if many of the variables are interrelated, as that relation will have a greater effect on the decision, for better or for worse. Naïve bayes classification does not allow for categorical output attributes.

Naïve Bayes classifiers have been proven in many domains, especially in text categorization, despite the simplicity of the model and restrictiveness of the independence assumptions it makes [Peng F. et.al., 2003].

The Naive Bayes algorithm calculates the conditional probability  $P(C_i|X)$ , where  $C_i$  is the class value and  $X$  is an instance of the sample. The  $X$  will be classified to the  $C_i$  if  $P(C_i|X)$  is the maximum one of all class values.

According to Bayes rule equation (3.3) , 
$$P(C_i | X) = \frac{P(X | C_i)P(C_i)}{P(X)}$$
 , as  $P(X)$  is constant, and assuming all the predictors are independent each other, the problem converts to

calculate the maximum  $P(X|C_i)*P(C_i)$ , which is equal to  $(\prod_{k=1}^n p(x_k | C_i))P(C_i)$  .

Bayesian classifiers are also useful in that they provide a theoretical justification for other classifiers that do not explicitly use Baye's theorem [Han and kamber,2006 ].



### 3.3.2.4 K-nearest Neighbor Classifiers(KNN)

The k-nearest-neighbor method was first described in the early 1950s. The method is labor intensive when given large training sets, and did not gain popularity until the 1960s when increased computing power become available [Han and Kamber, 2006].

KNN is an Instance-based learner. Instance-based learners are also referred to as lazy learners; because they wait until the test set is supplied, unlike eager learners.

Eager learners, when given a set of training tuples, will construct a generalization (i.e. classification) model before receiving a new (e.g. test) tuples to classify. We can think of the learned model as being ready and eager to classify previously unseen tuples. However, Lazy learners wait until the last minute before doing any model construction in order to classify a given test tuple. That is, when given a training tuple, a lazy learner simply stores it (or does only a little minor processing) and waits until it is given a test tuple. Only when it sees the test tuple does it perform generalization in order to classify the tuples based on its similarity to the stored training tuples [Han and Kamber, 2006]

KNN classifier is a learning algorithm that is based on a distance function for pairs of observations, such as the Euclidean distance. In KNN, “Closeness” is defined in terms of a distance metric used. The Euclidean distance between two points or tuples, say,

$\mathbf{X}_1 = (x_{11}, x_{12}, x_{13}, \dots, x_{1n})$  and  $\mathbf{X}_2 = (x_{21}, x_{22}, x_{23}, \dots, x_{2n})$ , is

$$\text{Dist}(\mathbf{X}_1, \mathbf{X}_2) = \sqrt{\sum_{i=1}^{i=n} (x_{1i} - x_{2i})^2} \quad (3.3)$$

In other words, for each numeric attribute, we take the difference between the corresponding values of that attribute in tuple  $\mathbf{X}_1$  and in tuple  $\mathbf{X}_2$ , square this difference, and accumulate it. The square root is taken of the total accumulated distance count.

For K-nearest classification, the unknown tuple is assigned the most common class among its k-nearest neighbors. When  $k = 1$ , the unknown tuple is assigned the class of the training tuple that is closest to it in pattern space.

In this classification paradigm, k nearest neighbors of a training data is computed first. Then the similarities of one sample from testing data to the k nearest neighbors are aggregated according to the class of the neighbors, and the testing sample is assigned to the most similar class.

One of advantages of KNN is that it is well suited for multi-modal classes as its classification decision is based on a small neighborhood of similar objects (i.e., the major class). So, even if the target class is multi-modal (i.e., consists of objects whose independent variables have different characteristics for different subsets), it can still lead to good accuracy. A major drawback of the similarity measure used in KNN is that it uses all features equally in computing similarities [Yancong Z. et.al., 2001].

### **3.3.2.5 Classification by decision tree induction**

A decision tree is a predictive model; that is, a mapping from observations about an item to conclusions about its target value. More descriptive names for such tree models are classification tree (discrete outcome) or regression tree (continuous outcome). In these tree structures, leaves represent classifications and branches represent conjunctions of features that lead to those classifications. The machine learning technique for inducing a decision tree from data is called decision tree learning or decision trees.

A decision tree is a flow-chart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions. The top most node is the root node [Han and Kamber, 2006].

In order to classify an unknown sample, the attribute values of the sample are tested against the decision tree. Testing begins from the root, decisions are made at the branches and class prediction is arrived at the leaf node. Decision tree classifiers can handle high dimensional data and have good accuracy. A path is traced from the root to a leaf node that holds the class prediction for that sample. Consequently, decision trees can easily be converted to classification rules.

Decision trees have been used in many application areas ranging from medicine to game theory and business. They are the basis of several commercial rule induction systems [Han and Kamber, 2006].

The tree starts as a single node representing the training sample. If the sample are all of the same class, then the node becomes a leaf and is labeled with that class, otherwise the algorithm uses an entropy-based measure known as **information gain** as a heuristic for selecting the attribute that will best separate the samples into individual classes. This attribute becomes “test” or “decision” attribute at the node. A branch is created for each known value of the test attribute, and the samples are partitioned accordingly. The algorithm uses the same process recursively to form a decision tree for the samples at each partition. Once an attribute has occurred at a node, it need not be considered in any of the node’s descendants. The recursive partitioning stops only when any one of the following conditions is true:

- a) All samples for a given node belong to the same class.
- b) There are no remaining attributes on which the samples may be further partitioned.
- c) There are no more samples for the branch.

The growing of Decision Tree from data is a very efficient technique for learning classifiers. The selection of an attribute used to split the data set at each Decision Tree node is fundamental to properly classify objects; a good selection will improve the accuracy of the classification. When a decision tree is built, many of the branches will reflect anomalies in the training data due to noise or outliers. Tree pruning methods address this problem; attribute selection and tree pruning will be discussed next.

### **I. Attribute Selection**

Attribute selection measure helps to find the best partition by selecting the criteria that best separates a given data partition of class-labeled training data instances into individual classes.

Attribute selection method provides a ranking for each attribute describing the given training instances. The attribute having the best score for the measure is chosen as the splitting attribute for the given instances. Different attribute selection measures are used in decision tree induction including information gain, gain ratio, and Gini index. The decision tree algorithms compute the information gain of each attribute.

The attribute with the highest information gain is chosen as the test attribute for the given set. A node is created and labeled with the attribute, branches are created for each value of the attribute, and the samples are partitioned accordingly.

## II. Tree pruning

When a decision tree is built, many of the branches will reflect anomalies in the training data due to noise or outliers. Tree pruning methods address this problem of over fitting the data. Such methods typically use statistical measures to remove the least reliable braches. Generally, this results in faster classification and improvement in the ability of the tree to correctly classify independent test data that is not seen before.

There are two common approaches to tree pruning. They are prepruning and postpruning. In the **prepruning** approach, a tree is pruned by halting its construction early. If partitioning the samples at a node would result in a split that falls below a pre specified threshold, then further partitioning of the given subset is halted. In this case upon halting, the node becomes a leaf. The leaf may hold the most frequent class among the subset samples or the probability distribution of those samples.

The second approach, **postpruning** , removes braches from a “ fully grown” tree. A sub-tree at a given node is pruned by removing its branches and replacing it with a leaf. The lowest unpruned node becomes a leaf and is labeled by the most frequent class among its former branches.

### **III. Extracting classification rules from Decision trees**

The knowledge represented in decision trees can be extracted and represented in the form of classification IF-Then rules. One a rule is created for each path from the root to a leaf node. Each attribute-value pair along a given path forms a conjunction in the rule antecedent (IF part). The leaf node holds the classification prediction, forming the rule consequence (Then part). The IF-THEN rules may be easier for humans to understand, particularly when the given tree is very large [Han and kamber, 2004].

Among the advantage of decision trees are, they are easy to understand, easily converted to a set of production rules, and can classify both categorical and numerical data, but the output attribute must be categorical. The limitations that are mentioned on decision trees are, decision tree algorithms are unstable. Slight variation in the training data can result in different attribute selections at each choice point within the tree. The effect can be significant since attribute choices affect all descendent sub trees. Trees created from numeric data sets can be quite complex since attribute splits for numeric data are binary.

#### **3.3.2.6 Major Challenges of text classifications**

Text categorization faces two particular challenges. First, the scale of text categorization applications causes problems for many standard learning algorithms. Documents are represented by vectors of numeric values, with one value for each word that appears in any document. Document feature vectors therefore are typically of high dimension.

High dimensionality both increases processing time and increases the risk of over fitting, i.e. that the learning algorithm will induce a classifier that reflects accidental properties of the particular training examples rather than systematic relationships between the words and categories [Eyheramendy et al., 2003].

The second challenge is how to incorporate human understanding of the categorization problem into the learning process. For instance, people with a need for text categorization almost always have some sense of words that would be good predictors for each category. Textual descriptions of the category content, or just the category name itself, can also provide clues for human. However, in machine learning, we need greater number of labeled training data to induce knowledge for classification for the machine.

The result is to increase the expense of using text categorization, since larger amounts of training data must be labeled. Further, the most interesting categories in intelligence applications often have few known example documents, so unless prior knowledge can be combined with these documents, it may not be possible to learn a classifier with good effectiveness [Eyheramendy et al., 2003].

Therefore, the choice of learning algorithm for text classification considers both the challenge of high dimensionality and the challenge of integrating knowledge with learning.

Referring to different researchers, [Eyheramendy et al., 2003] mentioned the following:

The challenge of high dimensionality led early work in text categorization to focus on learning algorithms that were both computationally efficient (for speed) and very restricted in the classifiers they could produce (to avoid overfitting). These include Naive

Bayes [Lewis, 1998] and the Rocchio algorithm [Rocchio, 1971]. More recently, increased computing power and a better theoretical understanding of classifier complexity have enabled algorithms to learn less restricted and thus more accurate classifiers while simultaneously avoiding overfitting and maintaining sufficient speed. Examples include support vector machines [Joachims, 1998], [Lewis, 2002], and [Lewis et al., 2003] and ridge logistic regression [Zhang and Oles, 2001].

Document classification is a domain with a large number of attributes. The attributes of the examples to be classified are words, and the number of different words can be quite large indeed. While some simple document classification tasks can be accurately performed with vocabulary sizes less than one hundred, many complex tasks on real-world data from the Web, UseNet and newswire articles do best with vocabulary sizes in the thousands. Naive Bayes has been successfully applied to document classification in many research efforts [McCallum et.al., 1998]

This research is about automatic classification of text documents, specifically about Afaan Oromoo news items classifications using machine learning approach. Text documents are represented by many attributes (word vectors), therefore learning algorithms which entertain high dimensionality such as SVM, Bayesian classifiers, Decision trees are experimented with. Learner algorithms which proves the application of machine learning approach to the Afaan Oromoo news items classification with better accuracy was reported.



### **3.3.3 Classifier Evaluation**

There are many learning algorithms which are useful for text classification purpose as discussed above.

Classification methods can be compared and evaluated according to the following criteria [Han and Kam, 2004].

- Predictive accuracy: This refers to the ability of the model/algorithm to correctly predict the class label of new or previously unseen data.
- Speed: This refers to the computation costs involved in generating and using the model.
- Robustness: This is the ability of the model to make correct predictions given noisy data or data with missing values.
- Scalability: This refers to the ability to construct the model efficiently given large amounts of data.
- Interpretability: This refers to the level of understanding and insight that is provided by the model.

#### **3.3.3.1 Classifier Performance measures**

According to [Sebastiani,2002] The evaluation of document classifier is typically conducted experimentally, rather than analytically. The reason is that, in order to evaluate a system analytically (e.g. providing that the system is correct and complete), we would need a formal specification of the problem that the system is trying to solve (e.g. with respect to what correctness and completeness are defined), and the central notion of text classification (namely, that of membership of a document in a category) is, due to its subjective character, inherently non formalizable.

