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

LIST OF TABLES

LIST OF APPENDICES

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.

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.

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[1] 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.

[1] 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

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

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 REIVEW 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]

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: X** $\rightarrow$ 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

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

$Doc_i = (term_{i1}, term_{i2}, term_{i3}, \ldots , term_{it})$

$Doc_j = (term_{j1}, term_{j2}, term_{j3}, \ldots , term_{jt})$

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 **C** and for the set of documents **D** where :

$\mathbf{C} = \{c_1, c_2, c_3, c_4, c_5, c_6, \ldots C_m\}$

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

We can consider the following, for a given integer **k**

Exactly **k** elements of **C** are assigned to each $\mathbf{d_j} \in \mathbf{D}$. The case in which exactly one category (**k** = 1) must be assigned to each $\mathbf{d_j} \in D$ 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 $d_j \in D$ 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 $d_j \in D$ belongs to a category $c_i \in C$. 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].

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.

aerospace discipline, or the MeSH[4] 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.

[4] 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:

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

- 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[2]) 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

[2] WEKA : **W**aikato **E**nvironment for **K**nowledge **A**nalysis. Developed at the University of Waikato in New Zeland.

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 $\mathbf{d}_j$ 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 $\mathbf{d}_j$ is typically represented as a vector of term weights

$$\mathbf{d}_j \text{ vector} = (w_{1j}, w_{2j}, w_{3j}, w_{4j}, \ldots, w_{|t|j}) \qquad (3.1)$$

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.

> 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 **ω** in a document **d** is given by:

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

Where:

- ω_{ik} is the weight of term i in the k^{th} document
- f_{ik} is the frequency of the i^{th} term in the k^{th} document
- N is the number of documents in the collection
- n_i is the number of documents in which the i^{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, ... , 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: x** $\rightarrow$ **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 tile 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[5] 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.

[5] 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.

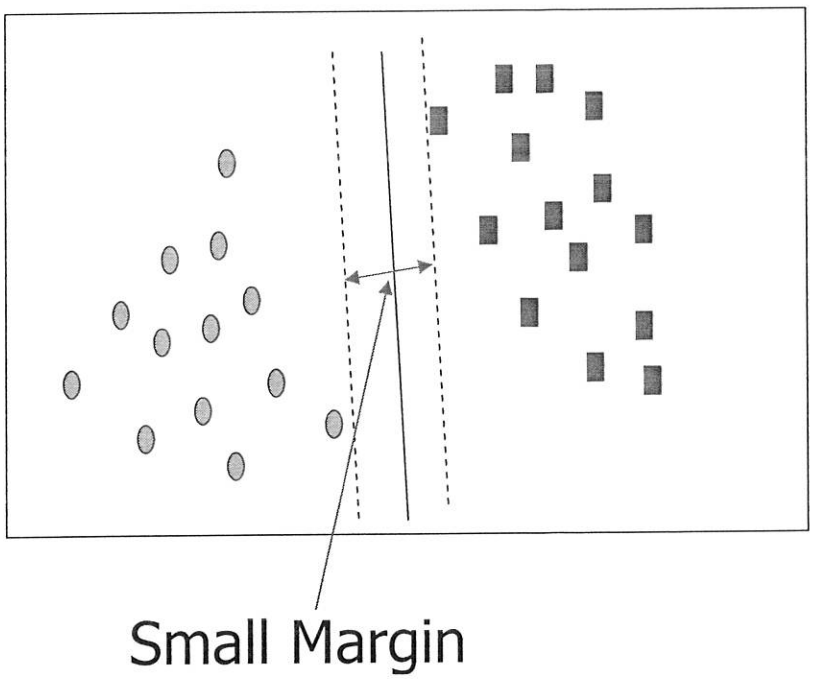

Figure 3.1 SVM Small Margin

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.

