ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
FACULTY OF INFORMATICS
DEPARTMENT OF INFORMATION SCIENCE

AUTOMATIC AMHARIC TEXT NEWS CLASSIFICATION: A NEURAL NETWORKS APPROACH

BY

WORKU KELEMEWORK

A THESIS SUBMITTED TO
THE SCHOOL OF GRADUATE STUDIES OF ADDIS ABABA UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE IN INFORMATION SCIENCE

ADDIS ABABA, ETHIOPIA
OCTOBER 2009

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
FACULTY OF INFORMATICS
DEPARTMENT OF INFORMATION SCIENCE

AUTOMATIC AMHARIC TEXT NEWS CLASSIFICATION: A NEURAL NETWORKS APPROACH

BY

WORKU KELEMEWORK

APPROVED BY

EXAMINING BOARD:

MILLION MESHESHA (PhD), ADVISOR _____

LEMLEM HAGOS (M.I.Sc.), CO-ADVISOR _____

WORKU ALEMU (PhD), EXAMINER _____

Henock Lulseged, Chairperson _____

_____

DEDICATION

This work is devoted to my mother, Maydersa Ayele Teka (Shashe) and the Motherland Ethiopia.

ACKNOWLEDGEMENT

The first thank is for God. The next deep gratitude into my heart, words have no power to express, is to Saint Michael.

I am extremely grateful for my advisor Dr. Million Meshesha for his positive encouragement before the work was started and for his constructive comments and guidance after the work has been started. So, I gladly acknowledge a special debt of thanks for Dr. Million. The valuable comments and the welcoming behavior of my co-advisor, Lemlem Hagos, have been the major elements for the accomplishment of this study. I express thanks for my father Kelemework Birhanie Ayele (Gashe) for giving aids on Amharic language.

I thank all my friends who wish my best and helping me by providing suggestions, directions, encouragements, etc. Specially, Teshome Alemu for providing various kinds of helps, Assefa Misganaw who initiate me to do research on this topic, and Abinew Ali and Alemayehu Golla Guallu who have edited the draft report and provide important comments. I also thank Ebrahim Chekol for editing some portion of the draft report.

Worku Kelemework Birhanie

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF APPENDICES

Page

LIST OF ACRONYMS

ANN	Artificial Neural Network
ASCII	American Standard Code for Information Interchange
BPN	Back Propagation Network
DF	Document Frequency
ENA	Ethiopian News Agency
ICT	Information and Communication Technology
IDF	Inverse Document Frequency
IG	Information Gain
GUI	Graphical User Interface
KNN	K-Nearest Neighbor
LVQ	Learning Vector Quantization
MSE	Mean Square Error
SERA	System for Ethiopic Representation in ASCII
SOM	Self Organizing Map
SVM	Support Vector Machine
TF	Term Frequency
TF*IDF	Term Frequency by Inverse Document Frequency
VLSI	Very Large Scale Integrated

ABSTRACT

Text classification is one of the methods used to organize massively available textual information in a meaningful context to maximize utilization of information. Automatic text classification is the preferred method for accomplishing classification in large volumes of information. Research works on automatic classification is flourishing in the context of other languages; whereas, research on automatic Amharic text classification is in its infancy stage and very few attempts have been made till now. This study puts forward its own contribution for automatic Amharic text classification.

Before the classifier is constructed, preprocessing has been done on the data to make it ready for the learning algorithm including changing various Amharic characters with the same sound to one common form; stemming word variants; and removing stop words, punctuation marks and numbers. And Document Frequency (DF) threshold is applied to select features of news items.

Two weighting schemes, Term Frequency (TF) and Term Frequency by Inverse Document Frequency (TF*IDF), are used so as to weight the features in news documents to construct news by features matrix, which is fed to the learning algorithm. This study considers one of the neural networks learning methods called Learning Vector Quantization (LVQ), to see its suitability for automatic Amharic text news classification. In the course of this study, it is found that TF weighting scheme outperforms TF*IDF weighting scheme by 3.54% on average. Using the TF weight method, 94.81%, 61.61% and 70.08% accuracies are obtained at three, six and nine categories experiments respectively with an average of 75.5% accuracy. For similar experiments, the application of TF*IDF weight method resulted in 69.63%, 78.22% and 68.03% accuracies with an average of 71.96% accuracy.

Previous research works on Amharic text classification show that, accuracy decreases consistently with the increase in categories. The result of this study shows that accuracy does not depend on the number of news items and categories considered; rather, representing each category with enough number of subclasses determines accuracy. Therefore, further works focusing on finding the optimum number of subclasses is the major direction of research with regard to Amharic text news classification using LVQ.

CHAPTER ONE

INTRODUCTION

1.1 Background

During the last twenty four years, the number of documents in the digital form has grown enormously in size with the introduction of Internet as a medium of information transfer and sharing. As a result, it is really advantageous to be able to automatically organize and classify documents (Novovicová, 2005). As Klein (2004) noted, the dramatic increase in textual information forced us to use automated classification system for the management of textual information. Automatic text classification is attractive because it frees organizations from the need of manually organizing documents, which can usually be too expensive, error prone or may not totally be feasible within the given time (Sebastiani, 2005).

Supervised learning and unsupervised learning are the two common methods for classifying textual information. We may be given a set of documents with the aim of establishing the existence of classes or clusters in the document. Or, we may know for certain that there are classes and the aim is to establish a rule where by we can classify a new document into one of the existing classes. The former type is called unsupervised learning and the later is supervised learning (Giorgino, 2004; Michie, Spiegelhalter and Taylor, 1994; Skarmeta, Bensaid and Tazi, 2000).

Text categorization is now a fairly mature technology that has delivered working solutions in a number of application contexts. But, a number of challenges remain (Sebastiani, 2006).

According to Sebastiani (2006), delivering high accuracy in all application contexts is one of the challenges for text classification. Effective classifiers have been produced for application domains such as the thematic classification of professionally authored texts such as newswires. But accuracies are poor in some application domains like classification of web pages, where the use of text is more versatile; spam filtering, spammers adapt their spamming strategies to the latest spam filtering technologies; and authorship attribution. Another challenge is the labeling of categories according to Sebastiani (2006); that is, manually classifying documents for use in the training phase is costly. Because text classification needs large amount of preclassified documents to build classifier for new unlabeled documents.

According to Sebastiani (2002), for accomplishing text classification task, there are a number of learning techniques. The common ones include Probabilistic methods, Regression methods, Decision Tree and Decision Rule learners, Neural Networks, Batch and Incremental learners of linear classifiers, Example-based methods, Support Vector Machines, Genetic Algorithms, Hidden Markov models and Classifier committees. This study considers neural networks algorithm called Learning Vector Quantization.

According to Demuth and Beale (2004), neural networks can be trained to solve problems which are difficult for conventional computers or human beings. Neural networks can be trained to perform complex functions in various fields of application including, pattern recognition, writer identification, text classification, speech recognition, computer vision and control systems.

Neural networks can be used for text classification tasks among others. Algorithms include Back Propagation Network (BPN), Self Organizing Map (SOM), Learning Vector Quantization (LVQ), etc (Martín-Valdivia, Ureña-López and García-Vega, 2007). The authors added that the most widely used algorithm is Back Propagation Network (BPN) and a number of text classification tasks are carried out using Self Organizing Map. They said that research results show that LVQ is better than SOM. This idea is also expressed by Haykin (1999). In fact, Michie, Spiegelhalter and Taylor (1994) and Anderson (2006) indicated that LVQ is the most efficient classification algorithm. Heuristic simplicity in LVQ algorithm can be adapted to text classification tasks though the algorithm is not exploited for such task as noted by Martín-Valdivia, Ureña-López and García-Vega (2007).

1.2 Statement of the Problem and Its Justification

The automated classification (categorization*) of texts has been flourishing in the last decade or so due to incredible increase in electronic documents on the Internet; this renewed the need for automated text classification (Klein, 2004). Extensive study is done for English language (Maly and et. al., 2007). But booming interest is shown even if extensive study is made (Sebastiani, 2005).

Amharic is technologically under resourced language (Solomon and Menzel, 2007); therefore, a lot has to be done. Only three researches have been tried on the area till now as to the researcher's knowledge. Effort has to be exerted to come up with the best classification performance.

* Text categorization and text classification are used to refer the same concept.

Amharic is the native language of people living in the north central part of Ethiopia. The language is also spoken as a second language in many parts of the country. Significant number of immigrants in the Middle East, Asia, Western Europe and North America also speak Amharic (Encyclopedia Britannica, 1992). Amharic language has its own writing system that uses Ge'ez alphabet. Recently, there are so many Amharic electronic documents such as web pages, word documents, articles, etc. So, it is helpful to undergo research on this important language to contribute in the process of enriching the language with technology.

News articles are among the most popular and regularly accessed content on the web. There is a need to have automatic document classification among others for managing large number of news articles (Calvo, Lee and Li, 2004). Most researches are carried out on Reuter's dataset (Maly and et. al, 2007) for the categorization of news articles into sport, politics, economics, social, etc.

There are also so many news articles which are produced and stored in Amharic. More specifically, Ethiopian News Agency (ENA) produces and stores so many news articles. They have web site that releases news in Amharic and English. Now, the agency uses ENASoft software for the management of news. But the classification task is done manually. Currently, there are 116 categories available in ENA. Among these, 13 are major categories and 103 are sub categories. Using manual classification is challenging for these large number of classes*.

Timeliness property of news can be maintained by using automated classification and the burden of human experts can be avoided (Zelalem, 2001; Surafel, 2003). Surafel added that it would be possible to store old news for further retrieval. Besides, Zelalem said, time is wasted in the process of training reporters for news articles classification purpose. In accordance with

* Category and class are used to refer the same concept.