The experimental evaluation of a classifier usually measures its effectiveness (rather than its efficiency), that is, its ability to take the right classification decisions.

The category assignment of a binary classifier can be evaluated using a confusion matrix. Confusion matrix is a tool for analyzing how well a classifier recognizes instances of different classes.

Category $C_i$		Classifier Judgments	
		Yes	No
Expert Judgments	Yes	t-pos	f-neg
	No	f-pos	t-neg

*Table 3.1: Two Class Confusion Matrix*

Where:

t-pos: is true positives

t-neg: is true negatives

f-pos: is false positive

f-neg: is false negative

In the table t-pos refers to the instances which are classified positive by the classifier as same as the judgment of an expert. f-neg are instances classified as negative by the classifier but which are positive by the expert judgment.

f-pos are instances classified as positive by the classifier but are actually negative by the expert judgment. t-neg are instances classified as negative by the classifier which are also negative by the expert judgment.

A measure of classification effectiveness is based on how often the classifier decision values match the expert decisions. Classification effectiveness is usually measured in terms of the classical IR notions of precision and recall, adapted to the case of text classification [Sebastiani,2002].

In this study classifiers are used for automated text classification, i.e. for binary classification tasks. Therefore, the following focuses on binary classification performance measures.

The accuracy of a classifier is estimated from the classifier's performance on a test data and is the percentage of test set instances that are correctly classified by the classifier.

In general the class assignment of a binary classifier can be evaluated using a confusion matrix which is a tool for analyzing how well a classifier recognizes instances of different classes.

The table above shows possible outcomes of a binary classifier. That is a two-class confusion matrix which shows positive instances (documents of the class of interest, i.e., 'yes' for class) versus negative instances ('no' for class). A perfect classifier would have a value of 0 for f-pos and f-neg, that is, it assigns positive instances to positive and negative instance to negative with out error.

Using the confusion matrix we define the three performance measures common in document categorization literature [ Rafael et al. ,2004]

These are Recall, Precision and F-score. Recall measure contain information about whether classification errors are dominated by f-negative. Precision measure contains information about whether the classification errors are dominated by f-pos. The trade-off between Recall and Precision can be controlled by setting classifier parameters. Both measures should typically be used to describe the overall performance, as neither is particularly informative by itself. The third measure F-score is an average R and P.

**Recall (R)** is the percentage of the documents for a given category that are classified correctly.

$$R = t\text{-pos} / (t\text{-pos} + f\text{-neg}) \quad (3.4)$$

**Precision (P)** is the percentage of the predicted documents for a given category that are classified correctly.

$$P = t\text{-pos} / (t\text{-pos} + f\text{-pos}) \quad (3.5)$$

**F-Score (F)** is an average of R and P given by

$$F = 2PR / P + R \quad (3.6)$$

Classification accuracy is also the other method of measure of performance represented by  $c / n$  where  $n$  is the total number of test instances and  $c$  is the number of test instances correctly classified by the system [Sebastiani, 2002].

$$\mathbf{Accuracy} = (t\text{-pos} + t\text{-neg}) / (t\text{-pos} + t\text{-neg} + f\text{-pos} + f\text{-neg}) \quad (3.7)$$

Accuracy (error rate) is the rate of correct (incorrect) predictions made by the model over a data set. The average results of accuracy can also be represented in confusion matrix form.

$$\text{Sensitivity (Recall)} = t\text{-pos} / \text{pos} \quad (3.8)$$

Where t-pos is true positive and pos is the number of positive documents

$$\text{Specificity} = \text{t-neg} / \text{neg} \quad (3.9)$$

Where t-neg is true negative and neg is the number of negative instances

$$\text{Precision} = \text{t-pos} / (\text{t-pos} + \text{f-pos}) \quad (3.10)$$

Therefore, in terms of sensitivity, specificity and precision accuracy will be:

$$\text{Accuracy} = \text{Sensitivity} * (\text{pos}/(\text{pos} + \text{neg})) + \text{Specificity} * (\text{neg} / (\text{pos} + \text{neg})) \quad (3.11)$$

### **ROC (Receiver Operating Characteristics) curves**

An ROC curves shows the trade-off between the true-positive rate or sensitivity (proportion of positive tuples that are correctly identified) and the false-positive rate (proportion of negative tuples that are incorrectly identified as positive) for a given model. That is, given a two-class problem, it allows us to visualize the trade-off between the rate at which the model can accurately recognize 'yes' cases versus the rate at which it mistakenly identifies 'no' cases as 'yes' for different 'portions' of the test set. Any increase in the true-positive rate occurs at the cost of an increase in the false-positive rate. The area under the ROC curve is a measure of the accuracy of the model. To assess the accuracy of the model, we can measure area under the ROC curve. The closer the area is to 0.5 the less accurate the corresponding model is. A model with perfect accuracy will have an area of 1.0 [Han and Kamber, 2006].

### 3.3.3.2 Estimating Classifier Accuracy

Using training data to derive a classifier and then to estimate the accuracy of the classifier can result in misleading overoptimistic estimate due to overspecialization of the learning algorithm (or model) to the data. Holdout and cross-validation are common techniques for assessing classifier accuracy, based on randomly sampled partitions of the given data [Han and Kamber, 2004].

In the **holdout** method, the given data are randomly partitioned into two independent sets, a training set and test. Typically, two thirds of the data are allocated to the training set, and the remaining one third is allocated to the test set. The training set is used to derive the classifier, whose accuracy is estimated with the test set. The estimate is pessimistic since only a portion of the initial data is used to derive the classifier.

**Random sub-sampling** is a variation of the holdout method in which the holdout method is repeated  $k$  times. The holdout accuracy estimate is taken as the average of the accuracies obtained from each of the iteration.

In **k-fold cross-validation**, the initial data are randomly partitioned in to  $k$  mutually exclusive subsets or folds, each of approximately equal size. Training and testing is performed  $k$  times. Unlike the random sampling method here each partition is used equal number of times for training and once for testing. Classification accuracy is estimated by dividing the overall number of correct classification from the  $k$  iterations by the total number of instances. In stratified cross-validation, the folds are stratified so that the class distribution of the sample in each fold is approximately the same as that in the initial data.

In general, stratified 10-fold cross-validation is recommended for estimating classifier accuracy (even if computation power allows using more folds) due to its relatively low bias and variance [Han and Kamber, 2006].

In this research 10-fold cross-validation is used for all experiments.

### **3.4 Summary**

In this chapter, machine learning approach to text classification has been discussed. Importance of machine learning was mentioned briefly. Inductive Learning algorithms in general and learner algorithms used in this thesis in particular are presented. It has been indicated that, to apply automatic classifiers for the classification purpose, the source documents need to pass through a number of pre-processing steps. The source documents are first represented by content bearing words. Content bearing words are further selected (feature selection) based on some parameters for example using words frequency (tf), and inverse document frequency (idf). The weight or importance of these feature in a collection will be calculated (for example using equations 3.2), which changes the features in to numerical values. The document is then represented by these vectors of feature words numerically. Using these vectors of terms with their pre-classified categories, the learning algorithm will be trained and its performance will be evaluated.

## CHAPTER FOUR

### THE AFAAN OROMOO LANGUAGE

#### 4.1 INTRODUCTION

Afaan Oromoo is the language of the Oromo people who comprises the largest ethnic group in Ethiopia. The 2007 census report of the Central Statistics Office for instance, estimated Oromos to be 25,489,024 (about 35% of the total population of Ethiopia).

Afaan Oromo is an official language of the Oromiya regional state and is the medium of instruction for the primary schools in the region. Currently there are a growing number of publications in hard copies and vast amount of information in electronic formats for Afaan Oromo.

According to [Stroomer, 1987] , quoted in [Wakshum, 2000] Afaan Oromo linguistically, belongs to the Cushitic branch of the Afro-Asiatic language family along with Somali, Afar and a number of other languages.

The adoption of the Latin script for Afaan Oromo writing system was made officially since 1991. This was made for the convenience of the Latin script for the writing of Afaan Oromo from the linguistics, pedagogic, and practical reasons [Tilahun, 1994]. It is believed that many fold more text were written in Afaan Oromo since then than before.

Afaan Oromo is a language that is used in a wide area in the country; as a result there is a dialectical variation. According to [Gragg, 1976], quoted in [Wakshum, 2000], four major categories can be identified. These are:

Western (Wellega, Iluababor, Kaffa and parts of Gojjam), Eastern (Harar, Eastern showa, and parts of Arsi and Bale), Central (Central Showa, Western Showa and possibly Wollo) and Southern ( Parts of Arsi, Sidamo and Borena).



## **4.2 The Oromo Alphabet**

The alphabets of Afaan Oromo is often called “Qubee Afaan Oromoo”, alphabets of the Oromo language. The major representatives of sources of the sound in a language are the vowels and consonants.

Afaan Oromo has 36 basic sounds (10 vowels and 26 consonants) [Wakshum, 2000]. Afaan Oromo is a phonetic language, which means that is spoken in the way it is written.

The Afaan Oromoo vowels represented by letters (a, e, o, u and i) are called “Dubbiiftuu” in Afaan Oromo and the consonants known as “dubbifamaa” in Afaan Oromo are shown in the following tables 4.1 and 4.3 respectively, together with their main articulators.

According to [Ladefoged,1955], quoted in [Morka, 2001] Some of the finer anatomical feature involved in speech production include the vocal cords, velum, tongue, teeth, palates, the alveolar ridge, the mouth, and lips. These anatomical components move to different positions to produce various sounds and are referred in articulators. Most of the characters of a sound are determined by the position of these articulators in the oral tract.

### **4.2.1 The Afaan Oromo Vowels**

Afaan Oromo basically has 10 phonemic vowels, five short and five long vowels, indicated in the orthography by doubling the five vowel letters. These vowels (Table 4.1) can be divided into different categories depending how they are formulated: Front/back indicates the position of tongue that give sound in mouth, while High/Low is place of the

tongue with respect the palate (the top part of the inside of the mouth) during articulation.

Vowel can appear in initial, medial and final positions in a word in Afaan Oromo language. A long vowel is interpreted as a single unit and occurs everywhere a short vowel can occur.

The following examples show some of long vowels at word initial, medial and final positions.

Initial positions: **uumaa** to mean ‘nature’, **eelee** to mean ‘pan’,

Medial position: **keenna** to mean ‘gift’, **leexaa** to mean ‘single’

Final position: **garaa** to mean ‘belly’, **daaraa** to mean ‘ash’

The difference in length is contrastive, for example consider, ‘lafa’ in Afaan Oromoo which is to mean 'land', and ‘laafaa’ in Afaan Oromoo which is to mean 'weak'. The difference between the words ‘lafa’ and ‘laafaa’ is the length of vowel they have. Two vowels in succession indicate that the vowel is long (called “Dheeraa” in Afaan Oromoo), while a single vowel in a word is short (called “Gababaa” in Afaan Oromoo).

#### Afaan Oromoo Vowels

	Front	Central	Back
High	<b>i , ii</b>		<b>u, uu</b>
Mid	<b>e, ee</b>		<b>o, oo</b>
Low		<b>a, aa</b>	

*Table 4.1 The Afaan Oromo vowels*

Afaan Oromo vowels are pronounced in sharp and clear fashion which means each and every word is pronounced strongly.