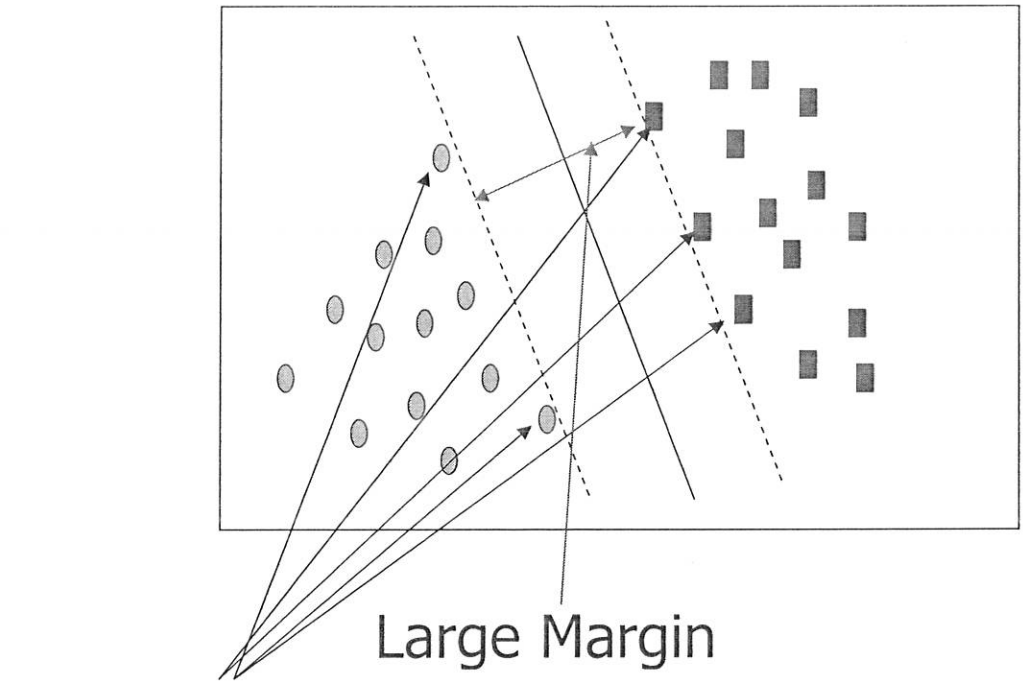

Figure 3.2 SVM Large Margin

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.

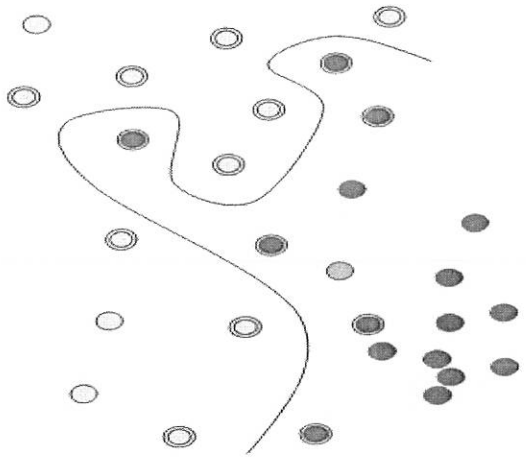

Figure 3.3 Non-Linear region

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.

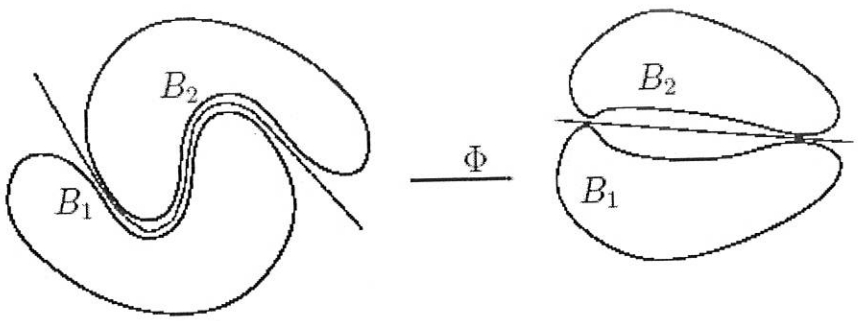

Figure 3.4 Kernel function transform data

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

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, $\mathbf{X_1} = (x_{11}, x_{12}, x_{13}, \ldots, x_{14}, x_{1n})$ and $\mathbf{X_2} = (x_{21}, x_{22}, x_{23}, \ldots, x_{24}, x_{2n})$, is

$$\text{Dist (X1, X2)} = \sqrt{\sum_{i=1}^{i=n} (x_{1i} - x_{2i})^2} \qquad (3.3)$$

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.

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.

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.

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

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

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

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

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

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].

Accuracy = (t-pos + t-neg) / (t-pos + t-neg + f-pos + f-neg) (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.