Blumberg and Atre (2003), manual classification can achieve a high degree of accuracy though domain experts occasionally disagree on how to categorize a document, which is not the case in automatic classification. And manual classification is labor intensive than automated technique.

Hence, more effort is required by researchers and developers in the area of text classification to favor technologically under resourced language-Amharic. Research has to show which technique, tool, etc is best for Amharic text classification. It was with this view that Zelalem (2001), Surafel (2003) and Yohannes (2007) had done their research using, Statistical method, K-Nearest Neighbor (KNN) and Naïve Bayes, and Support Vector Machine (SVM) and decision tree respectively. Previous research works on Amharic text classification are discussed in Chapter Two of Section 2.7. This work also aimed at contributing in the area of Amharic text classification.

The great problem of news classification is the decrease in accuracy as the number of categories and news increases. Previous researches by Surafel (2003) and Yohannes (2007), who have made experiments on increasing levels of categories, prevail the idea as indicated in Table 2.6 of Chapter two. Since many, 13 major and 103 sub categories involved in news articles of ENA, such problem is the major challenge to implement automatic text classification system for news articles. Surafel (2003) and Yohannes (2007) recommended a different approach for automatic Amharic text classification. Hence, the aim of this study is, to see the performance of a neural network based approach, LVQ, on automatic Amharic text news classification at increasing levels of categories and news items.

This research work attempts to answer the following questions.

- Is neural network approach using LVQ learning method feasible for automatic Amharic text news classification?
- Can we reduce the effect of increasing number of categories and news items on Amharic text news classification performance using LVQ learning method?
- What is the effect of TF and TF*IDF weighting methods on Amharic text news classification performance?

1.3 Objective of the Study

The general and specific objectives of this research work are described here under.

1.3.1 General Objective

The general objective of this study is to investigate automatic Amharic text news classification using neural networks based approach with specific consideration of LVQ algorithm.

1.3.2 Specific Objectives

The following specific objectives are incorporated for achieving the general objective.

- To preprocess news documents so as to make the document ready for classification task.
- To build classifier based on the Learning Vector Quantization so that classification is applied on test dataset.
- To evaluate the performance of the Amharic news classifier.
- To provide concluding remarks and recommendations for further research.

1.4 Methodology

Methodology provides an understanding of how a research is conducted and organized in order to obtain information that is helpful for developing the design of a research (Monica Ines, 2001). The methods used in this research are described below.

1.4.1 Literature Review

Literature from books, journals, Internet, etc, have been reviewed. The review is to understand the concept of text classification, approaches of text classification, the methods used for preprocessing and classifier construction, concept of neural networks and LVQ, and text classification using neural networks in general and LVQ in particular. Amharic writing system has also been reviewed to understand Amharic Language characteristics which are helpful for preprocessing Amharic news.

1.4.2 Data Source and Datasets Preparation

Structured interview with the Information and Communication Technology (ICT) coordinator of ENA was conducted concerning the data (news), how news are created and distributed, and any software that they use for news management. The interview guide is attached in Appendix 1.

The data source selected for this study is Ethiopian News Agency (ENA). ENA is selected as a data source because there is no standard Amharic news corpus ready for text classification task as to the researcher's knowledge. And the other reason is that, the data is categorized into predefined categories manually by the experts which are suitable for supervised learning approach used in this study. Six years data, from 2003-2008, have been considered for analysis.

But the majority of news items are from 2007 and 2008. According to the ICT coordinator of ENA, most news items of the four years from 2003-2006 are lost due to damage.

Nine categories have been selected based on random sampling. Since each category is equally important for this study, every category is given equal chance to be selected based on random sampling. The nine categories have 1, 762 news items totally. After preprocesses are carried out all the remaining data have been used for experimentation, which accounts 1, 463.

After the data have been preprocessed, training and test sets are prepared; 66.67% constitutes the training dataset and 33.33% comprises the test dataset.

1.4.3 Automatic Amharic Text News Classification Design

The three important steps used in this study are, preprocessing, classifier building and classifier evaluation.

Removing irrelevant news items due to various errors is the first step. The next preprocesses aims in selecting important features that can represent and discriminate news items. These include normalization (removing punctuation mark and converting varying Amharic characters with the same sound to one common form), tokenization, removing stop words and numbers, stemming, term weighting and dimension reduction for feature selection.

Matrix is generated with features representing rows and news as columns and the value of each cell is either TF or TF*IDF weight methods. Hence, for each experiment, there are two matrices, one is using TF weighting scheme and the other is using TF*IDF weighting scheme.

After preprocessing is done, the data is partitioned into training and test sets. The training dataset is used to build (construct) the Amharic text classifier or model based on learning. Finally, evaluation of the system is made using the test dataset. Accuracy is used to evaluate the system, which is computed as the percentage of correctly classified news items.

1.4.4 Development Tools and Experimentation Method

For developing Amharic news text classification, a number of tools are used, which are discussed here under.

Visual Basic 6 is used to export the contents of Head line, Keyword and slug attributes from SQL server to a folder named 'NEWS', in doing so file names have been generated automatically for the text files. Visual Basic 6 is used because of its flexibility to deal with databases and researcher's familiarity.

For Amharic text preprocessing tasks, such as normalization, tokenization, stemming, stop word and number removal and weighting words, Python 3.0, integrated with NLTK (Natural Language Tool Kit), is employed. The reason to use Python is its convenient nature and powerfulness to work on text processing.

Learning Vector Quantization (LVQ) algorithm is used to train neural network classifier. LVQ is first proposed by Kohonen in 1990. It is the class of supervised learning algorithms employed to find the elements of which will "best" represent the classes called prototypes*. The number of classes and the number of prototypes are predefined. A labeled training dataset is used in LVQ learning method. LVQ is easy to implement and intuitively clear. This makes it interesting for

* In LVQ, subclass and hidden neuron refer the same concept and subclasses of a certain class are prototype or codebook for that class.

researchers and practitioners who are searching for robust classification schemes without the black box character of many neural network methods (Ghosh, 2007).

MATLAB 7.0 is used to build classifier using the selected algorithm. The tool is selected because of its availability. Additionally, the Neural Network toolbox of MATLAB provides the following functionalities among others (MathWorks, 2009).

- Graphical user interface (GUI) for creating, training and simulating neural networks.
- Provides support for the most commonly used methods of supervised and unsupervised network architectures.
- Modular network representation that enables an unlimited number of input setting layers and network interconnections including graphical view of network architecture.
- Visualization functions and GUI for viewing network performance and monitoring the training process.
- Preprocessing and post-processing functions and Simulink blocks for improving network training and assessing network performance.

1.5 Scope and Limitation of the Study

The scope of the study is to investigate the potential application of neural networks approach-LVQ to automatic categorization of Amharic text news articles of ENA. Nine categories have been considered, these are Bank and insurance, Tourism development, Mines and energy, Information and communication technology (ICT), Art, Educational coverage, Weather forecast, Religious assemblies and reports, and Creativity work.

Corpus other than news texts is not considered. From the news agencies available in Ethiopia (ENA and Walta Information Center), only ENA news are considered. In ENA, there are 13 major and 103 sub categories but only nine categories are selected.

LVQ works better if the raw data is normalized between -1 and 1 according to Thulasiraman (2005). Russell, Eberhart and Shi (2007) recommended that the raw data has to be tried and analyzed before the normalized one is tried. Their suggestion is followed but time frame inhibits to try the normalized one. In MATLAB, there are two types of LVQ, LVQ 1 and LVQ 2.1. According to Demuth and Beale (2004) of MATLAB, performance may be improved if LVQ 2.1 is applied after LVQ 1. But only LVQ 1 is used in this study.

The reason, for not taking into account the above considerations, is to focus on the impact of two weighting schemes on Amharic text news classification. Additionally, based on the result of the first experiment, three categories experiment using TF weight method, switching epoch gives better result. Hence, the trend is followed and nine epoch levels are tried for all experiments using TF and TF*IDF weight methods. 54 experiments are carried out for the nine epoch levels in the three, six and nine categories experiments using both TF and TF*IDF weight methods without the preprocessing experiments. The consideration of all these disallow to try the fore mentioned important considerations, given time constraint.

1.6 Application of the Study

Document classification has practical importance. The applications of text categorization listed by Novovicová (2005) and Sebastiani (2005) testify its great value. These are spam filtering, mail routing, e-mail filtering, news monitoring, selective dissemination of information to information

consumers, automated indexing of scientific articles, automated population of hierarchical categories of web resources, identification of document genre, authorship attribution, survey coding, and so on. According to Sebastiani (2005), its application can be extended to speech categorization by combing speech recognition and text categorization for phone call routing, image classification using captions, for question answering system by classifying questions so that the overall system can be enhanced.

The study conducted on news articles of ENA, can be extended for similar industries in the country. The result can be used as a starting point to do research in those domain areas.

1.7 Organization of the Thesis

This thesis is organized into five chapters. The First Chapter (this Chapter) is the Introduction that contains background, statement of the problem and its justification, objective of the study, methodology, scope and application of the study.

Chapter Two describes the concept and approaches of text classification, steps involved in text classification and networks in general and LVQ learning method in particular. Chapter Two also provides the details of Amharic writing system and the tool called System for Ethiopic Representation in ASCII (SERA), which facilitates Amharic computerization.

Chapter Three is Methodology discussion. It elaborates the methods and algorithms applied for designing the system. Chapter Four is the Experimentation and Performance Evaluation. It presents implementation details, experimental results, analysis and finding of the study. Chapter Five pinpoints concluding remarks and the recommendations forwarded for further research work.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

The task of classification occurs in a wide range of human activity. The term could cover any context in which some decision or forecast is made on the basis of currently available information. As a result, classification procedure is required for making judgments in new situations based on the information provided (Michie, Spiegelhalter and Taylor, 1994). Textual information is the one that has to be classified for efficient and effective utilization for a given purpose.