For example:

A: Fardda, Haadha

E: Gannale, Waabee, Noole, Roobale, colle

I: Arsii, laali, Rafi, Lakki, Sirbbi

O: Oromo, Cilaalo, Haro, caancco, Danbidoollo

U: Ulfaadhu, Gudadhu, dubadhuu, arbba guugu, Ituu

[Morka, 2001] referring to [Mongham, n.d] , In order to produce sounds, the oral tract must involve an **active articulator**, which is raised to form the stricture, as well as a **passive articulator** towards which the active articulator is raised.

The major places of articulations, Table 4.2 below, taken from [Morka, 2001], is the description of places of sound production system together with their active and passive articulators.

Place	Active Articulator	Passive Articulator
Bilabial	Lower lip	Upper lip
Labio-dental	Lower lip	Upper teeth
Dental	Tip of tongue	Upper teeth
Alveolar	Blade of tongue	Alveolar ridge
Retroflex	Tip of tongue	Hard palate
Palatal	Front of tongue	Hard palate
Velar	Middle of tongue	Velum (Soft Palate)
Uvular	Back of tongue	Uvula

*Table 4.2 Major places of articulations*

#### 4.2.2 The Afaan Oromo Consonants

The Afaan Oromo consonants are shown in table 4.3 below. All Afaan Oromo consonants except the combination consonants ny, dh, ph, and sh have double consonant combinations if the syllable is stressed. Failure to make this distinction results in miscommunication. For examples; the word “Walqixumma”, which is to mean ‘Equality’ is different from “Walqixuma” which is “it is equal”.

		Bilabial / Labiodental	Alveolar / Retroflex	Palato-Alveolar / Palatal	Velar	Glotal
Stops and	Voiceless	( p )	t	ch	k	?
	Voiced	b	d	j	g	
Affricates	Ejective	ph	x	c	q	
	Implosive		dh			h
Fricatives	Voiceless	f	s	sh		
	Voiced	( v )	( z )			
Nasal		m	n	ny		
Approximants			l	y		
Flap / Trill		w	r			

*Table 4.3 The Afaan Oromo consonants*

The consonants p,v, and z only occur in loan words.

**Gemination** happens when a spoken consonant is pronounced for an audibly longer period of time than a short consonant.

Gemination is also significant in Afaan Oromo. That is, consonant length can distinguish words from one another;

for example, the word “baruu” in Afaan Oromo is to mean ‘to learn’ is ungeminated (called ‘laafaa’ in Afaan Oromo) , whereas the word “barruu” in Afaan Oromoo is to mean ‘palm of hand’ is geminated (called ‘jabaa’ in Afaan Oromo). The difference between the two words “baruu” and “barruu” is the number of consonants appearing together, which makes a difference in meaning.

Some Afaan Oromo words are pronounced with the stress on the last syllable. Examples: fanno, harre, gaarre.

On the other hand, few words are stressed on the first syllable. These words always have a combination consonant.

Examples: Nyaadhu, Nyaara , Dhugaa,

### **4.3 Morphology**

Natural Language Processing is the application of computational models to tasks involving human language text. Automatic text classification therefore requires the studying of the morphology of the language on which it applies.

In every language, whether it is spoken or written, every meaningful pattern has its own structure and the elements of language of language related to each other in understandable manner. Words are the basic elements of a language and are formed from morphemes which constitute the smallest meaningful unit of speech in a language. This is also true for Afaan Oromoo which has its own rules of words and/or sentence structure [Assefa, 2005].

Morphology is a way of studying the language words structure. It is about the way words are put together, their internal structure. It is a level at which the structure of language is analyzed. It deals with the analysis and examination of meaningful units of forms which make up sentences. The smallest meaningful units of forms are called "morphemes" which are either "free" or "bound." A free morpheme can occur on its own whereas bound morphemes do not occur alone. They accompany free morphemes in the formation of words and constituents. The English forms like Book, man, etc., are free morphemes and -ish (as in Bookish) and -ly (as in manly) are bound morphemes. Bound morphemes are of three types, these are, prefix attached to the initial positions, infix inserted in the middle and suffix attached to the final position of the word. All the three types of bound morphemes are called "affixes".

In linguistics agglutination is the morphological process of adding affixes to the base of a word. Languages that use agglutination widely are called agglutinative languages.

While document representation for classification, different forms of the same words due to number, tenses, gender, etc may need to be reduced in to its stem or root word (For example the words connects, connected , connecting, connection etc... may need to be reduced to the word connect). Therefore the consideration of the morphology of the language during the automatic classification process helps.

The two main branches of morphology are **inflectional** and **derivational** morphology

[Richard Hudson, 2003]

### 4.3.1 Inflectional morphology

Inflectional Morphology studies the inflectional changes in words that generally do not result in changing the classes of words. Rather, the inflection changed indicates tense (present, past, far past, future), number (singular, plural), gender/class (masculine, feminine, neuter), person (first, second, third) etc ...

Examples:

Present → Past                      look → looked

Present → Present continuous    look → looking

Singular → Plural                    car → cars

Male → Female                        actor → actress

First, second person → third person    give → gives

From the example above, Inflectional morphology deals with the combination of stems grammatical makers of suffixes such as –s , -ess, -ed and –ing in English language.

Generally inflectional morphology is very productive as all nouns have singular/plural distinctions; most verbs have tense distinctions, etc... [Wakshum, 2000]

### 4.3.2 Derivational Morphology

Derivational morphology results in changing classes of words, for example, a noun(n) may be derived from a verb(v), or adjective(adj) can be derived from a verb. Examples;

Create (v) + ive (adjective maker) → creative (adj)

Creative (adj) + ity (noun maker) → creativity (n)

Nation (n) + al (adjective maker) → national (adj)

National (adj) + ize (verb maker) → nationalize (v)

Nationalize (v) + ation (noun maker) → nationalization (n)

#### 4.4 The Afaan Oromo Morphology

Every language has its own morphological structure that defines rules used for combining the different components the language may have. The English language for instance is basically different in its morphological structure from French, Arabic or Afaan Oromoo [Wakshum, 2000].

There are a number of word formation processes in Afaan Oromoo. Affixation and compounding are among these word formation processes.

**Affixation** is generally described as the addition of affixes at the beginning, in between and/or at the end of a root/stem depending on whether the affix is prefix, infix or suffix. Attaching one or more prefixes and/or suffixes to a stem may form a word. The word *durbumma* ‘girlhood’ for instance is formed from the stem *durb-* ‘girl’ and the suffix – *umma*.

**Compounding** is the joining together of two linguistic forms, which functions independently. Examples compound nouns include; *abbaa-buddenaa* ‘step father’ from *abba-* ‘father’ and *buddena* ‘food’.

Like a number of other African and Ethiopian languages, Afaan Oromo has a very complex and rich morphology. It has the basic features of agglutinative languages involving very extensive inflectional and derivational morphological processes. In **agglutinative languages** like Afaan Oromo, most of the grammatical information is conveyed through affixes (i.e. prefixes and suffixes) attached to the root or stem of words. Obviously, these high inflectional forms and extensive derivational features of the language are presenting various challenges for text processing and information retrieval experiments in Afaan Oromo [Kula et al, 2007].



Although, Afaan Oromo words have some prefixes and infixes, suffixes are the predominant morphological features in the language. Almost all Oromo nouns in a given text have persons, number, gender and possession makers which are concatenated and affixed to a stem or singular noun form. In addition, Afaan Oromo noun plural markers/forms can have several alternatives. For instance, in comparison to the English noun plural marker s(-es), there are more than ten major and very common plural markers in Afaan Oromo including: -oota, -wwan, -lee, -an, -een, -eeyyii, -oo, etc..). As an example, the Afaan Oromo singular noun “mana” (house) can take the following different plural forms: Manoota (mana + oota), manneen (mana + een), manawwan (mana + wwan. The construction and usages of such alternative affixes and attachments are governed by the morphological and syntactic rules of the language [Kula et al, 2007].

## **CHAPTER FIVE**

### **AUTOMATIC CLASSIFICATION OF AFAAN OROMOO TEXTS**

#### **5.1 INTRODUCTION**

This chapter presents data source for the research, the data processing carried out, the testing procedures followed for the automatic classification of Afaan Oromo news items, and the results of the experiment on the thesis. The preprocessing tasks are performed on the basis of the concepts discussed in earlier chapters. The performance of the automatic classifiers used in classifying Afaan Oromo news items will be shown and compared on various sets of categories.

As it has been indicated, most of the pre-processing tasks that will be performed on the natural language is that language dependent. The pre-processing for the automatic classification of the Afaan Oromo text starts from the identification of each individual word in a text, that is, it involves word-level processing of the source dataset with the ultimate aim of identifying feature words that are representatives of the document dataset.

Figure 5.1 is the description of Machine Learning architecture for text document classification and the summaries of main activities performed in the thesis for automatic classification process. The major components are data pre-processing, classifier construction and document classification. The sub-sections of this chapter present the activities performed at each component in more detail manner.

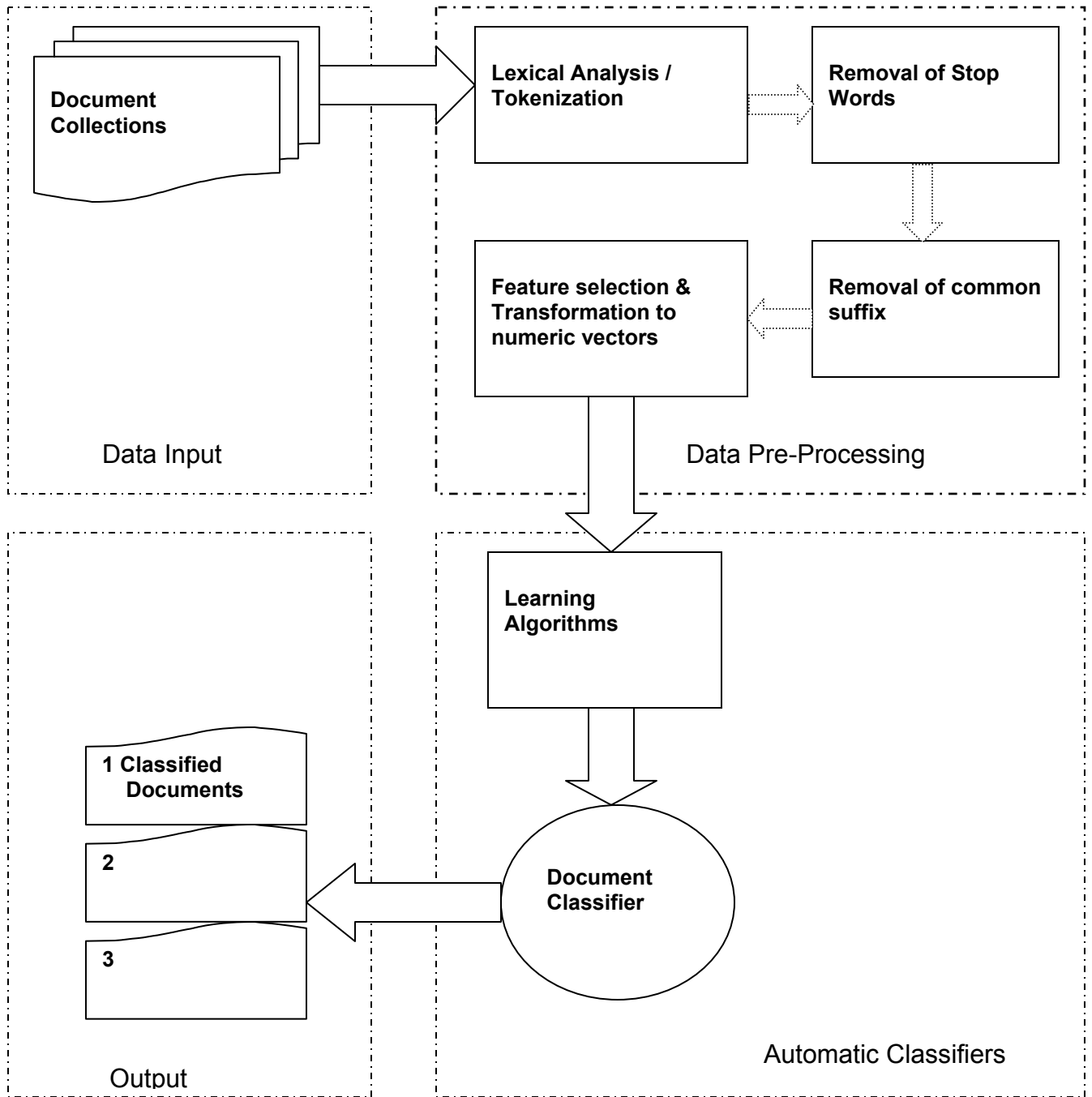


Figure 5.1 Machine Learning text classification Architecture

## **5.2 Data Source**

The data source for this research is the electronic Afaan Oromo news items collected from the Radio Fana Share Company. As it has been discussed in section (1.2) the company is a multi-lingual radio station. The choice for the company is the availability of the electronic news items for the Afaan Oromo. The news items from April 2006 to October 2008 are used as data source for the experiment of this thesis.