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

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

Specificity = t-neg / neg (3.9)

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

Precision = t-pos / (t-pos + f-pos) (3.10)

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

Accuracy = Sensitivity * (pos/(pos + neg) + Specificity * (neg / (pos + neg) (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-poistive 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 asses 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 "Dubbifttuu" 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: **uu**maa to mean 'nature', **ee**lee to mean 'pan',

Medial position: k**ee**nnaa to mean 'gift', l**ee**xaa to mean 'single'

Final position: gar**aa** to mean 'belly', daar**aa** 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	**i , ii**		**u, uu**
Mid	**e, ee**		**o, oo**
Low		**a, aa**	

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, ltuu

[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 Affricates	Voiceless	(p)	t	ch	k	?
	Voiced	b	d	j	g	
	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.

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 → actoress
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 language**s 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.

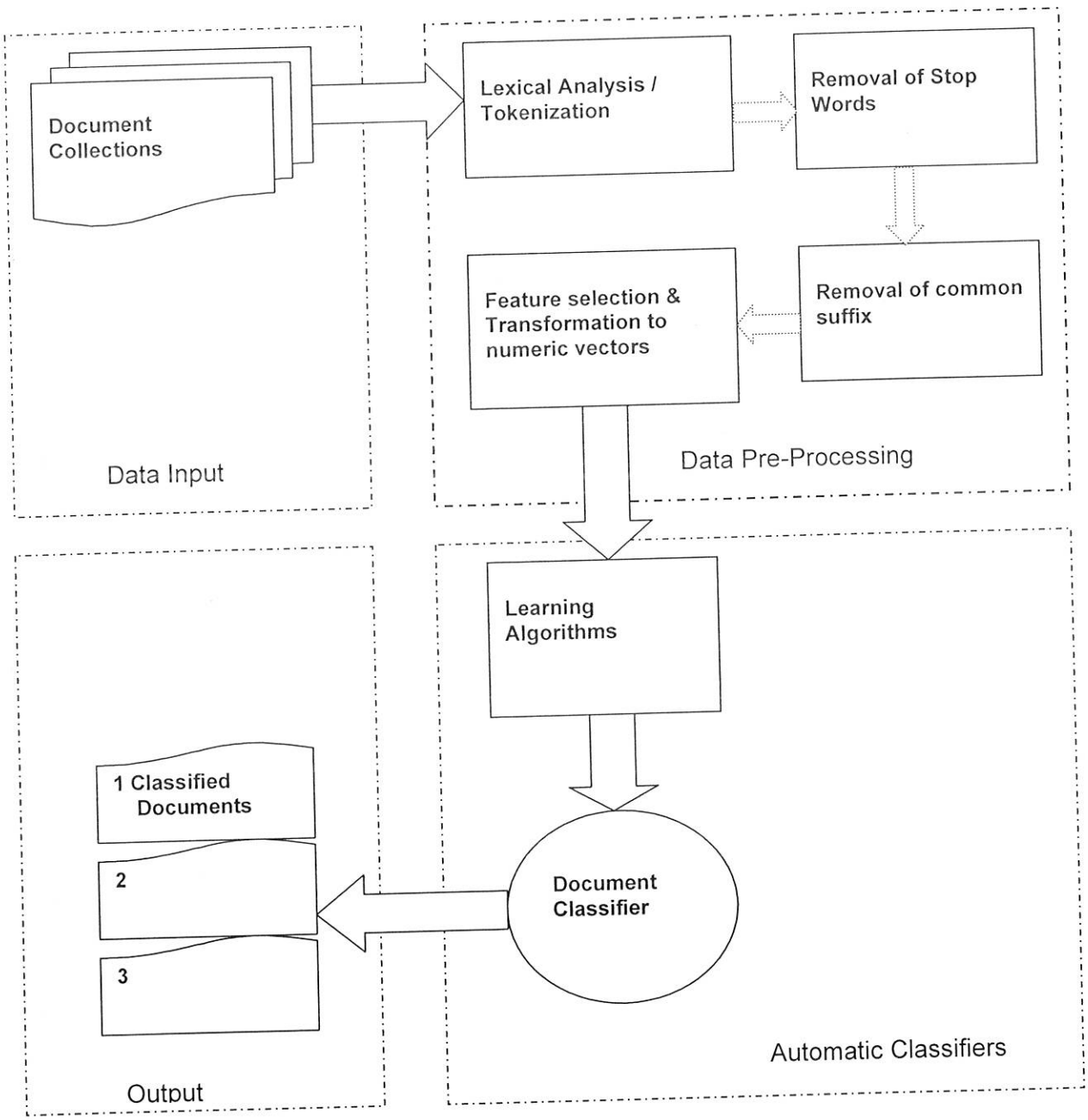

Figure 5.1 Machine Learning text classification Architecture

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.