The volume of text content grows continuously online and in corporate domains. Such information available in digital form has to be organized in a manner suitable for use; such need has generated and progressively intensified the interest in automatic text categorization. Hence, automatic text categorization, acting as a way to organize the text content, becomes interesting not only from academic but also from industrial point of view (Giorgino, 2004; Klein, 2004; Liao, Alpha, and Dixon, 2003). It is with this consideration that the task of text classification has got attention from researchers and developers in the last fifteen years though its history dates back in 1960 (Klein, 2004; Sebastiani, 2005).

The subsequent sections discuss the concept of text classification, approaches used in text classification, steps used in text classification, neural networks in general and Learning Vector

Quantization in particular, Amharic writing system and previous research works on automatic Amharic text classification.

2.2 Meaning of Text Classification

The concept of text classification is defined by a number of authors in similar way. For example, Bi, Murtagh and Anderson (1999); Blumberg and Atre (2003); Giorgino (2004); Ifrim, Theobald and Weikum (2005); Klein (2004); Liao, Alpha and Dixon (2003); Martín-Valdivia, Ureña-López and García-Vega (2007); Michie, Spiegelhalter and Taylor (1994); Sebastiani (2005); Skarmeta, Bensaid and Tazi (2000); Wang and et. el. (2005); and Yi and Beheshti (2004) defined text classification as the task of automatically assigning a set of documents into categories (or classes, or topics) from a predefined set. The definition given by Sebastiani (2002) and Klein (2004) clarifies the concept of text classification more.

Klein (2004) and Sebastiani (2002) defined text categorization as a mapping of text documents to categories. To clarify, if C= $\{c_1, c_2 \ldots c_m\}$ is a set of categories (classes) and D= $\{d_1, d_2, \ldots, d_n\}$ is a set of documents, the purpose of text classification is assigning c_i to d_j ($1<=i<=m$ and $1<=j<=n$) a value of 0 if the document d_j does not belong to c_i; otherwise a value of 1. The mapping is sometimes referred to as the decision matrix (Klein, 2004) and it is depicted in Table 2.1.

	$\mathbf{d_1}$	...	$\mathbf{d_j}$	...	$\mathbf{d_n}$
$\mathbf{c_1}$	A_{11}	...	a_{1j}	...	a_{1n}
...	...	...	...	...	...
$\mathbf{c_i}$	a_{i1}	...	a_{ij}	...	a_{in}
...	...	...	...	...	...
$\mathbf{c_m}$	a_{m1}	...	a_{mj}	...	a_{mn}

Table 2. 1: Document to Category Matrix

In Table 2.1, $d_1 \ldots d_n$ refers set of documents, $c_1 \ldots c_m$ refers set of categories and $a_{11} \ldots a_{mn}$ represent a value of 0 if the document does not belong to that category, otherwise a value of 1.

Depending on the application, text classification may be either a single-label task or multi-label task. Single-label task is assigning exactly one category to a document. And multi-label task is assigning one category or more categories for a given document. A special kind of single-label classification is binary text classification, in which a document is going to be classified in either of the two available categories (Sebastiani, 2005).

Based on Sebastiani (2005), text classification is a subjective task in the sense that two experts, human or artificial, may disagree on the decision of the category to be assigned for a document. A news article could be filed under Politics, Finance, Sport, or any other category, or even under neither, depending on the subjective judgment of the expert. Because of this, the meaning of a category is subjective.

2.3 Text Classification Approaches

In accordance with Blumberg and Atre (2003), there are four approaches to text classification, which are manual classification, rule-based classification, supervised learning and unsupervised learning.

2.3.1 Manual Classification

Manual classification is the assignment of one or more categories to documents by human experts. These experts have domain knowledge and familiar with the category structure being used (Blumberg and Atre, 2003).

2.3.2 Rule-based Classification

In this approach, keywords or Boolean expressions are used to categorize a document. Such method is more convenient if a category can be described using few words. The method is uncommon to large scale classification system (Blumberg and Atre, 2003).

2.3.3 Supervised Learning

This is a learning mechanism which uses manually classified documents for training purpose. From the training, a model or classifier is constructed so that new unseen document will be classified based on the model constructed for each category (Blumberg and Atre, 2003; Giorgino, 2004; Michie, Spiegelhalter and Taylor, 1994; Skarmeta, Bensaid and Tazi, 2000).

2.3.4 Unsupervised Learning

With this learning approach, preclassified documents are not required since the method tries to exploit regularities found in the document and make group or cluster based on similarity. The method, also called clustering; it may not found categories which are intuitive to humans (Blumberg and Atre, 2003; Giorgino, 2004; Michie, Spiegelhalter and Taylor, 1994; Skarmeta, Bensaid and Tazi, 2000).

For both supervised and unsupervised learning, classification must be accomplished only on the basis of knowledge extracted from the documents themselves because the categories tell no meaning or do not contain any knowledge like publication date, document type, publication source, etc (Sebastiani, 2005).

2.4 Text Classification Phases

The phases with regard to text classification are, feature preparation, term weighting, dimension reduction, classifier learning and classifier evaluation (Sebastiani, 2005).

2.4.1 Feature Preparation

The text has to be converted into features in order to transform a document into a feature vector, which is an input for the classifier learning. Hence, this phase is basically the first step in the preprocessing stage of the data. Feature formation must be performed with reference to the definition of the features. Features may be tokens (single stemmed or non stemmed words) or phrases. Features such as single tokens or single stemmed tokens are most frequently used in text categorization. In this bag-of-words representation, information about dependencies and the relative positions of different tokens are not used. Phrasal features consisting of more than one token are one possible way to make use of the dependencies and relative positions of component tokens. However, tokens or stemmed tokens show better classifier performance than phrases (Liao, Alpha and Dixon, 2003).

For both classification and retrieval of natural language text documents, the standard document representation is a term vector where a term is simply a morphological normal form of the corresponding word (Ifrim, Theobald and Weikum, 2005). According to Skarmeta, Bensaid and Tazi (2000), a word is a string of characters delimited by spaces in the context of English language. The representation of a document using token is said to be document indexing from Information Retrieval point of view (Baeza-Yates and Ribeiro-Neto, 1999).

Document indexing denotes the activity of mapping a document d_j into a compact representation of its content that can be directly interpreted by a classifier building algorithm and by a classifier, once it has been built. The document indexing methods usually employed in text classification are borrowed from Information Retrieval, where a text d_j is typically represented as a vector of term weights. Indexing method is characterized by defining what a term is, and a method to compute term weights (Sebastiani, 2005).

Before a document is indexed, it passes through one or more text processing steps. These steps as dealt in Baeza-Yates and Ribeiro-Neto (1999) are, lexical analysis, stop words removal, identification of noun groups, stemming and finally indexing. The steps are not sequential. For example, words without removing stop words can be taken as index terms. Figure 2.1 clarifies this fact.

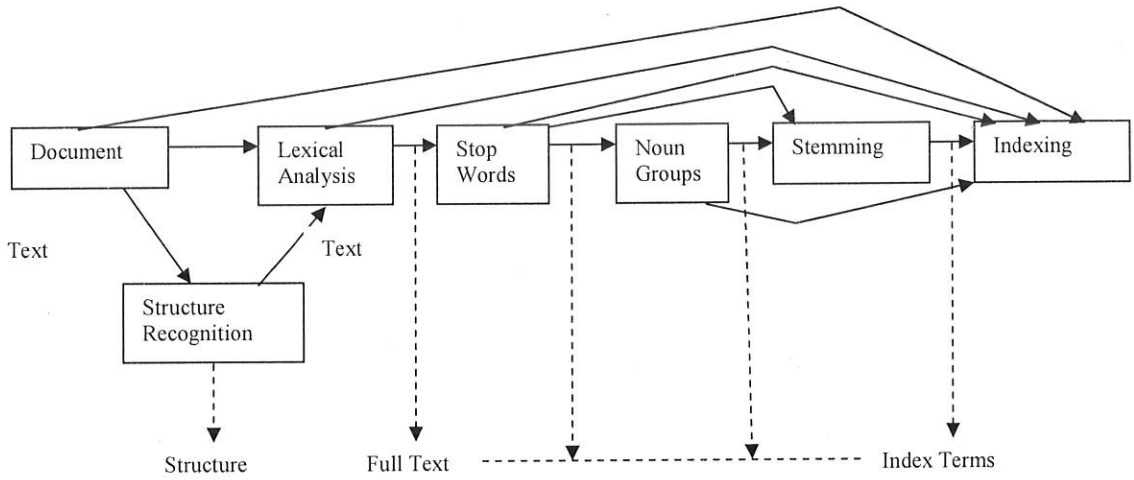

Figure 2. 1: Steps for Document Representation

The steps depicted in Figure 2.1 are discussed in the subsequent part based on Baeza-Yates and Ribeiro-Neto (1999).

Lexical Analysis: concerned about converting a given document to individual words or tokens by the process known as tokenization. It deals with the identification of term separators, accents, spacing etc. It also involves a decision on considering or disregarding special characters like hyphens, digits, punctuation marks, spaces, letter cases, etc. Usually punctuation marks, numbers, spaces, etc are ignored.

Stop Words Removal: Stop words are non-content bearing words in a document. They are words which occur frequently and have less power in discriminating one document from another. Some are used for syntactic purpose; for instance, articles, prepositions, conjunctions, etc. Stop words are normally removed from a document to facilitate effective representation of a document.

Noun Groups: deals with the identification of nouns since nouns convey more meaning in a document collection. In this step, correcting spelling errors and combining similar words using a thesaurus are also included.