The data obtained was SQL server database which comprises the news items for both the Amharic and Afaan Oromo together, in to a single database.

The database was converted in to MS-Access database format for the ease of processing, and the Afaan Oromo news items which is required in this research is separated in to different database for further processing. As it has been discussed in section (1.4) the Radio Fana Share company uses a computerized electronic system for the efficient processing of the news items, which was once considered as best. However, Radio Fana uses a manual classification system. As it has been discussed in chapter two, Manual classification involves human experts to classify documents based on classification rules and subjective judgment. There are 11 main categories (Table 5.1) in to which the news editors of the company, has to assign the news items in order to save it to the database.

No	Category Code	Category Name
1	1	Infrastructure
2	2	Sport
3	3	Economy
4	4	Social Affairs
5	5	National Politics
6	27	Law and Justice
7	31	Associations (Organizations)
8	34	Defense and Security
9	37	Others(Different) Affairs
10	40	Foreign related news
11	52	Governmental organizations

*Table 5.1 Main News Categories*

In addition to the main categories, Radio Fana has also sub-classifications under each of the main classifications, however only the main class classifications are considered in this research.

### **5.3 Creating the Afaan Oromo news items database**

The focus of this study is the Afaan Oromo news items at Radio Fana. Therefore, it was necessary to create the Afaan Oromo database from the main database of the company. To do this the researcher used the information “StoryOwner” field of the news story table in main database and the information, names of the Afaan Oromo editors since the beginning of the current system to date. A program which reads the main database of the company selects the news items which are prepared by the Afaan Oromo news editors and writes it to the newly created Afaan Oromo database, is

developed by the researcher to split the Afaan Oromo news items from the main database.

The database then created for the Afaan Oromo news items from the main database contained 3,710 news items in the news story table. Creation of the Afaan Oromo news items database alone is not the only thing that makes the data ready for classification. The data was impossible to use as it was, because it creates a problem for the experimentation process. Hence, manual data cleaning was done by the researcher in consultation with the editors of the news items at the company before further pre-processing steps. Duplicate records entered to the database more than once were dropped except for one record. There were records which are blank, or do not give meaning in Afaan Oromo language. This was because as there are required fields in the news editing software currently used, but the editors simply putting any thing they like to this fields in order to save the news items to the database; also the editors some times types any character to the database for the checking purposes. Such records were removed during the preprocessing phases. Samples of kinds of records removed are given at appendix 1.

The number of the news items left after the first instance manual cleaning for the data is 2,268 Afaan Oromo news items comprising all the main categories. The researcher created a table in the Afaan Oromo database for each of the mentioned main categories and assigned each of the news item left after the first instance manual cleaning to its respective categories by a program developed by the researcher using the category code of the news items which are given by the company.

The number of news items for each category obtained for further processing and the table name for each category which is created in the database by the researcher is shown at table 5.2 below. The sample of news items head lines in the database is presented in appendix 2.

No	Category Code	Category Name	Table Name	Number of News items
1	1	Infrastructure	tblInfra	70
2	2	Sport	tblSpo	63
3	3	Economy	tblEco	415
4	4	Social Affairs	tblSoc	106
5	5	National Politics	tblPol	255
6	27	Law and Justice	tblJus	71
7	31	Associations (Organizations)	tblOrg	139
8	34	Defense and Security	tblDef	369
9	37	Others(Different) Affairs	tblDif	59
10	40	Foreign related news	tblFor	415
11	52	Governmental organizations	tblGov	306

*Table 5.2 Number of news items by categories*

#### **5.4 Data Preparation**

The news story table in the database has 27 fields for each news item, though some of the fields are not used at all. For example “TimeNewsCreated” , “Location” and “WICTimeStamp” fields are not used at all and some fields are used as needed by the editors, for example “NewsSource”, “AllowedUsers” fields are rarely used. Therefore, all the news item fields are not considered equally important currently at the company. The researcher found out that, fields considered most important and basic are the following:

#### **5.4.1 The Header of the news items**

These are the fields which give a description about each news items such as the news id, date the news created, the story owner and the classification codes.

#### **5.4.2 The summary of the news items**

These are the fields which gives the summarized descriptions about the full story of the news items. The headline field is a one or two sentence that pictures about the full story in a concise way. The Key words and the slug fields contain words and phrases considered specific to that news items.

#### **5.4.3 The body of the news items**

This is a full story of the news items. It contains every necessary detail about the topic elaborating on the head line. Every part of it explains the information contained in the headline from its source, location, time, actor, date, etc... point of views.

In most cases the body of the items contains details of the news such as the name of a place, person, or office, the time or date when the event took place, and the numeric amount of an event or a thing mentioned in the headline [Surafel, 2003].

The headline, keywords and slug fields were used for the further processing towards the Afaan Oromo news items classification process in this research. These words are assumed to most discriminate the categories of the news items. The words in the body of the news items are the elaborators of the words, and therefore may not discriminate the categories of the news items.



## **5.5 News text Pre-Processing for classification**

As it has been detailed in Section (3.3.1) automatic text classification requires pre-processing steps. In natural language processing for classification, language dependent text preprocessing should be performed before the implementation of the automatic classification; these are the steps at which the text is made available to the learning algorithms by removal of non-informative words or characters from the text to be classified.

### **5.5.1 Tokenization**

As it has been discussed in section (3.3.1.1) lexical analysis or tokenization is the process of segmenting a string of characters into words, and words are the bases for document representation. In Afaan Oromoo writing system strings that are delimited by white spaces is a word token.

The following algorithm from [Zelalem, 2001] is used to identify words and is adapted in Afaan Oromo text.

- 1. Initialize the variable to hold the word*
- 2. Read a character from the sentence(document)*
- 3. Check if the character is Afaan Oromoo word delimiter*
- 4. If not, concatenate the character to the variable,*
- 5. Else if the number of characters is above two character report the word*
- 6. If there is more data to process go to step 1*

After analyzing the Afaan Oromoo news items, the researcher found that the words whose character length is less than or equal to two characters can not be a candidate word to represent the document. For example the words “ti”, “ol”, “fi” etc... are ignored

during the feature selection for classification purpose. The cases of all the words in the document are converted in to the same lower case. Similarly the digits in texts are not accepted as a good representation terms for documents and therefore, are not taken as words. For example the number like 20ffaa to mean '20<sup>th</sup>', 2000tii to mean 'in 2000' is not taken as a valid word. In general extraneous characters which don't help as good representation of a document are removed. The implementation of these activities is done in visual basic programming.

### **5.5.2 Removal of the stop words**

Stop words are the words which are very common in every document, discussed in section (3.3.1.2). The words which occur frequently in almost all documents are non informative for a document representation in preparing for the automatic classification system. Therefore a list of stop words in Afaan Oromo, particularly news specific stop words are compiled during data preparation and saved to the stop words list table in a database. Like the English language prepositions "the" , "and", "is" etc... which are frequent, Some words in Afaan Oromo are also very frequent in a document, for example "tti", "tiin", "itti" etc... are common in a document. In addition to this there are news specific stop words like "ibsame", "ibسانی", "beeksisee", "beeksisaan" , "jedhan" etc... to mean 'notify', 'notified' etc... The list of the stop words used in this research is shown in appendix 3.

These stop words are ignored in the documents during the feature selection process for the automatic classification of the documents.

An algorithm for stop words removal from a given documents

1. *Get the next word until the last word in the document*
2. *Check the word against the stop words list*
3. *If not a word exists in the stop words list then write it as a candidate for document representation*
4. *Else drop it*
5. *Go to step 1.*

Implementation of the algorithm is done using visual basic programming language.

### **5.5.3 Word Stemming**

Varies forms of the same word may appear in a document, this may be due to number, tense, gender and parts of speech agreement in a sentence for a specified language. Stemming is discussed in section (3.3.1.3) of Chapter 3. This is a language dependent as it requires the morphology of the given language. Stemming by itself is a big task in research area. It has been seen from the literature that a stemmer for Afaan Oromo text is developed by [Wakshum, 2000], however despite the effort made to get the full stemmer, it was not found. Afaan Oromo is a morphologically complex and rich language through which most of the grammatical information is conveyed through affixes, especially suffixes discussed in section (4.4). In order to incorporate the advantage of simple stemming in automatic text classification, the researcher developed a suffix removal tool for some of the suffixes observed in the Afaan Oromo news items to aid the stemming process. The stripping off the suffixes is done on the suffixes identified not to change the linguistic meaning of the words in the news items, for otherwise a detailed study of the morphology is required, which has a time constraint on

the researcher. For example in a word “akkoowwan” to mean ‘grandmothers’ the suffix “wwan” can be stripped to give “akko” to mean ‘grandmother’, in a word “lammiileen” to mean ‘citizens’ the suffix “leen” can be stripped to give “lammi” to mean ‘citizen’, similarly, in word “biyyalee” to mean ‘countries’ the suffix “lee” can be stripped to give “biyya” to mean ‘country’ without changing the linguistic meaning of the words. The procedure refers to the suffixes list table in the database to strip off suffixes from the words in a document.

#### **5.5.4 Data transformation and scaling**

After representative words preparation, a document is treated as a collection of words, bag of words, which are the candidate representatives of a given document. In bag of words representation the relative position of words are not used. The feature which represents a document is prepared in a format that is used by the application package, i.e. the comma separated value (CSV) or Arff (Attribute relation file format). The application package accepts the minimum threshold frequency, which is 2 in this research, a point at which good accuracy is observed, to select feature words. Feature words now become the attributes of the documents, then the frequency of a word in a document (tf), in how many documents in a collection the word appears, inverse document frequency (idf) are used to calculate the weight of each word in a document, and normalization to this calculation is done to find how good the word represent the document both with respect of the document itself and with respect to the document collections, the details of term weighting is discussed in section (3.3.1.5) of chapter 3. The features which represent the document are then changed to word vector which represent the document.

## **5.6 The news item classification process**

In this study the Weka application package is used to classify the Afaan Oromo news documents. Weka is open source software developed by the University of Waikato in NewZealand. The choice for this package is that, the package is freely available, it provides many different algorithms for machine learning and the whole package is written in Java, so it can be run on any platform.

## **5.7 Testing classifier algorithms**

Testing the classifier algorithms is important because it allows evaluating how reliably a given classifier will label future data, that is, data on which the classifier has not been trained. Classifier evaluation has been discussed in general in section (3.3.3).