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.

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 "lammii" 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 coma 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.

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.

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.

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].

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.

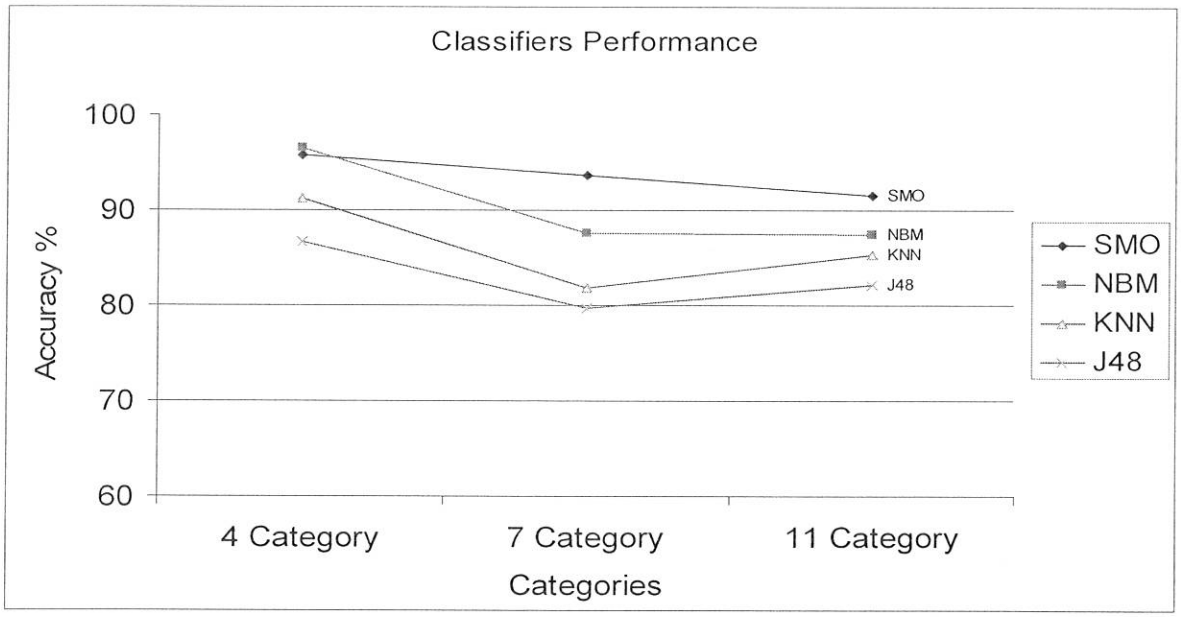

Figure 5.2 Average accuracy of SMO, NaiveBayesMultiNominal, KNN and J48 classifiers.

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 form 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(IBK), 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.

Nils J.Nilsson, 1996, Introduction to machine learning: An Early Draft of a proposed textbook, available at http:/robotics.standord.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 Cost.

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

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, 22nd 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

ID	Head Line/News Story
372	ዜና ማንበብ ዜና ?? ?? ????? ?? ?????? ???? ???
373	ZENA ዜና ???? ?????? ? ??
380	YE MISRAK OROMIA ????? ???? ????? ??? ??? 20
391	የዛሬ ዜና አለ ??? ??
411	ygara simiminti
413	ye dinber simmeminet
447	መልካም አስተዳደር
452	በኦሮሚያ የፍትህ ስልአት ማሻሻያዉ እየተተገበረ ነዉ
458	በኦሮሚ ያክልል 920ሺህ ቤቶች ሊገነቡ ነዉ
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 ኤክሰቴንሽነ 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	ኘመነጀ ?????? ???? ????? ?? ??? ??????
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 LAMMIIWWAN A DHALATTOONNII FI ITOOPHIYAA
33202	መመመመመመ መመመ ????? ???? ??????
34430	CCC YTLYAYU W-ETOC

Appendix 2 Samples of the news head line from the database

ID	Head Line	Class
261	komishiniin Invastimanii 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 indastrii 12 keessatti hojiin bu'ura misoomaa guutuu hoojjetamaa jira	Economy
6293	Itoophiyaan tarkaanfiinnIttisaa 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 cimanii 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	jioota	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	keenyaa	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	tooatame	
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