Stemming: The same word normally permitted to take various forms for language usage based on the grammatical rule of the language. Stemming therefore, deals about converting these varying forms of the same word into one-root form. Figure 2.2 shows Taxonomy of stemming algorithms by Frakes and Baeza-Yates (2002).

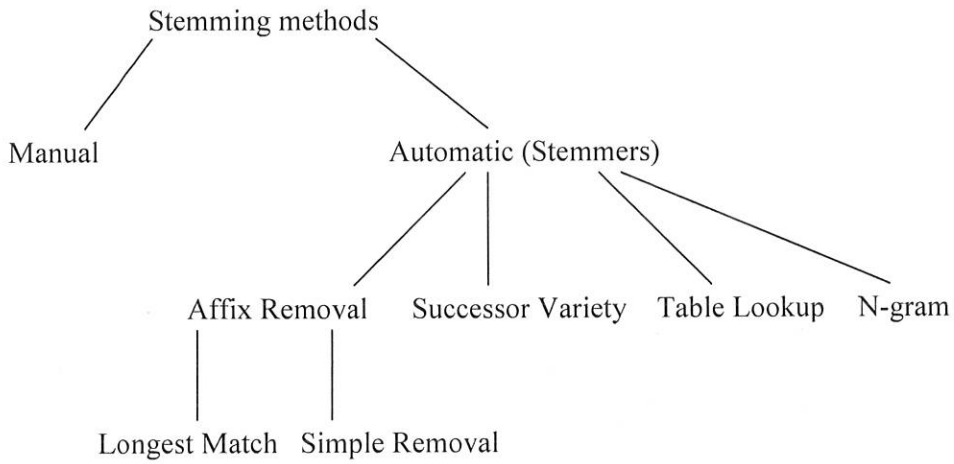

Figure 2. 2: Taxonomy of Stemming Algorithms

According to Frakes and Baeza-Yates (2002), affix removal algorithms, the method used in this study, remove suffices and/or prefixes from terms. Porter Stemmer is an example of affix removal algorithm. According to Hooper and Paice (2005), Porter Stemmer is the most common widely used affix removal algorithm, which applies rules on the word to remove affixes in order to find its stem.

After carrying out at least the first step (lexical analysis), index terms can be generated as a representation of a document.

2.4.2 Term Weights

The importance of an index term to a document is shown by using weight (Giorgino, 2004; Liao, Alpha and Dixon, 2003). Term Frequency (TF), Inverse Document Frequency (IDF) and Term Frequency by Inverse Document Frequency (TF*IDF) are common weighting methods to show the importance of a term (Manning, Raghavan and Schütze, 2008; Baeza-Yates and Ribeiro-Neto, 1999). The three weighting methods are discussed based on the two sources.

TF: is the number of occurrences of a term in a document. The weight of term k in document i, is given by:

$$TF = FREQ_{ik}$$

Formula 2. 1: Term Frequency

In Formula 2.1, $FREQ_{ik}$ is the number of occurrence of term k, in document i. TF is zero if the term does not appear in document i.

IDF: is a measure of the general importance of the term. Formula 2.2 depict IDF of a term.

$$IDF = \log_2^{\frac{N}{d_k}}$$

Formula 2. 2: Inverse Document Frequency

In Formula 2.2, N is the total number of documents in the collection, d_k the number of documents in which term k occurs.

TF*IDF: As the name implies, TF*IDF is the combination of TF and IDF weighting methods. TF*IDF incorporates two intuitions.

a) If an index term occurs more frequently in a document, the index term is more important for that document, the Term Frequency intuition.
b) If more number of documents contain the index term, the index term is less discriminating between the documents, the Inverse Document Frequency intuition.

Formula 2.3 shows TF*IDF weight of term k.

$$TF * IDF = FREQ_{ik} * \log_2^{\frac{N}{d_k}}$$

Formula 2. 3: TF*IDF

In Formula 2.3, $FREQ_{ik}$ is the number of occurrence of term k in document i, N is the total number of documents in the collection, d_k the number of documents in which term k occurs.

2.4.3 Dimension Reduction

A major difficulty in text categorization is the high dimensionality of the feature space. The feature space comprises one new dimension for each unique term that occurs in the text documents, which can lead to tens of thousands of dimensions for even a small-sized text collection. The dimensionality of the feature space often exceeds the number of available training documents; this is an obstacle for learning algorithms. So, it is desirable to integrate a dimensionality reduction phase with text classification task (Skarmeta, Bensaid and Tazi, 2000; Yi and Beheshti, 2004).

Feature selection is often used for dimension reduction. Among feature selection methods Document Frequency thresholding (DF) and Information Gain (IG) are frequently used methods for choosing a subset of the available features (Krishnakumar, 2006). According to Krishnakumar (2006), DF, used in this study, is the number of documents in which the word occurs. Predetermined threshold is used to remove words which have document frequency less than the threshold value. And IG is the number of bits of information obtained for category prediction by knowing the presence or absence of a word in a document.

2.4.4 Text Classifier Learning

A text classifier for a category is automatically generated by a general inductive process (the learner) by observing the characteristics of a set of preclassified documents, which dictates the characteristics that a new unseen document should have in order to belong to a certain category. So as to build classifiers for a category, there is a need to have a set of documents for which the category is known. In experimental text classification, it is customary to partition the set of text documents into training set and test set. The training set is the set of documents from which the learner builds the classifier and the test set is the set on which the effectiveness of the classifier is evaluated (Sebastiani, 2005).

2.4.5 Text Classifier Evaluation

The performance of text classifier can be evaluated by considering effectiveness or efficiency aspects. Efficiency depends on volatile parameters like computer hardware/software on the other hand effectiveness depends on the correctness of the classifier (Lavesson, 2003). In a single-label classification, accuracy is the major effectiveness measure (Sebastiani, 2005). In this single-label classification study, accuracy is used for evaluation of Amharic text classifier. According to Skarmeta, Bensaid and Tazi (2000), accuracy is computed as the percentage of correctly classified documents.

Table 2.2 is a contingency table (Goyal, 2009) for which the formula of classifier accuracy is based.

Category c_i		Expert Judgments	
		Class=Yes	Class=No
Classifier Judgments	Class=Yes	TP	FP
	Class=No	FN	TN

Table 2. 2: Contingency Table for Computing Classifier Effectiveness

In Table 2.2, TP (True Positive) is the number of test documents correctly classified, FP (False Positives) is the number of test documents incorrectly classified, FN (False Negatives) is the number of test documents incorrectly not classified and TN (True Negatives) is the number of test documents correctly not classified. Hence, accuracy, A, is computed using Formula 2.4.

$$A = \frac{TP + TN}{TP + TN + FP + FN}$$

Formula 2. 4: Accuracy of Text Classifier

2.5 Neural Networks

The intension of using neural networks is to model natural processing of the human brain. The brain is highly complex, nonlinear and parallel computing system. It has the capability to organize its constituent neurons, information processing elements, to perform computations like pattern recognition, perception, motor control, etc; faster than the digital computers with such tasks (Haykin, 1999).

The discussion on biological neural networks, artificial neural networks and Learning Vector Quantization (LVQ) follows.

2.5.1 Biological Neuron

A neuron has a cell body, a branching input structure: the dendrite, and a branching output structure: the axon, which connects to dendrites via synapses. Electro-chemical signals are propagated from the dendritic input, through the cell body, and down the axon to other neurons (Stergiou and Siganos, 1997). Figure 2.3 depicts the structure of biological neuron.

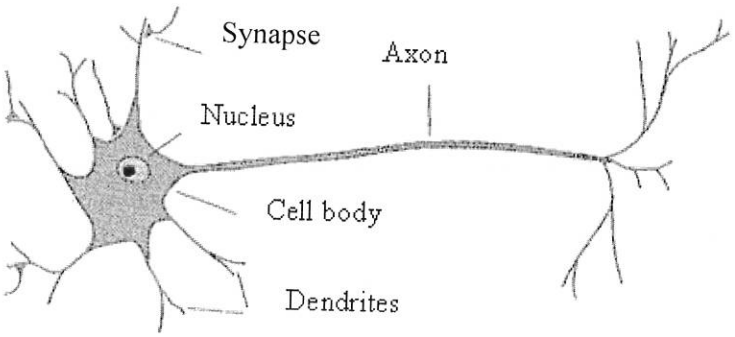

Figure 2. 3: Structure of Biological Neuron

A neuron fires, produces output, if its input signal exceeds a certain amount (the threshold), in a short period of time. Synapses vary in strength. Good connections allowing a large signal and slight connections allow only a weak signal. Each neuron produces only one output signal. The output signal is transmitted through the neurons outgoing connection. The outgoing connection splits into a number of branches. The outgoing branches terminate at the incoming connections of other neurons (Stergiou and Siganos, 1997).

2.5.2 Artificial Neural Networks (ANN)

A neural network is a massively parallel distributed processor made up of simple processing units called neurons. The processing units have a natural tendency for storing experimental knowledge

and making it available for later use. Two important things make neural networks similar to the brain. The first one is, it acquires knowledge from the environment through a learning process using the network. The other one is, the acquired knowledge is stored as weights, which shows interneuron connection strength (Haykin, 1999).

2.5.3 Model of Artificial Neuron

Neuron is information processing element that is fundamental to the operation of the network. There are three constituents of neural networks (Haykin, 1999).

1. **Synapses:** are connecting links, which have weights that shows its strength. Weights represented as w_{kj} refers k neuron with j weight.
2. **Adder:** that sum input signals from sending neuron/s to receiving signals.
3. **Activation Function (Squashing Function):** for limiting the amplitude of neuron output.