In this research the performance of selected automatic text classifiers, for the application of Afaan Oromoo news items classification is tested. For the reasons discussed in section (3.3.2.6) and to make closer observation on the characteristics of the automatic classifiers on each and every category, Sequential Minimal Optimization (SMO) algorithm from Support Vector Machine (SVM), BayesNaiveMultiNominal algorithm from Bayesian Classifiers are used in the experiment. To ensure the application of machine learning approach to the Afaan Oromo text classification, J48 algorithm from Decision trees and KNN from lazy classifiers was also employed in the experiment for performance evaluation and comparison. All the selected classifiers were compared on the same data and a set of categories.

In order to observe the effects of the number of categories and the number of instances on the learning algorithms, the researcher performed the experiment with different number of categories stage by stages. The testing started with data of four categories (Sport, Justice, Others (Different) and Infrastructure), seven categories (Sport, Justice, Others (Different), Infrastructure, Economy, Foreign and Defense & Security) and finally the all the eleven categories were tested together. The whole experiment dataset has two types of attributes, numeric attributes for the weight of the feature words and nominal attributes for the class labels. Regarding the training and test sets, in all the experiments 10-fold stratified cross validation is used.

### **5.7.1 Classification using Sequential minimal Optimization (SMO)**

Sequential Minimal Optimization (or SMO) is a fast method to train Support Vector Machines (SVMs). Training an SVM requires the solution of a very large quadratic programming (QP) optimization problem. SMO breaks this large QP problem into a series of smallest possible QP problems. These small QP problems are solved analytically, which avoids using a time-consuming numerical QP optimization as an inner loop. The amount of memory required for SMO is linear in the training set size, which allows SMO to handle very large training sets.

#### **5.7.1.1 Experiment on four categories**

The four classes 'Sport', 'Justice & Law', 'Others (Different)' and 'Infrastructure' were selected together for relatively having the same number of news items. The classifier performed with an average accuracy of 95.81%. The summary statistics shows:

Correctly Classified Instances	252	95.8175 %
Incorrectly Classified Instances	11	4.1825 %

The testing on the data is done using 10-fold stratified cross validation, Weka options for measuring the performance of a classifier, which gave the summary statistics shown above. The confusion matrix table 5.3 and detailed accuracy by class table 5.4 for SMO Classifier on the four categories of the experiment are shown below.

```

=== Confusion Matrix ===
  a  b  c  d  <-- classified as
70  0  1  0 | a = Justice
 5 65  0  0 | b = Infra
 3  1 55  0 | c = Different
 1  0  0 62 | d = Sport

```

*Table 5.3 The Confusion Matrix for the SMO Classifier on four categories*

The confusion matrix contains a row and a column where the row is the actual number of instances in the specified categories and the column is the predicted number of documents classified to the corresponding class by the classifier. For example the first row in the above table shows there are actually 71 documents in the justice & law class, out of which 70 are classified as justice & law class correctly (True Positive, TP) by the classifier while 1 document (False negative, FN) is classified as Others(different) class incorrectly by the classifier.

Looking down the column, for example, first column in the above table shows there are 79 documents predicted as Justice & Law class by the classifier, out of which 70 (True Positive, TP) predicted correctly by the classifier while 9 instances are (False Positive, FP) predicted as Justice & Law incorrectly by the classifier. The confusion matrix can be taken as a data source for measures used for calculating the detailed accuracy of each class as shown in table 5.4.

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.986	0.047	0.886	0.986	0.933	0.97	Justice
0.929	0.005	0.985	0.929	0.956	0.959	Infra
0.932	0.005	0.982	0.932	0.957	0.951	Different
0.984	0	1	0.984	0.992	0.997	Sport

*Table 5.4 Detailed accuracy by class for SMO classifier on four categories.*

Looking at the first column in detailed accuracy table 5.4, the True Positive rate (TP rate), that is, the instances classified correctly out of all instances of a category. For example, for the first row first column of the detailed accuracy table, TP rate for the Justice & law category, the number of instances classified correctly from the confusion matrix Table 5.3, is 70 and the total number of instances in a Justice & law category is 71, then TP rate will be  $70 / 71 = 0.986$

The False Positive rate, that is, the negative instances classified as positive out of all negative instances  $9 / (70+59+63) = 9 / 192 = 0.047$

From the detailed accuracy by class table, it can be seen that Justice & law class has the highest TP-rate of 98.6%, followed by the sport class 98.4%, the Others (Different) class 93.2% and the Infrastructure class 92.9%.

Considering only the TP-rate for measuring the performance of a classifier could be misleading. Therefore other measures like precision, the F-Measure (to optimize for either precision or recall), an ROC-area to see the trade-off between the true-positive (TP) rate or sensitivity (proportion of positive instances that are correctly identified) and the false-positive (FP) rate (proportion of negative instances that incorrectly identified as positive) which are also an important measures of performance are shown in the above detailed accuracy table. Classifier performance measures were covered in section 3.3.3.1 of chapter 3.



Considering precision, Recall, F-measure, and ROC-Area and observing the details of the class statistics in table 5.4, we can see that SMO algorithm can be applied for Afaan Oromo text classification of four categories.

### 5.7.1.2 Experiment on seven categories

The second experiment was performed on seven classes, the classes ‘Sport’, ‘Justice’, ‘Others(Different)’ and ‘Infrastructure’ which was tested in section 5.7.1.1 and the classes ‘Economy’, ‘Defense & Security’ and ‘Foreign related’ are used in an evaluation of the SMO algorithm. The testing on the data is done using 10-fold stratified cross validation. A total of 1,462 documents are used in the experiment. The summary statistics shows:

Correctly Classified Instances	1370	93.7073 %
Incorrectly Classified Instances	92	6.2927 %

```

=== Confusion Matrix ===
  a  b  c  d  e  f  g  <-- classified as
61  0  0  0  1  6  3 |  a = Justice
 0 56  0  0 10  4  0 |  b = Infra
 0  0 52  0  1  5  1 |  c = Different
 0  0  0 60  1  2  0 |  d = Sport
 0  0  2  0 395 13  5 |  e = Economy
 0  0  1  0  4 401  9 |  f = Foreign
 0  0  1  0  5 18 345 |  g = DefSecurity

```

*Table 5.5 The Confusion Matrix for the SMO Classifier on seven categories.*

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.859	0	1	0.859	0.924	0.943	Justice
0.8	0	1	0.8	0.889	0.949	Infra
0.881	0.003	0.929	0.881	0.904	0.948	Different
0.952	0	1	0.952	0.976	0.983	Sport
0.952	0.021	0.947	0.952	0.95	0.979	Economy
0.966	0.046	0.893	0.966	0.928	0.965	Foreign
0.935	0.016	0.95	0.935	0.943	0.964	DefSecurity

*Table 5.6 Detailed accuracy by class for SMO classifier on seven categories.*

Looking at the TP-Rate, the Precision, the F-measure, and the ROC-Area indicates that categories with relatively larger instance documents are classified more accurately than the other classes with lower instances (Justice, Infra and Different). Here we can see that SMO algorithm can be applied for Afaan Oromoo text classification purpose on seven categories.

### 5.7.1.3 Experiment on eleven categories

The third experiment was performed on eleven classes, the classes ‘Sport’, ‘Justice & Law’, ‘Others(Different)’, ‘Infrastructure’, ‘Economy’, ‘Defense & Security’, and ‘Foreign related’ which was tested in section 5.7.1.2 and the classes ‘Social affairs’, ‘Associations’, ‘National Politics’ and ‘Governmental organizations’ are used in an evaluation of the SMO algorithm. The testing on the data is done using 10-fold stratified cross validation. A total of 2,268 documents are used in the experiment. As it can be observed from table 5.2 of section 5.3 the number of news items in the eleven categories is not uniform.

The summary statistics shows:

```
Correctly Classified Instances      2075      91.4903 %
Incorrectly Classified Instances    193       8.5097 %
```

=== Confusion Matrix ===

```

  a  b  c  d  e  f  g  h  i  j  k  <-- classified as
63  0  0  1  5  0  0  0  0  2  0 |  a = Justice
 0 100  0  1  3  0  1  0  0  1  0 |  b = Social
 0  0 128  2  5  2  0  0  0  2  0 |  c = Org
 0  1  2 382 17  4  4  0  2  3  0 |  d = Economy
 1  1  2  3 398  1  1  0  0  8  0 |  e = Foreign
 0  0  0  4  5 240  2  0  1  3  0 |  f = Politics
 1  1  0 16  9  6 265  2  1  5  0 |  g = Gov
 0  0  0  8  3  0  2 57  0  0  0 |  h = Infra
 0  0  0  2  4  0  0  0 53  0  0 |  i = Different
 0  0  1  6 21  0  9  0  1 331  0 |  j = DefSecurity
 0  0  0  1  4  0  0  0  0  0 58 |  k = Sport
```

**Table 5.7** The Confusion Matrix for the SMO Classifier on eleven categories.

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.887	0.001	0.969	0.887	0.926	0.959	Justice
0.943	0.001	0.971	0.943	0.957	0.975	Social
0.921	0.002	0.962	0.921	0.941	0.938	Org
0.92	0.024	0.897	0.92	0.908	0.967	Economy
0.959	0.041	0.84	0.959	0.895	0.964	Foreign
0.941	0.006	0.949	0.941	0.945	0.984	Politics
0.866	0.01	0.933	0.866	0.898	0.956	Gov
0.814	0.001	0.966	0.814	0.884	0.927	Infra
0.898	0.002	0.914	0.898	0.906	0.966	Different
0.897	0.013	0.932	0.897	0.914	0.968	DefSecurity
0.921	0	1	0.921	0.959	0.989	Sport

**Table 5.8** Detailed accuracy by class for SMO classifier on eleven categories.

As it can be seen from the detailed accuracy by class for SMO classifier on eleven categories (Table 5.8), the FP-rate, for all the categories are very low, instead the TP-rate, Precision, F-measure and ROC-Area is higher. This classifier can be employed in automatic classification of Afaan Oromoo text with good acceptable accuracy.

### 5.7.2 Classification using NaiveBayesMultiNominal

Naïve Bayes Multi Nominal Classifier is a Bayesian Classifier model that captures the word frequency information in documents. A document is represented by the set of word occurrences from the document in which the number of occurrences of each word in the document is captured. In order to make the comparison of SVM algorithm and Bayesian algorithm easier, the experimentation in this section is performed on the same data, the same number of categories and the experiment was performed using 10-fold stratified cross-validation as is in section 5.7.1 above.

#### 5.7.2.1 Experiment on four categories

The four classes ‘Sport’, ‘Justice & Law’, ‘Others(Different)’ and ‘Infrastructure’ classification experimentation using the NaiveBayesMultiNominal (NBM) algorithm in Weka application package gave the following summary statistics:

Correctly Classified Instances	254	96.5779 %
Incorrectly Classified Instances	9	3.4221 %

```

=== Confusion Matrix ===
  a  b  c  d  <-- classified as
70  0  1  0 | a = Justice
 3 66  0  1 | b = Infra
 3  0 55  1 | c = Different
 0  0  0 63 | d = Sport

```

*Table 5.9 The Confusion Matrix for the NBM Classifier on four categories.*

Looking at the confusion matrix of the NiveBayesMultiNominal Classifier on four categories, all the documents of the Sports class are predicted correctly by the classifier.

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.986	0.031	0.921	0.986	0.952	0.996	Justice
0.943	0	1	0.943	0.971	0.997	Infra
0.932	0.005	0.982	0.932	0.957	0.996	Different
1	0.01	0.969	1	0.984	1	Sport

*Table 5.10 Detailed accuracy by class for the NBM Classifier on four categories.*

Observing the detailed accuracy by class of the NaiveBayesMultiNominal Classifier on four categories of the Afaan Oromoo news items, Recall, Precision and F-measures are higher whereas the FP-Rate is low. It can be seen that the algorithm can be applied for the Afaan Oromoo text classification of four classes.

### **5.7.2.2 Experiment on seven categories**

The second experiment using the NBM classifier was performed on seven classes, the classes ‘Sport’, ‘Justice & Law’, ‘Others(Different)’ and ‘Infrastructure’ which was tested in section 5.7.2.1 and the classes ‘Economy’, ‘Defense & Security’ and ‘Foreign related’ are used in an evaluation of the NiveBayesMultiNominal algorithm. The testing on the data is done using 10-fold stratified cross validation. The summary statistics shows:

Correctly Classified Instances	1281	87.6197 %
Incorrectly Classified Instances	181	12.3803 %
Total Number of Instances	1462	

```

=== Confusion Matrix ===
  a  b  c  d  e  f  g  <-- classified as
65  0  0  0  3  1  2 |  a = Justice
 2 62  0  1  4  1  0 |  b = Infra
 1  1 55  2  0  0  0 |  c = Different
 0  0  0 63  0  0  0 |  d = Sport
13 15 11  4 366  1  5 |  e = Economy
17  3  6  9  6 362 12 |  f = Foreign
17  1  2  7  7 27 308 |  g = DefSecurity

```

**Table 5.11** *The Confusion Matrix for the NBM Classifier on seven categories.*

Looking at the confusion matrix of the seven categories ( Table 5.11 ) we can observe that a number of documents from the economy class are predicted as ‘Infrastructure’ class and ‘Justice & Law’ class, also many documents of the ‘Defense & Security’ class are predicted as ‘Justice & Law’ class and ‘Foreign related’.

```

=== Detailed Accuracy By Class ===
TP Rate   FP Rate   Precision   Recall   F-Measure   ROC Area   Class
 0.915     0.036     0.565     0.915     0.699     0.98     Justice
 0.886     0.014     0.756     0.886     0.816     0.989     Infra
 0.932     0.014     0.743     0.932     0.827     0.99     Different
 1         0.016     0.733     1         0.846     1         Sport
 0.882     0.019     0.948     0.882     0.914     0.99     Economy
 0.872     0.029     0.923     0.872     0.897     0.986     Foreign
 0.835     0.017     0.942     0.835     0.885     0.983     DefSecurity

```

**Table 5.12** *Detailed accuracy by class for the NBM Classifier on seven categories*

Looking at the detailed accuracy by class for the NaiveBayesMultiNominal Classifier on seven categories (Table 5.12), the Classes with relatively greater number of instances

has greater precision, while classes with relatively lower number of instances has greater recall.

### 5.7.2.3 Experiment on eleven categories

The eleven classes, the ‘Sport’, ‘Justice & Law’, ‘Others(Different)’, ‘Infrastructure’, ‘Economy’, ‘Defense & Security’, and ‘International relations’ which was tested in (section 5.7.2.2) and the classes ‘Social affairs’, ‘Associations’, ‘National Politics’ and ‘Governmental organizations’ are used in an evaluation of the NaiveBayesMultiNominal algorithm. The testing on the data is done using 10-fold stratified cross validation. A total of 2,268 documents are used in the experiment.

The summary statistics shows:

Correctly Classified Instances	1983	87.4339 %
Incorrectly Classified Instances	285	12.5661 %
Total Number of Instances	2268	

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	j	k	<-- classified as
64	0	0	2	2	2	0	0	1	0	0	a = Justice
0	101	2	0	1	0	1	0	0	1	0	b = Social
0	0	128	4	0	4	0	0	1	1	1	c = Org
13	7	8	362	1	7	3	5	6	3	0	d = Economy
24	6	1	4	364	3	2	1	3	7	0	e = Foreign
4	3	4	5	1	226	6	1	0	4	1	f = Politics
2	4	2	10	4	8	261	8	0	5	2	g = Gov
1	3	0	2	1	2	1	60	0	0	0	h = Infra
2	0	0	0	0	1	0	1	54	0	1	i = Different
13	4	3	6	24	6	7	1	1	300	4	j = DefSecurity
0	0	0	0	0	0	0	0	0	0	63	k = Sport

*Table 5.13 The Confusion Matrix for the NBM Classifier on eleven categories.*

Looking at the confusion matrix (Table 5.13) the instances predicted out of their class are mostly between the classes 'International relation' and 'Justice & Law', 'Defense & Security' and 'Justice & Law', 'Defense & Security' and 'International Relations'. This is because these classes have many attributes in common.

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.901	0.027	0.52	0.901	0.66	0.993	Justice
0.953	0.012	0.789	0.953	0.863	0.989	Social
0.921	0.009	0.865	0.921	0.892	0.978	Org
0.872	0.018	0.916	0.872	0.894	0.979	Economy
0.877	0.018	0.915	0.877	0.895	0.985	Foreign
0.886	0.016	0.873	0.886	0.879	0.991	Politics
0.853	0.01	0.929	0.853	0.889	0.971	Gov
0.857	0.008	0.779	0.857	0.816	0.978	Infra
0.915	0.005	0.818	0.915	0.864	0.99	Different
0.813	0.011	0.935	0.813	0.87	0.976	DefSecurity
1	0.004	0.875	1	0.933	1	Sport

*Table 5.14 Detailed accuracy by class for the NBM Classifier on eleven categories.*

Looking at the detailed accuracy by class for the NaiveBayesMultiNominal Classifier on eleven categories (Table 5.14), The area under the ROC-curve for all the classes are near 1.0, showing good accuracy for prediction of each classes by the NBM classifier.

## 5.8 Automatic text classifiers and Afaan Oromo texts

Text classification systems, i.e. systems which can make distinction between meaningful classes of texts, have been widely studied in information retrieval and natural language processing [Lewis, 1991].



Natural language processing aims at the development of tools allowing to retrieve, extract, and classify linguistic information, and, to a certain level, to understand language and provide translation or relevant feedback. Automatic text classifiers (the learning systems) aims at determining a description of a given concept from a set of concept examples (train set) provided by the teacher and from the background knowledge [kazakov, 1999].

As it has been discussed in section 5.5; pre-processing on the Afaan Oromo news items texts was done before the experimentation processes conducted in using the classifier algorithms. In order to see closely, the effect of the two selected algorithms, SMO implementation of Support Vector Machine (SVM) from the functions and NaiveBayesMultiNominal from Bayesian Classifiers on a category of 4, 7 and 11 classes, the confusion matrix and the detailed accuracy by class of the experimentation was reported in section 5.7 of this chapter.

In addition to this, the learning algorithms which are practical on many attribute data, that is, algorithms applicable to high dimensionality data, available in Weka application package was experimented on the Afaan Oromo news text, for example, from the Decision trees, J48 algorithm; from Lazy learners, IBK(KNN) were seen to be applicable to the Afaan Oromoo news items with good accuracy.

Accordingly, Table 5.15 summarizes the average accuracy of SMO, BayesMultiNominal, J48 and KNN classifiers (learners) on 4, 7 and 11 categories of the Afaan Oromo news items.

Classifier	Average Accuracy for 4-category data	Average Accuracy for 7-category data	Average Accuracy for 11-category data
SMO	95.82%	93.70%	91.49%
NaiveBayesMultiNominal	96.58%	87.62%	87.43%
KNN	91.25%	81.87%	85.27%
J48	86.69%	79.69%	82.05%

*Table 5.15 Average accuracy of SMO, NaiveBayesMultiNominal, KNN and J48 classifiers.*

NaivBayesMultiNominal seems to give better classification accuracy on a category with lesser number of classes (for example better in 4 category), whereas SMO in general, gives better classification accuracy on all categories without much difference as the number of categories increases when compared to others.

The performance of all the classifiers on 4 category is greater than their respective performance on 11 category. The performance of the classifiers was with respect to accuracy. It does not consider classifier's performance with respect to time and memory requirement.

From the experiments carried out in this research, it can be seen that when categories having relatively equal number of documents tested together, it has resulted in better classification accuracy (experiment on four categories of section 5.7.1.1 and 5.7.2.1) than categories with different number of instances (documents) tested together (experiment on seven categories of section 5.7.1.2 and 5.7.2.2). Moreover, as the number of instances in a given class increases the classifier accuracy increases.

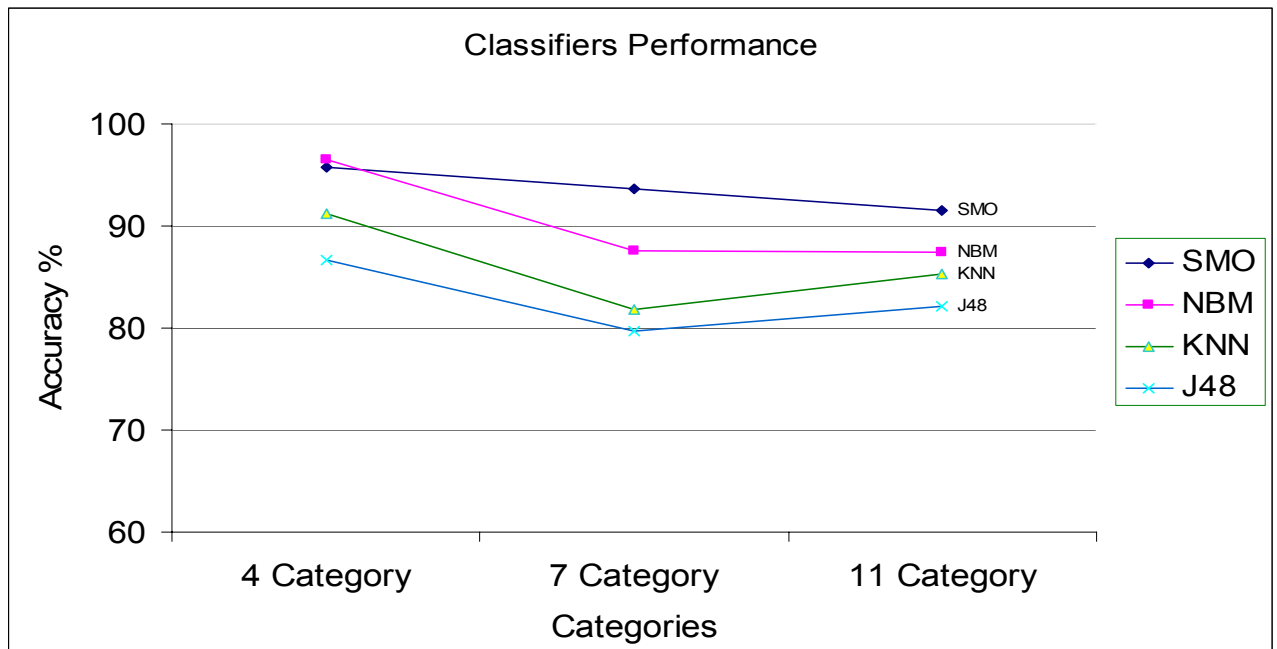


Figure 5.2 Average accuracy of SMO, NaiveBayesMultiNominal, KNN and J48 classifiers.

The results of the experimentation have shown that SMO algorithm tends to be the best classifier for Afaan Oromo news items as the number of classes in a category increases. The accuracy obtained from others classifiers are also encouraging as it can be observed from figure 5.2 shown above.

In general, this research asserts that, Machine learning approach can be applied to the automatic classification of Afaan Oromo news items.