Figure 2.4 illustrates the details of neuron model

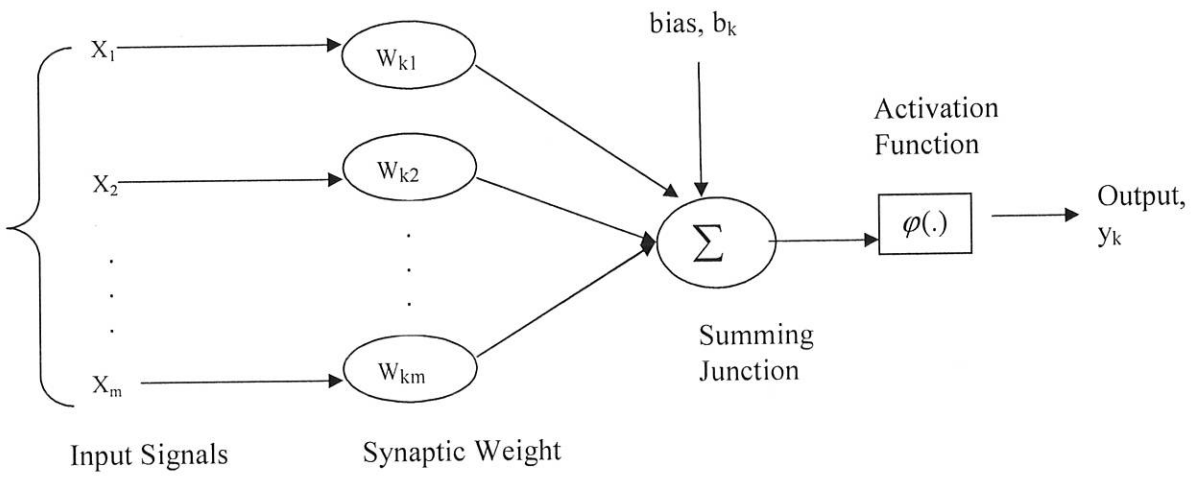

Figure 2.4: Neuron Model

In Figure 2.4, x_1, x_2,…,x_m refer to input signals; w_{k1}, w_{k2}, …, w_{km} are synaptic weights of neuron k; bias b_k has the effect of, increasing: for positive values or lowering: for negative values, the net input of the activation function.

Formula 2.5 shows u_k, which is linear combiner due to input signals.

$$u_k = \sum_{j=1}^{m} w_{kj} . x_j$$

Formula 2. 5: Linear Combiner due to Input x_i

In Formula 2.5, x_1, x_2,…,x_m refer to input signals; w_{k1}, w_{k2}, …, w_{km} are synaptic weights of neuron k.

After applying bias b_k, input to the activation function v_k is calculated as:

$$v_k = u_k + b_k$$

Formula 2. 6: Linear Combiner with Bias

The bias is an external parameter that we can ignore, in such condition $v_k=u_k$.

Finally, the output of the neuron y_k becomes:

$$y_k = \varphi(v_k)$$

Where, φ is the activation function.

Formula 2.7: Output of Neural Network

2.5.4 Architecture of Artificial Neural Network

Neurons are normally grouped into layers; layers are groups of neurons which perform similar functions. Usually, there are three layers of neurons; input layer, hidden layer, and output layer. The layer of neurons that receive input from the user program and sends the pattern to the next layer (hidden layer) is the input layer. Hidden layer (one, more or none in some learners), is between input layer and output layer. This layer presents information to the succeeding layer, output layer. Finally, pattern is generated by the output layer (Heaton, 2005).

Normally, there are two basic topologies of neural networks feed-forward networks and recurrent networks. In feed-forward networks, the data flows through zero, one or more succeeding hidden layers and then to the output layer in one direction. On the other hand, in recurrent networks, the data flows to all adjacent connected units and circulates back and forth until the activation of the units is stabilized (Artificial Neural Networks, 2008). The layered architecture of feed-forward neural networks, used in this study, is shown in Figure 2.5 based on Curtis (2002).

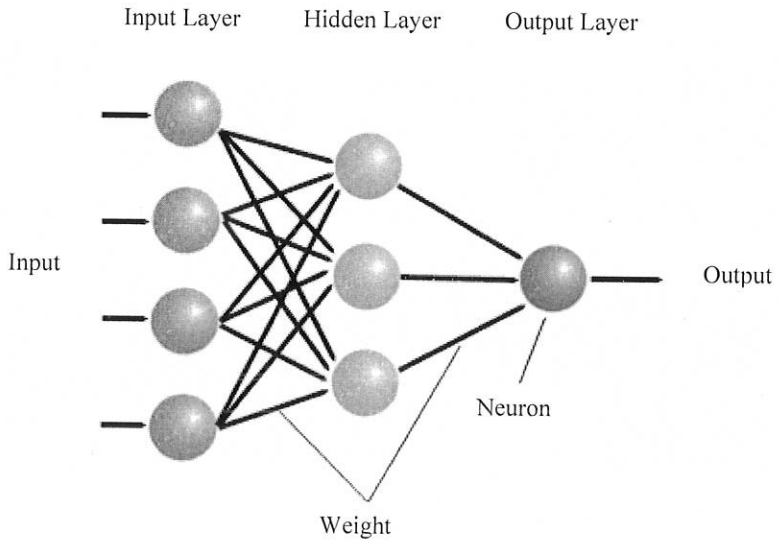

Figure 2.5: Architecture of Feed-Forward Neural Network

2.5.5 Learning in Artificial Neural Networks

One of the characteristics of neural networks is the ability to learn from its environment and to improve its performance accordingly. Learning is the process by which the free parameters of neural networks are adapted by simulation of the environment in which the network is embedded (Haykin, 1999). Haykin identified five types of learning rules. These are Error-Correction learning, Memory-Based learning, Hebbian learning, Competitive learning and Boltzmann learning.

Error-Correction Learning: is the technique of comparing the system output to the desired output value. In this learning, a neuron k is driven by a signal vector produced by one or more layers of hidden neurons, which are driven by input vector applied to the input layer. The output of the neural network is compared to the target output and an error is computed as a result.

Memory-Based Learning: All or most of past experience are explicitly stored in a large memory of correctly classified input-output examples. When classification of a test vector that is not seen before is encountered, the algorithm retrieves the training data and analyzes the neighborhood with the test data.

Hebbian Learning: states that if two neurons on either side of a synapse (connection) are activated synchronously (simultaneously), then the strength of that synapse is selectively increased. On the other hand, if two neurons on either side of a synapse are activated asynchronously, then that synapse is selectively weakened or eliminated.

Competitive Learning: The output neurons compete among themselves to become active (fired). In Hebbian learning, several output neurons may fire; whereas, in competitive learning only one

output neuron is active. Thus, competitive learning is suitable to classify input patterns. There are three basic elements of competitive learning.

1. A set of neurons that are all the same except synaptic weights; hence, neurons respond differently to a given set of input patterns.
2. A limit imposed on the strength of each neuron.
3. A mechanism that permits the neurons to compete to a given subset of inputs; as a result one output neuron is active at a time. The neuron that wins the competition is called a winner-takes-all neuron.

Boltzmann Learning: can be viewed as neurons with either of the two states s_i=+1 or s_i=-1. Every pair of neurons is connected by the bidirectional weights w_{ij}; if a weight between two neurons is zero, then no connection is drawn. The optimization problem is to find a configuration (assessment of all neurons) that minimizes the energy described by:

$$E = -\frac{1}{2}\sum_{i,j=1}^{N} w_{ij}.s_i.s_j$$

Where, E is the Energy, s_i and s_j are two neurons, and w_{ij} is the weight between neurons i and j.

Formula 2.8: Energy in Boltzmann Learning

2.5.6 Advantages of Using Neural Networks

Parallel architecture, noise tolerance, self-organization and generalization are the characteristics of neural networks that its benefits are based (Martín-Valdivia, Ureña-López and García-Vega, 2007). Haykin (1999) lists out the benefits of neural networks as nonlinearity, input-output

mapping, adaptivity, contextual information, fault tolerance, Very Large Scale Integrated (VLSI) implementation, uniformity of analysis and design, and neurobiological analogy.

Nonlinearity: is an important feature if the mechanism responsible for generation of the input signal is inherently nonlinear, for example speech signal.

Input-Output Mapping: Neural networks work in a set of examples to map input to an output. Each example consists of a unique input signal and a corresponding desired response. The network adjusts synaptic weights (free parameters) to produce the desired output. The training of the network continues until no change is observed in synaptic weights.

Adaptivity: Neural networks trained in a specific environment can be retrained to deal minor changes in the operating environment. If the system is non-stationary, adapting the system to the new situation is very fundamental to ensure stability and to robust performance.

Contextual Information: Every neuron in the network is affected by the global activity of all other neurons in the network. Thus, contextual information is dealt naturally by the neural network.

Fault Tolerance: A neural network degrade in performance instead of catastrophic failure due to the distributed nature of information stored.

VLSI Implementation: Due to parallel nature of neural networks it is possible to implement VLSI technology, which is helpful for capturing complex behavior.

Uniformity of Analysis and Design: The same notation is used in all domains involving the application of neural networks and neurons are ingredient for all types of neural networks.

Neurobiological Analogy: base for fault tolerant parallel processing. Additionally, fast and powerful processing is also possible. Hence, it provides a research tool for neurobiological and complex problem domains.

2.5.7 Learning Vector Quantization (LVQ)

Learning Vector Quantization, known for twenty years, is a competitive supervised learning algorithm that deals with selecting prototypes. The task is to find a map from R (inputs) into a finite set of labels Y (outputs). The classifiers are parameterized by a set of points, $\mu 1,....,\mu k \in R$, which is referred as prototypes formed from the training dataset. Each prototype is associated with a label $y \in$ Y. Given a new instance (input) $x \in R$, the label of the closest prototype will be assigned. The goal of the learning process in this model is, to find a set of prototypes which will predict accurately the labels of unseen instances. LVQ iterates over the training data and updates the prototypes position (Crammer, Gilad-Bachrach, and Navot, 2003).