## CHAPTER SIX

### CONCLUSIONS AND RECOMMENDATIONS

#### 6.1 CONCLUSIONS

The rapid expansion in Information and Communication Technology has resulted in the creation of large volume of text in electronic form. As the volume of information continues to increase, there is growing interest in helping people better find, filter, and manage these resources. Text categorization - the assignment of natural language documents to one or more predefined categories based on their content is an important component in many information organization and management tasks.

The automated categorization (or classification) of texts into pre-specified categories, although dating back to the early 1960s, has witnessed a booming interest in the last ten years, due to the increased availability of documents in digital form and the ensuing need to organize them [Sebastiani,2002].

Currently, Machine Learning (ML) is one of the major popular approaches to text categorization. According to ML approaches to text categorization, a general inductive process automatically builds an automatic text classifier by learning from a set of pre-classified documents, the characteristics of the categories of interest.

The objectives of this research has been to prepare processing tools for Afaan Oromo text classifications, and test the applicability of automatic text classifiers for Afaan Oromo text classification activities based on document content.

Many factors affect the success of Machine learning (ML) on a given task. The representation and quality of the instance data is first and foremost. If there is much irrelevant and redundant information present or noisy and unreliable data, then

knowledge discovery during the training phase is more difficult. It is well known that data preparation and filtering steps take considerable amount of processing time in ML problems. Data pre-processing includes data cleaning, normalization, transformation, feature extraction and selection, etc. The product of data pre-processing is the final training set [Kotsiantis S., et al., 2006].

To this end, much attention was given on the pre-processing of the source data by developing language dependent pre-processing tools, for the Afaan Oromo news items, obtained from the Radio Fana for the classification purpose. As mentioned in section 5.2 news items database of Radio Fana incorporate both Amharic and Afaan Oromo news items together, therefore Afaan Oromo news items are prepared in different database. Only the headline, the slug and the keywords were considered to build models assuming that they contain features which represent the document and the processing and learning time of the algorithms can also be reduced. Therefore, procedures to facilitate these things were prepared. Furthermore the following tools were developed:

- A tool for tokenization , Afaan Oromo words identification
- Removal of extraneous characters, numeric characters
- Stop words removal tool from the news items texts
- Attempt to remove affixes from Afaan Oromo words

The use of these tools helped the reduction of features and data cleaning for the automatic text classification process.

After making the Afaan Oromo news items text comfortable to the tool used in this research for classification (Weka application package) feature selection, term weighting and data transformation to Arff (Attribute relation file format) was employed before

training and classification tasks. During the experimentation, 10-fold cross-validation was used to estimate the classifier accuracy.

Finally, from this work on automatic classification of Afaan Oromo news items, it can be concluded that:

- Proper data pre-processing techniques which considers the Natural Language Processing increases the effectiveness of the automatic classification.
- The best result (accuracy) obtained from both the SMO and Bayes MultiNominal classifiers was when the number of instances are approximately equal in each class and the accuracy is 95.82% and 96.58% respectively for the category of 4 classes.
- Relatively lower accuracy obtained is for J48 on category of 7 classes 79.69% and on category of 11 classes 82.05%
- Both SMO and BayesMultiNominal showed high accuracy, SMO tends to have better accuracy over BayesMultiNominal for the Afaan Oromo news items classification.
- Both SMO and BayesMultiNominal showed better accuracy over others Lazy(IBM), Dtree(J48) classifiers for Afaan Oromo news items.
- Generally in all classifiers uneven distribution of instances in classes tend to decrease the accuracy of the classifiers when taken together as shown in the experimentation of eleven categories.
- This study demonstrates that Machine Learning approach can be applied to Afaan Oromo news items classification tasks.

## 6.2 RECOMMENDATIONS

The results of this research indicate machine learning techniques are applicable for automatic classification of Afaan Oromo news items. Success of machine learning is foremost affected by the representation and quality of instance data. Therefore much more has to be done to ensure automatic processing of Afaan Oromo texts under all circumstances.

- The stop-word list used in this research is compiled during the data preparation and mostly is news specific. The availability of standard stop-word list would definitely facilitate researches in the area of automatic classification; therefore a standard Afaan Oromo stop-word list should be developed.
- The availability of standardized text corpus promotes text classification researches, nevertheless, there is no established text corpus for Afaan Oromo text classification purpose and checking, hence there is a need for standardized text corpus.
- In this research, the researcher tried to correct some of the spelling errors manually which is not exhaustive for the purpose of this research. Spelling errors has effect in attribute reduction and selection for automatic classification, indicating a need for Afaan Oromo spell checker.
- Afaan Oromo is a morphologically rich language; as a result there are many variants of words in Afaan Oromo, a construction of a thesaurus for the language helps to bring the variants together and standardization.

- Some of the machine learning algorithms (SMO, NaiveBayesMultiNominal, IBK, J48) has been tested for classification of Afaan Oromo news items, testing of machine learning algorithms on other Afaan Oromo texts and other machine learning algorithms could help.
- This research considers only the classification of main classes, it does not consider the classification of sub-classes, and therefore, research in this direction would also improve the classification quality.
- This research considers the single-label classification; it assigns an instance (document) to one of the pre-defined classes. The case where assigning an instance to more than one class (multi-label classification) is an issue that has to be studied.



## BIBLIOGRAPHY

- Adriani M. and Croft, W.B.. (1997), Retrieval Effectiveness of Various Indexing Techniques on Indonesian News Articles, CIIR Technical Report, IR-169, University of Massachusetts, Amhrest.
- Assefa Woldemariam, (2005), Development of Morphological Analyzer for Afaan Oromo text, Master thesis at Faculty of Informatics, Addis Ababa University, Addis Ababa.
- Bacan H., Pandzic, I. S., and Gulija, D., (2005), Automated News Item Categorization, In Proceedings of the 19<sup>th</sup> Annual Conference of the Japanese Society for Artificial Intelligence, pp 251-256, Kitakyushu, Japan.
- Beza-Yates, Ricardo and Berthier Ribeiro-Neto, (1999), Modern Information Retrieval, Harlow, Addison Wesley, England
- Diriba Megersa, (2002), An Automatic Sentence parser for Oromo language, Master thesis at school of Information studies for Africa, Addis Ababa University, Addis Ababa.
- Dumais S. Platt J, Heckerman D and Shami M., (1998), Inductive Learning Algorithms and Representations for Text Categorization. In Proceedings of the Seventh International Conference on Information And Knowledge Management , pp. 148-155.
- El-Kourid M., Bensaid A., and Rachidi T., (2004), Automatic Arabic document categorization based on the Naïve-Bayes Algorithm, Workshop on computational Approaches to Arabic Script-based languages, COLING-2004, University of Geneva, Switzerland.

- Eyheramendy S., Genkin A., Wen-Hua Ju, David D. , Lewis and Madigan D., (2003), Sparse Bayesian Classifiers for text Categorization, Journal of Intelligence Community Research and Development, Available at <http://www.stat.rutgers.edu/~madigan/PAPERS/jicrd-v13.pdf>, visited on December 10, 2008.
- Gumii Qormaata Afaan Oromoo, (1996 E.C), Caasluga Afaan Oromoo, Komishinii Aadaaf Turizmii Oromiyaa, Jildii – 1. Finfinnee
- Han and Kamber, (2004), Data mining: concepts and techniques, First edition, Morgan Kaufmann Publishers, San Francisco, California.
- Han and Kamber, (2006), Data mining: concepts and techniques, Second edition, Morgan Kaufmann Publishers, San Francisco, California.
- Joachims T., (2002) , Learning to classify text using Support Vector Machine, Methods, Theory and Algorithms, Kluwer Academic publishers, Ah Dordrecht, The Netherlands.
- Joachims T., (1998), Text categorization with support vector machines: Learning with many relevant features, In Proceedings of ECML-98, 10<sup>th</sup> European Conference on Machine Learning, (Chemnitz, Germany, 1998), 137-142.
- Kazakov, (1999), Natural language processing application of machine learning, doctoral thesis, Czech Technical University, Prague
- Kotsiantis S., Kanellopoulos D. and Pintelas P., (2006) , Data Preprocessing for Supervised Learning, International Journal of Computer Science, Vol. 1 No. 2, pp 111-117

- Kotsiantis S., Ikonomakis M. and Tampakas V., (2005), Text Classification using Machine Learning Techniques, Wseas Transactions on Computers, Issue 8, Volume 4, pp. 966-974
- Kotsiantis S., S.B., (2007), Supervised Machine Learning: A review of Classification Techniques, Informatica, Vol. 31, No. 3, pp 249-268, Slovene Society Informatika.
- Kula K. and Varma V., (2006), Oromo-English Information Retrieval Experiments at CLEF 2006, Working Notes. Hyderabad, India.
- Kula K., Varma V. and Pingali P., (2007), Evaluation of Oromo-English Cross-Language Information Retrieval. In IJCAI 2007 Workshop on CLIA, Hyderabad, India.
- Lam W., Ruiz ME and Srinivasan P., (1999), Automatic text categorization and its application to text retrieval, IEEE Trans. Knowl. Data Engin. 11, 6, p865-879.
- Lewis, (1991), Evaluating text categorization, In Proceedings of the speech and Natural Language Workshop, Morgan Kaufmann, pp 312-318.
- Mark W. Pope, (2007), Automatic classification of online news headlines, a master's paper for the M.Sc in IS degree, Chapel Hill, North Carolina
- McCallum A and Nigam K., (1998), A Comparison of Event Models for Naive Bayes Text Classification, In Working notes of the ICML/AAAI workshop on Learning for text Categorization, Pittsburgh,
- Mitchell S., (2007), The History of News - 3rd Ed, Oxford University Press, New York.
- Morka mekonnen. (2001), Text-to-Speech System for Afaan Oromo, Master thesis at school of Information studies for Africa, Addis Ababa University, Addis Ababa.

- Nils J. Nilsson, 1996, Introduction to machine learning: An Early Draft of a proposed textbook, available at <http://robotics.stanford.edu/people/nilsson/mlbook.html>
- Olivier, D.V. , 2000, Mining e-mail Authorship, Proceedings of Sixth ACM SIGKDD, International Conference on Knowledge Discovery and Data Mining, Boston, USA.
- Peng F. and Schuurmans D., (2003), Combining naïve bayes and n-gram language models for text classification, European Conference on IR Research, ECIR'03. Pisa, Italy.
- Platt, J. C., (1998), Fast training of support vector machines using sequential minimal Optimization (SMO), Support Vector Learning, MIT Press, Cambridge.
- Radio Fana: Available at <http://www.radiofana.com/> Visited in December, 2008
- Rafael A. and Calvo, (2003), Coping with the news: the machine learning way. In A. Treloar and A. Ellis, editors, Proceedings of Ausweb 2003 Conference, Gold Coast.
- Rafael A. Calvo, Jae-Moon L, and Xiaobo L , (2004), Managing content with automatic document classification, Journal of digital information, volume 5 issue 2
- Richard Hudson, (2003), An encyclopedia of English Grammar and Word Grammar, Available at <http://www.phon.ucl.ac.uk/home/dick/enc-gen.htm> visited in November, 2008.
- Roberts A., (2003), Machine Learning in Natural Language Processing, Available at

[http://www.comp.leeds.ac.uk/andyr/misc/latex/sessions/bibtex/bib\\_example.pdf](http://www.comp.leeds.ac.uk/andyr/misc/latex/sessions/bibtex/bib_example.pdf), visited in December, 2008.

Robinson H., (2003), Feature Selection and representation in text categorization,

Available at:

<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.1.7936> seen on November 22, 2008.

Salton G, 1983, Introduction to Modern Information Retrieval, McGraw-Hill, Inc. New York

Sebastiani F, (1999), Machine Learning in Automated text categorization, a survey: Technical report IEI-B4-31, consiglio nazionale delle ricerche, Italy

Sebastiani F., (2002), Machine Learning in Automated text categorization, ACM computing Surveys, vol. 34, No. 1, pp 1-47.