2.5.7.1 LVQ Architecture

LVQ is supervised version of Kohonon neural network (Martín-Valdivia, Ureña-López and García-Vega, 2007).

LVQ network has two layers (Demuth and Beale, 2004), competitive layer and linear layer. The competitive layer learns to classify input vectors. The linear layer transforms the competitive layer's classes into target classes defined by the user. The classes learned by the competitive layer are referred as subclasses and the classes of the linear layer are called target classes. The subclasses are always larger than the target classes.

All the input units are fully-connected by feed-forward connections to all output units. The output units are interconnected through lateral inhibitory connections so that, when an output unit is activated, it uses the lateral connections to send inhibition signals and to be able to deactivate the rest of the output units (Martín-Valdivia, Ureña-López and García-Vega, 2007). Figure 2.6 indicates the architecture of LVQ according to Demuth and Beale (2004).

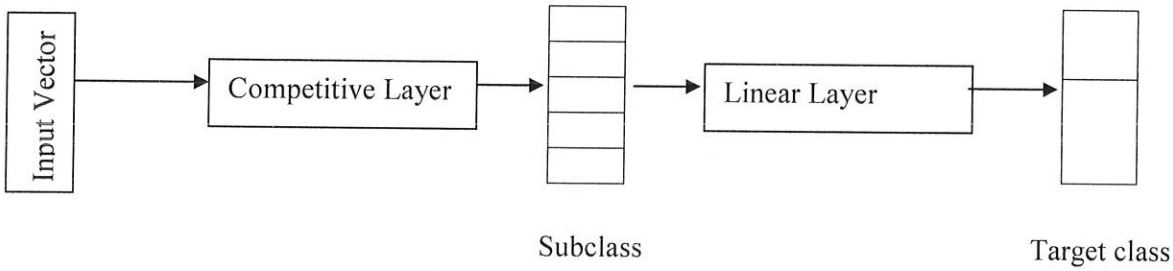

Figure 2.6: Architecture of LVQ

2.5.7.2 Types of LVQ

According to Umer and Khiyal (2007), the types of LVQ algorithms are LVQ 1 (Kohonen, 1990b), LVQ2.1 (Kohonen, 1990a), LVQ3 (Kohonen, 1990b) and the Optimized Learning rate algorithms OLVQ 1 (Kohonen, 1992). MATLAB 7.0 incorporates two of these algorithms, LVQ 1 and LVQ 2.1.

LVQ 1: According to Demuth and Beale (2004), an input vector P is presented, and the distance from P to each row of the input weight matrix is computed with the function ndist, a function to compute Euclidean distance. Layer one neurons (subclasses) compete for the input. The Euclidean distance of an input vector is computed with each codebook vector and the nearest codebook vector is declared the winner (Umer and Khiya, 2007; Martín-Valdivia, Ureña-López

and García-Vega, 2007). Demuth and Beale added that the class of the winning neuron (neuron with the smallest Euclidean distance) is assigned to the input as a target class. The algorithm for LVQ 1, which is used in this study, is presented on Chapter Three of Section 3.8.2.3.

LVQ 2.1: Similar to LVQ 1, the difference is, two vectors of layer one which are closest to the input vector may be updated providing that one belongs to a correct class and the other belongs to a wrong class. LVQ 2.1 is used only after LVQ 1 has been applied (Demuth and Beale, 2004).

2.5.7.3 Text Classification Using LVQ

A text classifier based on a neural network approach is a network of units, where the input units represent terms, the output units represent the categories of interest and the weight connections represent dependence relations. The neural classifier is trained by using a learning algorithm in order to modify and adjust the weights appropriately (Sebastiani, 2002).

LVQ algorithm is the supervised version of the Kohonen model and it was especially designed to accomplish pattern classification tasks (Martín-Valdivia, Ureña-López and García-Vega, 2007).

According to Martín-Valdivia, Ureña-López and García-Vega (2007), there are two layers of the LVQ algorithm-input layer and output layer. Codebooks are the weight vectors associated with each output units. The input layer has as many units as features retained (dimensions of the input vectors) and there is one output unit for each possible class. Each class of input space is represented by its own codebook. For text classification purpose, one codebook vector per class is used.

Since LVQ algorithm is a competitive network, output units compete among themselves in order to find the winner according to Euclidean distance to assign a category for each training vector (Martín-Valdivia, Ureña-López and García-Vega, 2007).

2.6 Amharic Writing System

Amharic is the working language of the Federal Government of Ethiopia; it is one of the Semitic languages. Twenty seven million people speak Amharic; with this, it is the second largest Semitic language next to Arabic. One of the major differences between Amharic and Semitic languages like Arabic and Hebrew is that Amharic is written from left to right as of English (Wapedia, 2009).

The subsequent parts discuss Amharic characters, punctuation marks and number systems. The problems in the processing of computerizing Amharic are also dealt. Finally, System for Ethiopic Representation in ASCII (SERA), which is a system to facilitate Amharic computerization, is described.

2.6.1 Amharic Characters

A character or a symbol, Fidel (ፊደል) in Amharic, is used to represent a phoneme, which is a combination of a vowel and a consonant (Tewodros, 2003).

The present writing system of Amharic is taken from Ge'ez alphabet. Ge'ez was used for literature in the early time in Ethiopia. The Amharic writing system consists of a core of thirty three characters each of which occur in basic form and in six other forms called orders. As shown in Table 2.3, the six orders represent the different forms of Amharic basic characters. Each form is made in accordance with the sound that goes with the symbol. Each character designates a

consonant together with its vowel; the vocalic symbol cannot be detached from the consonant element. Thus, Amharic does not use independent symbols for vowels (Bender and et. al., 1976). Bender and et. al. added that the non-basic forms are derived from the basic forms by more or less regular modifications.

First Order	Second Order	Third Order	Fourth Order	Fifth Order	Sixth Order	Seventh Order
ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ
Hä	Hu	Hi	Ha	He	H	Ho
ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ
Lä	Lu	Li	La	Le	L	Lo
መ	ሙ	ሚ	ማ	ሜ	ም	ሞ
Mä	Mu	Mi	Ma	Me	m	Mo

Table 2. 3: Amharic Characters Example

Five order characters known as labio-vellars are also there, for example, ኰ, ኱, ኲ, ኳ, and ጕ. Additionally, labialized consonants like ሏ, ሟ, ቧ, ኗ, etc are included in the character set (Ethiopia, 2002; Zelalem, 2001).

In Amharic characters, there is no capital letter and small letter distinction (Bender and et. al., 1976). Amharic characters are attached in Appendix 2.

2.6.2 Amharic Punctuation Marks

Identifying punctuation marks is vital to know word demarcation for natural language processing. According to Daniel (1994) cited in Tewodros (2003), the punctuation marks in Amharic are about ten though few of them used in computer writing system. 'Hulet Neteb' (':')-word separator and 'Arat Neteb' ('::')-sentence separator are the major punctuation marks. But, space

is mostly used instead of Hulet Neteb (':') specially in computer writing system. The punctuation marks together with their use are presented in Appendix 3.

2.6.3 Amharic Number System

The Amharic number system consists of twenty characters. They represent numbers one to ten, multiples of ten (twenty to ninety), hundred and thousand. The numbering system is not suitable for arithmetic computation because there is no representation for zero (0) symbol, no place value, no comma and no decimal point. Amharic numbering system is used in dates specially calendar; otherwise western numerals are used in most literature these days (Bender and et. al., 1976). Amharic numbers are presented in Appendix 4.

2.6.4 Problem of Amharic Writing System

There are a number of problems associated with Amharic writing system which are challenging natural language processing of Amharic documents; which are dealt below.

2.6.4.1 Redundancy of some Characters

Sometimes more than one character is used for similar sound in Amharic (Ethiopia, 2002; Zelalem, 2001). Though the various forms have their own meaning in Ge'ez, there is no clear cut rule that shows its purpose and use in Amharic according to Bender and et. al.(1976). Table 2.4 illustrates the different forms of Amharic characters with similar sound.

Character	Other form/s of the character
ሀ (hä)	ሐ and ኀ
ሠ (sä)	ሰ
አ (ä)	ዐ
ጸ (tsä)	ፀ

Table 2. 4: Amharic Characters with Different Forms of the Same Sound

The problem of the same sound with various characters is not only observed with core characters, but also exhibited in the same order of characters. For example, ሀ and ሃ, ጎ and ጋ; አ and ኣ; etc (Tewodros, 2003).

The use of various forms of characters for the same sound poses a problem in the process of feature preparation for the classifier learning since the same word is represented in different forms. For example, the word 'ጸሀይ' ('sun') can be represented in Amharic as ጸሀይ, ጸሐይ, ጸጎይ, ፀሀይ, ፀሐይ, ፀጎይ, etc.

2.6.4.2 Compound Words

There is no standard way of writing Amharic compound words (Bender and et. al., 1976). Space or hyphen is used between two words in a compound word; sometimes the words are merged together. According to Tewodros (2003), there is a meaning difference when compound words separated by space are treated separately. For example, the word 'ሆድ-ሰፊ' ('tolerant') formed from the words 'ሆድ' meaning 'stomach' and 'ሰፊ' meaning 'wide'. One can imagine how the meaning of the original word is diverted to different contexts.

2.6.4.3 Spelling Variation of the Same Word

The same word is written in various forms (Ethiopia, 2002; Tewodros, 2003; Zelalem, 2001). For example, the word 'ሰምዮአል' ('he hears') can be written in Amharic as ሰምዮአል, ሰምኢል, ሰምትዋል, etc. Spelling variation may happen also in the case of translating foreign word to Amharic. For instance, the word 'ቴሌቪዥን' ('television') can be written as ቴሌቭዥን, ቴሌቭዥን, ቴሌቪዥን, etc.