Sebastiani F., (2005), Text Categorization, In Alessandro Zanasi, editor, Text mining and its application to Intelligence, CRM and Knowledge Management, WIT press, pp 109-129, Southampton, UK.

Tokunaga,(1994), Text categorization based on weighted inverse document frequency, Special Interest Groups and Information Process Society of Japan (SIG-IPJS), pp 33-39, Tokyo, Japan  
visited in November, 2008.

Tilahun Gemta, (1994), Afaan Oromoo, Journal of Oromoo Studies

Available at

[http://www.africa.upenn.edu/Hornet/Afaan\\_Oromo\\_19777.html](http://www.africa.upenn.edu/Hornet/Afaan_Oromo_19777.html)  
seen on November 22, 2008.

- Wakshum mekonnen. (2000), Development of a stemming algorithm for Afaan Oromo text, Master thesis at school of Information studies for Africa, Addis Ababa University, Addis Ababa.
- Yancong Z. and Hyuk cho., (2001), Classification Algorithms on text document.  
Available at  
<http://www.cs.utexas.edu/users/hyukcho/classificationAlgorithm.html>,  
Visited in December, 2008.
- Yang Y. and Liu X., (1999), A re-examination of text categorization methods. In proceedings of SIGR-99, 22<sup>nd</sup> ACM International conference on Research and Development in Information Retrieval, pp 42-49, Berkeley, CA.
- Yang Y., (1999), An evaluation of statistical approaches to text categorization. Information Retrieval, 1, 1-2, pp 69-90.
- Yang Y. and Thorsten J. , (2008), Text categorization, Scholarpedia journal, vol.3 No.5,4242
- Zelalem Sintayehu, (2001), Automatic classification of Amharic News items: The case of Ethiopian news agency, Master thesis at school of Information studies for Africa, Addis Ababa University, Addis Ababa
- Zhang Y, Gong L. and Wang Y., (2005), An improved TF-IDF approach for text classifications, J Zhejiang Univ SCI 2005,6A(1), pp49-55, China.

**APPENDICES**

**Appendix 1 Samples of Records removed during manual scanning**

<b>ID</b>	<b>Head Line/News Story</b>
372	ዜና ማንበብ ዜና ?? ?? ????? ? ? ?????? ????? ???
373	ZENA ዜና ????? ?????? ? ??
380	YE MISRAK OROMIA ????? ????? ????? ??? ??? 20
391	የዛሬ ዜና አክ ??? ??
411	ygara simiminti
413	ye dinber simmeminet
447	መልካም አስተዳደር
452	በኦሮሚያ የፍትህ ስልጠና ማሻሻያው አየተተገበረ ነው
458	በኦሮሚያ ያክልል 920ሺህ ቤቶች ሊገነቡ ነው
882	kiliiniika dhunfaa naannoo oromiyaa
887	kiliiniika dhunfaa naannoo oromiyaa
888	WATATOCH AD ????? ????? ????? ????? ?????
1055	y zare zena
1147	የኦሮሚያ መምህራን ማህበር የተለያዩ ስራዎችን ሰራ
1891	AKKAM JIRTA BRE OGEESSI TEEKNIKAA-----=----AN
2044	?
2143	linsesat hkmina tst
2169	?Godina Harargee bahaatti manni maree aanaa Jaarsoo bajata
2572	በኦሮሚያ የጤና ኤክስፔንሽን አገልግሎት ?????? ??? ?????? ??? ? 1
2863	L72 KTMOCI PILAN LISRA NWU DB ?????? ?????
3183	Sagantaan nyaataa addunyaa gargaarsa nyaataa
3184	Sagantaan nyaataa addunyaa gargaarsa nyaataa
3638	YMIRT BIKNT DB ????? ??? ????? ????????? ??? ??? ????? ?????? ??????
3878	LASHAS S,XAS ????? ??? ????????? ??? ????? ?100 ????? ??? ???
5471	ኸመነጅ ??????? ????? ?????? ?? ??? ???????
9088	በኦሮሚያ ክልል ?????? ?????? ????? ??? ?? ????? ????????? ?????? ? 8
12316	?Ayyaana injifannoo caamsaa 20 waggaa 16ffaa sirni dargii i
12446	?Caamsaan bultiin 20n mootummaan abbaa irree kan itti barbadaa
12557	ymikr beet zeena
17569	DHALATTOONNII FI LAMMIWWAN A DHALATTOONNII FI ITOOPHIYAA
33202	መመመመመመ መመመመ ?????? ????? ????????
34430	CCC YTLYAYU W-ETOC

## Appendix 2 Samples of the news head line from the database

ID	Head Line	Class
261	komishiniin Investimani Oromiyaa abbootii qabeenyaa 732 barana hayyama kenne.	Economy
477	piiroojektoota biishaan	Infrastructure
677	xuumuraa marii qaamolee seera	Justice&Law
882	kiliiniika dhunfaa naannoo oromiyaa	Social
1151	Israa'eelii fi Liibaanoos wal waraansa eegalan cimsanii itti fufaniiru.	Foreign
1153	waldaaleen hojii gamtaa Oromiyaa	Associations
1589	balaa lolaa dirree dawaa	Governmental
1951	Otoobusni deeddeebisa ummataa Egypt tokko garagaluusaatiin namoonni 9 akka du'an dhaga'ame	Foreign
1990	abban taayitaa daandilee baadiyyaa pirojaktiiwwan 28 raawwachuu isaa beeksise.	Infrastructure
2288	koreen sadaarkaa biyyaalessaatti galii walitti sassaabu hundeeffame.	Economy
2547	USAID GARGAARSA MILIYOONA 20 KENNUUF	Foreign
2676	Ministirri muummichi Mallas Zeenaawii duree ADWUI ta'anii lammata filatamanii	Def&Security
3712	Busaa ittisuuf sochiin cimaan taasifamaa jira	Social
4282	sochii dhaabbilee mit-mootumma damee fayyaa	Social
4457	mana marii islaamummaa	Associations
4609	Yuniivarstiiwwan haroomaayaa ,Goondarii fi Baahirdaar	Infrastructure
4647	Waldaan hojii gamtaa qonnaan bultootaa miliyoona afurii ol fayyadamoo taasise.	Associations
5780	zoonii industirii 12 keessatti hojiin bu'ura misoomaa guutuu hoojjetamaa jira	Economy
6293	Itoophiyaan tarkaanfiinnlittisaa birmadummaashee kabajisiisuuf fudhachaa jirtu seera	Governmental
12316	Ayyaana injifannoo caamsaa 20 waggaa 16ffaa	Politics
13778	Ajandaa jeequmsaa mootummaan Ertiraa	Politics
15592	Wal morkii kuphaa miilaa dubartootaa	Sport
17106	Bar kumee haaraatti dargaggoo cimani hojjachuu qabu jedhame.	Governmental
17947	Ayyaanni Irreechaa sirna oowwaan ayyaaneeffamaati jira.	Social
20767	Ministirii dhimma alaa Ameriikaa raayiis finfinne seenan.	Def&Security
24992	Mana maree federeshiinaa	Politics
25252	Instituutii qorannoo qonna Itoophiyaa waggaa 40ffaa isaa ayyaaneeffate.	Different
25675	Poolisii ni Oromiyaa shoora irraa eeggamu gumaachu qaba jedhame.	Justice&Law
26041	Ibiddi balaa godinaalee sadi keessatti ka'ee ture too'atame.	Def&Security
27519	shaampiyoonni atileetiksii afrikaa sportii 16ffa	Sport
32553	Hoojjettoonni mootummaa duula harama haramuu gaggeessan	Different
33649	Adabamtoota seera naannoo Oromiyaa 4830f dhiifamni taasifame.	Justice&Law
34013	Haala qilleensaa godduu kanaa maal fakkaata.	Different



### Appendix 3 List of Afaan Oromoo Stop Words used in the research

abbaan	bakkee	dhufu	gara
adda	balleese	dhufuu	garaagar
addunyaa	bar	dubbataniiru	gargaarsa
adeemsifame	bara	dura	garuu
akka	bar-kume	duraan	gauf
akkan	barkumee	dursee	giddu
akkas	beeksifte	duula	gidduu
akkasuma	beeksisaan	ebila	gidduutti
akkuma	beeksisan	ebla	guddaa
al	beeksise	eblaa	guddina
alaa	biiliyoona	eega	guyyaa
amma	bira	eegalame	haa
ammas	biratti	eegalee	haala
ammoo	biyya	eegamee	haalli
an	biyyaa	egalame	haaraa
an	biyyaaleessaa	ejansii	hal
ana	booda	erga	hanga
anaa	boodatti	ergamaa	hara
anaaf	bor	ergamerraa	hawaasa
anaafi	boru	farra	heddu
ani	boruu	federaalaa	hedduu
argaman	bre	fi	hedduun
armaan	buuraa	fuafa	hi
arra	carraa	fuula	hime
ati	dandau	fuuldura	hin
baatii	dha	gabaafame	hirmaana
babalisuuf	dha	gad	hojii
bahaa	dhaabuu	gadi	hubachiise
bakka	dhufan	gahe	ibsame
bakkan	dhufeenya	gama	ibsan

### Appendix 3 List of Afaan Oromoo Stop Words used in the research (cont...)

idil	jalqabde	kkf	mormii
immoo	jalqqabame	koo	muraasa
inni	jedhame	kootiin	murtesse
iraa	jedhan	koottu	na
irraa	jjoota	kudhan	naaf
irraaf	jiran	kuma	naan
irran	kaabaa	kumee	nan
irratii	kaabaarratti	kun	ni
isa	kabajame	kunneen	oboo
isaa	kaleessa	kuntalaa	odoo
isaaf	kam	kurnan	oduu
isaan	kan	kutaa	of
isaani	kana	lafa	ol
isaanii	kanaaf	lafaa	olaanaa
isaanis	kanaafuu	lafaa	olaanaan
isaati	kanaan	laga	oliin
isaatiin	kanan	lakkoofsa	olitti
isaatu	kanarrat	lakkoofsi	oota
ishee	kee	lama	osoo
isheen	keenya	lamaa	osoon
isheetii	keenyyaa	lamaan	qaba
itiyoophiyaa	keessa	lixaatti	qaban
itti	keessaa	maal	qabdu
ittiin	keessatii	malee	qabne
ittinuu	keessatti	mana	qofa
ittisuuf	kenne	manneen	qophiin
ittuu	kennu	manni	rratti
jala	kibbaatti	marti	sababa
jalqaba	kiyya	miti	sadaasa
jalqabama	kiyyaan	mitii	sadaffaa

### Appendix 3 List of Afaan Oromoo Stop Words used in the research (cont...)

sadan	teenya	yemmuu
sadarkaa	teenyatti	yeroo
sadarkaa	tii	ykn
saddeet	tiin	yoo
sadeen	tokkicha	yoon
sadii	tokko	
sagantaa	tokkofa	
sagantaan	tokkoo	
sagataa	tokkotti	
san	tootame	
sana	tumsi	
sanatti	tumsuu	
sanii	turan	
shan	ture	
shanan	waa	
shawaa	waan	
si	waaraa	
soddoma	waaraafi	
sun	wagaa	
taasifamaa	waggaa	
taasifamera	waggota	
taasifamuu	wajjin	
taasisee	wal	
taayitaa	waliin	
tae	walitti	
tajaajila	wanta	
tamsasuu	wantoota	
tasiisama	warra	
tattaafiin	xumurame	
tau	yaa	

## **DECLARATION**

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

---

Abera Diriba Gemechu

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

---

Dejene Ejigu (Ph.D.)

March, 2009