2.6.4.4 Abbreviation

No consistency is kept in abbreviating Amharic words (Ethiopia, 2002; Zelalem, 2001). The word '**ዓመተ ምህረት**', meaning 'AD', can be abbreviated as **ዓም**, **ዓ.ም**, **ዓ.ም.**, **ዓ/ም**, etc.

All the aforementioned problems pose challenges since the same word is treated in different forms in the process of feature preparation for text classifier. So, care should be taken to solve such problems.

2.6.5 System for Ethiopic Representation in ASCII (SERA)

SERA is a tool for processing Amharic documents. The fundamentals of SERA are discussed based on Daniel (1996).

SERA is a convention for translation of Fidel (**ፊደል**) script into Latin script that insures the integrity of the format and content of the original document, and that can be fully transportable across all computer mediums. SERA has been under continued development since early 1993.

The crux of the problem that demands translation is due to ASCII size. ASCII uses 7-bit encoding of computer letters, which means there are 128 addresses available to assign for letters that people and computers may use. But Fidel requires a 9-bit system for more than 360 addresses than ASCII can hold.

SERA is used for this research work to translate Visual Ge'ez Unicode format of Amharic script to Latin Script for processing of Amharic documents (news). The translation of Amharic script to Latin script is presented in Appendix 5 based on Daniel (1996).

2.7 Review of Related Research Works on Amharic Text Classification

As to the knowledge of the researcher, three researches have been done on automatic Amharic text classification by Zelalem (2001), Surafel (2003) and Yohannes (2007).

Zelalem (2001) used statistical method for the classification of Amharic news. He used Cosine similarity function as a matching function for classification. Zelalem used TF*IDF to show the weight of features selected to represent news. According to Zelalem, classification error by experts is the major factor that reduces the performance of the classifier. Categories which contain high discriminating terms show better performance according to Zelalem. Zelalem's recommendations gives due attention on the development of standard Amharic preprocessing tools like spell checker, thesaurus and stop words so that researchers on Amharic text classification can focus on one aspect at a time.

The methods employed in Surafel's (2003) research work are Naïve Bayes and KNN. Surafel used Rainbow tool. And he said that Rainbow reads the datasets and writes/indexes to disk or archives a 'model' containing their statistics on documents like its category and the number of times words appear in a document. According to Surafel, the classification accuracy using Naïve Bayes and KNN decreases if the categories in the training data contain fewer documents. Surafel added that if the categories in the training data are not evenly distributed, classification affected negatively. Surafel indicated that KNN is poor for large dataset and there is a difficulty in determining the value of K. Amharic preprocessing tools development is one of the issues in Surafel's recommendation like Zelalem. Surafel also recommends other methods to work on automatic Amharic text classification.

Yohannes (2007) used SVM and Decision Tree methods. The weighting methods employed in Yohannes's research work is TF*IDF. Yohannes said that decision tree and SVM classifiers showed better accuracy for categories with large number of documents in the training set than fewer documents. He also noted that SVM and Decision Tree classifiers are good in accuracy at the expense of performance cost. Therefore, Yohannes recommends other classifiers with less processing cost and better accuracy. Like Zelalem and Surafel, Yohannes also recommended the development of standard Amharic preprocessing tool that aid text classification task.

Table 2.5 summarizes the results obtained by previous research works.

Name	Categories considered	Method used	Accuracy
Zelalem Sintayehu (2001)	3	Cosine Similarity	85.05%
Surafel Teklu (2003)	16	KNN	64.4%
		Naïve Bayes	78.48%
Yohannes Afework (2007)	15	LMT	79.72%
		LibSVM	81.15%

Table 2. 5: Previous Research Works on Automatic Amharic Text Classification

Surafel and Yohanes made experiments at increasing number of categories. And they found that accuracy decreases as the number of categories increase as depicted in Table 2.6.

Name	Category	Method	Accuracy
Surafel Teklu (2003)	3	KNN	89.61%
		Naïve Bayes	95.73%
	4	KNN	84.51%
		Naïve Bayes	93.86%
	7	KNN	75.27%
		Naïve Bayes	89.93%
	16	KNN	64.4%
		Naïve Bayes	78.48%
Yohannes Afework (2007)	5	LMT	93.45%
		LibSVM	95.21%
	10	LMT	89.98%
		LibSVM	91.36%
	15	LMT	79.72%
		LibSVM	81.15%

Table 2. 6: Previous Research Works Accuracy at Increasing Category Level

Two weighting schemes are used in this study, which are TF and TF*IDF. For the purpose of feature selection, Document Frequency (DF) thresholding is employed. This study considers one of the neural network algorithms called LVQ for Automatic classification of Amharic news. The method is tested with increasing number of categories and news items at various levels of epochs for both TF and TF*IDF weighting schemes.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

The tasks that are done for preprocessing of Amharic news includes tokenization, stop words and number removal, stemming, index term weight, dimension reduction and matrix generation. After the accomplishment of preprocesses, the classifier is constructed by Learning Vector Quantization (LVQ) learning method using MATLAB as a tool. Finally, the system is evaluated based on the results obtained using accuracy. The subsequent sections discuss the methods used for preprocessing the data to make it ready for classification task and the methods with regard to classifier building and evaluation.

3.2 Tokenization

Tokenization is the process by which tokens are identified as candidates to be used as features. Candidates in the sense that stop words and numbers are removed from tokens. And tokens which do not satisfy Document Frequency (DF) thresholding are not considered.

In this study, words are taken as tokens. All punctuation marks are converted to space and space is used as a word demarcation. Hence, if a sequence of characters is followed by space, that sequence is identified as a word.

During the translation of Amharic news using SERA, punctuation marks are also translated to their matches, accordingly. Table 3.1 shows some examples of translation of punctuation marks

from Amharic script to Latin script. Translation of all the characters including punctuation marks is attached in Appendix 5.

Amharic character	Translated to
፣	,
፤	;
፧	?
፡	(space)

Table 3. 1 Example of Amharic Punctuation Marks Translation

The translated punctuation marks are considered for processing: for replacing punctuation mark with space and using space as word demarcation. Algorithm 3.1 illustrates the process of tokenization.

Algorithm 3.1: Tokenization

```
do
        Read the contents of a file
        If a character is a punctuation mark then
            Replace it with space
        end if
        If space is found then
            Assign the word to variable
        end if
while end of file
```

3.3 Stop Word and Number Removal

Stop words are non content bearing words, which are less discriminating among documents since they appear in most of them. There are common stop words in Amharic which are used for

grammatical purposes like ነዉ, ነበር, ሆኖም, እና, ነገርግን, etc, which are non informative to identify documents. In addition to the common stop words, there are also news specific stop words like ገለፁ, ዘግበዋል, አስታውቀ, etc; their use is for elaboration and common to all news in accordance with the reporters of ENA. Because of the unavailability of standard stop list done by previous researchers, the researcher of this study is obligated to develop stop list.

Since stop words are highly frequent words, total frequency of terms aided by manual inspection, is the method employed in the process of identification of stop words. Stop list is prepared after identifying stop words; the list that contains, words which have to be removed from tokens generated during the tokenization process. The need of manual inspection is, because of frequently occurring keywords. For example, the word 'ቱሪዝም' ('tourism') is the most frequent word in the category 'Tourism development', which is crucial in discriminating the category. Hence, such words are not included in the stop list.

The purpose of identifying stop words is, to remove such words from the list of index terms. Index terms are believed to represent news or discriminate one news item from the others; whereas, stop words are not. Hence, using those words in the list of index terms is unimportant. That is why their exclusion from index term list is vital.

In most cases numbers are less discriminant among documents (Baeza-Yates and Ribeiro-Neto, 1999). In this study also, numbers are not considered as index terms. So, index terms list does not contain any number.

Algorithm 3.2 demonstrates removal of stop words and numbers from token list.

Algorithm 3.2: Stop Word and Number Removal

If token is in stop list then
    Remove from token list
end if
If token is number then
    Remove from token list
end if

3.4 Stemming

Stemming is changing varying words, due to grammatical reasons, to the root form of the word. Stemming is one of the preprocessing made on Amharic text news for this study. Stemmer that can remove common Amharic prefixes and suffices is developed.

Table 3.2 shows an example of the prefixes and suffices removed and an example under each affix; these are not the only affixes removed, list of all affixes considered for stemming purpose are presented in Appendix 7.

Type	Affix	Example	
		Word	Translated to
Prefix	ለ	ለጂማ	ጂማ
	ስለ	ስለጂማ	ጂማ
	በ	በጂማ	ጂማ
Suffix	ም	ጂማም	ጂማ
	ና	ጂማና	ጂማ
	ን	ጂማን	ጂማ

Table 3. 2: Affix Removed During Stemming

The stemmer developed for this study is based on affix removal algorithm using the concept by Nega and Willett (2002). In such case, rules are applied to find the stem of Amharic words. The rules to remove prefix or suffix from a given word may not hold true always. For instance, removing '**ዉ**' ('wu') from the word '**ሰዉ**' ('person') would give '**ሰ**' ('se'), which is meaningless; and removing '**በ**' ('be') from '**በልግ**' ('autumn') gives '**ልግ**' ('lg'),which does not represent the original meaning. Hence, two exception lists are prepared for which affix removal rules do not applied.

- List of words that prefix removal rule does not hold true and
- List of words from which suffix removal rule is not applied.

The stemmer developed takes words as an input and removes prefix of the word. After the prefix is removed, the word is again checked if it lasts with suffix in the suffix list, if so, the suffix is removed from the word. Table 3.3 shows an example.

	Example 1	**Example 2**	**Example 3**	**Example 4**
Input:	**ጂማ**	**ጂማን**	**የጂማ**	**የጂማን**
Prefix:	No	No	**የ**	**የ**
Output1:	**ጂማ**	**ጂማን**	**ጂማ**	**ጂማን**
Suffix:	No	**ን**	No	**ን**
Final output:	**ጂማ**	**ጂማ**	**ጂማ**	**ጂማ**

Table 3. 3: Example of Stemming

Algorithm 3.3 shows the procedure of finding stem of a word.

Algorithm 3.3: Stemming

```
L= token length-1
If token starts with prefix then
   If token not in exceptional list then
         token=token[prefix length: L]
         update token length
         L=token length-1
  end if
end if
If token ends with suffix then
   If token not in exceptional list then
         length=L-length of suffix
         token=token[0:length]
    end if
end if
```

3.5 Index Term Weight

All index terms are not equally important in representing and discriminating a document; it is thus, required to measure how important a term is with regard to representation and discrimination of a document. TF and TF*IDF are the weighing schemes used in this study. The procedure is shown in Algorithm 3.4.

Algorithm 3.4: Index Term Weight

```
for n to the number of news
        do
                If token=token[i]
                        TF=TF+1
                        i=i+1
                end if
        while end of file (news item)
        i=0
        do
                If token in news item then
                        dk =dk+1 //dk is the no. of documents containing the token
                end if
        while end of all news items
        Compute TF*IDF using Formula 2.3 of Section 2.4.2
        n=n+1
end for
```

3.6 Dimension Reduction

After identifying the number of tokens generated during tokenization, stop words and numbers removal and stemming are applied to reduce the number of tokens to be used as features. But still the dimension has to be reduced so that the most important attributes of each category is identified. The need to reduce the dimension is:

- Irrelevant features are removed which may affect performance badly.
- For convenient computational complexity.

In this study, Document Frequency (DF) thresholding is used to reduce the dimension of features generated. DF is the number of documents that contain a certain feature. The procedure for computing DF is shown in Algorithm 3.4 as d_k. DF thresholding is employed for each category to select features. The system is supported by manual observation; whether stop words which are not eliminated during stop word removal are mixed and if there are important features which do not satisfy the threshold.

3.7 Matrix Generation

After all the processing on Amharic news documents have been finalized, matrices have been generated to be fed into the learning algorithm. The matrices constitute terms and class as rows. The value of terms is the result of TF or TF*IDF weight methods. Finally, class attribute has been assigned the category label of each news item. The category name and category label of news are depicted in Table 3.4.

Category Name	Category Label
Bank and insurance	1
Tourism development	2
Mines and energy	3
ICT	4
Art	5
Educational coverage	6
Weather forecast	7
Religious assemblies and reports	8
Creativity work	9

Table 3. 4: News Categories and Their Corresponding Label

The experiment is carried out for two weighting schemes (TF and TF*IDF), hence Algorithm 3.5 indicates the procedure for generating matrix representation of news using TF and TF*IDF weight methods.

Algorithm 3.5: Generating Matrix

```
for i to size of index terms
        Write index term
        do
                If index term is in file then
                        Write TF or TF*IDF value
                Else
                        Write 0
                end if
        while end of file
        Write category label
        Return to new line
        i=i+1
end for
```

3.8 Classifier Building and Evaluation

Classifier building is the construction of the classifier using training dataset that can predict on test dataset. For the purpose of classifier construction, the neural network toolbox of MATLAB 7.0 has been used with Learning Vector Quantization (LVQ) algorithm.

3.8.1 MATLAB 7.0

The neural network tool box of MATLAB 7.0 supports command line or Graphical User Interface (GUI) to create, train, simulate (test) the network. For this study, GUI has been used to import data and create networks. And command window has been used to initialize epoch, create target, train networks and simulate (test) networks.

The preprocessed data has been prepared in CSV (Comma delimited) file format; that MATLAB can read. Actually, MATLAB can accept other file types too.

3.8.2 LVQ Algorithm for Amharic Text Classification

The procedure of using LVQ for Amharic text classification is:

- Create LVQ network
- Create target of the training dataset
- Train the network using training dataset
- Simulate (test) the network using test dataset

3.8.2.1 Creating LVQ Network

LVQ network is created using GUI of MATLAB. For the creation of the networks, the following parameters are specified.

Input Ranges: is the maximum and minimum value of each input data and it can be automatically generated from input matrix.

Number of Hidden Neurons: is the number of subclasses, which is always larger than the number of categories. For this study, in each experiment double to the number of categories is taken as the number of hidden neurons with the intension of representing each class (category) with two neurons after carrying out preliminary trials. There is no mechanism to assign neuron for each class separately in MATLAB; that is, no mechanism to assure equal distribution of neurons across classes. In this study, various epochs are experimented rater than varying number of hidden neurons by observing the result in three categories experiment using TF weight method

because good result (94.81% accuracy) is obtained by changing epoch at this experiment. Hence, the trend is followed for all experiments.

Output Class Percentage: is the percentage of each category with respect to the total number of training dataset.

Learning Rate: In LVQ, the learning rate states that the fraction of the distance between data point and the closest reference vector. It determines how far the reference vector is moved towards or away from the data point. The default value is 0.01, that is, the reference vector is moved 0.01 times the distance to the data point. In this study, the default learning rate (0.01) is used.

Learning Function: is the learning rule in LVQ. In MATLAB, there are two learning rules learnlv1 (LVQ 1) and learnlv2 (LVQ 2.1), Discussed in Section 2.5.7.2 of Chapter Two. Applying learnlv1 is the prerequisite to use learnlv2; hence, learnlv2 can only be used after applying learnlv1. The learning rule employed in this study is learnlv1.

3.8.2.2 Creating Target

Since LVQ is supervised algorithm, targets are very fundamental to the network; which is created using the command T= ind2vec (Tc), where Tc is the output (class) of the training dataset and T is the target created. Targets have rows equal to the number of classes and column equal to the number of training dataset as shown in Table 3.5.

1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0

Table 3. 5: Targets for Nine Classes

Table 3.5 shows the targets generated for nine classes, nine rows equal to number of classes and nine columns (reduced for viewing) equal to the number of training dataset. A one on a certain column indicates input data (on that column) belongs to that class; the reverse is true for a zero in a target column.

3.8.2.3 Training the Network

The construction of the classifier has been done using Learning Vector Quantization (LVQ) in a supervised manner. Hence, the algorithm demands training and test datasets which are preclassified by the experts. LVQ algorithm uses training dataset for classifier construction; and test dataset for the evaluation of the classifier constructed.

The training dataset constitutes 66.67% of the total data. The data is in CSV file format and imported using GUI of MATLAB.

Before training is started, the following parameters are set.

epoch: is the learning steps. In this study, various levels of epochs are experimented. Nine epoch levels are used for training: 100, 500, 1000, 1500, 2000, 2500, 3000, 3500 and 4000. epoch is set in MATLAB as 'network.trainParam.epochs= 'value';'.

100 is the default epoch for LVQ algorithm. epoch lower than 100 are not selected based on preliminary trial. Thus, experiment is made from the default epoch level up to 4000 increasing at interval of 500 (except the first). Interval of 500 is selected to see the impact of higher epoch levels because if smaller interval is chosen it takes long time to reach to 4000.

show: is the interval of epoch number to display the performance of the network while training. In this study, the default 'show' is used, which is 25.

goal: is the error that is intended to achieve during training. The default 'goal' is used in this study, which is 0.

After the initialization of training parameters, the next step is training the network created using the command, 'network=train (network, P, T);'. P is input patterns and T is the target for each input pattern.

The default training function is used, which is trainr, which trains a network with weight and bias learning rules with incremental updates after each presentation of an input. trainr is not directly called, it is called using 'train' by setting network.trainFcn property to 'trainr'.

During training, MATLAB uses different kinds of functions. The functions that LVQ uses are the following according to Demuth and Beale (2004).

MIDPOINT: is a weight initialization function that sets weight (row) vectors to the center of the input ranges (minimum+maximum/2).

NETSUM: is a net input function. Net input functions calculate a layer's (competitive or linear layer) net input by combining its weighted inputs and biases.

COMPET: is a transfer function. Transfer functions calculate a layer's (competitive layer) output from its net input. COMPET (N) takes one input argument, N and returns output vectors with 1 where each net input vector has its maximum value, and 0 elsewhere.

PURELIN: is a transfer function. Transfer functions calculate a layer's (linear layer) output from its net input to decide on the predicted class.

The performance function used is MSE (Mean Square Error), which is the default performance function. MSE is the average squared error between the network output and the target output.

The learning rule or the type of LVQ employed is LVQ1 (learnlv1), the algorithm works in the following manner according to Martín-Valdivia, Ureña-López and García-Vega (2007).

Step 1 Initialize the codebook vectors w_i and the learning rate α

Step 2 Randomly select an input vector x

Step 3 Find the winner unit closest to the input vector (the codebook vector w_c with the smallest Euclidean distance with regard to the input vector x):

$$D = \sqrt{\sum_{i=1}^{n} (x - w_c)^2}$$

Formula 3.1: Euclidean Distance

In Formula 3.1, n is the number of vectors in a codebooks, D is Euclidean distance between the input vectors, x and their representatives, w_c (codebook). The smallest D will be considered to select label for the input vector.

Step 4 Modify the weights of the winner unit:

- If w_c and x belong to the same class (the classification has been correct), then

$$w_c(t+1)=w_c(t)+\alpha(t)[x(t)-w_c(t)]$$

- If w_c and x belong to different classes (the classification has not been correct), then

$$w_c(t+1)=w_c(t)-\alpha(t)[x(t)-w_c(t)]$$

Step 5 Reduce the learning rate α

Step 6 Repeat from step 2 until a fixed number of iterations have been carried out.

3.8.2.4 Testing the Network

The test dataset is prepared like the training dataset and constitutes 33.33% of the total data. The test dataset is used to evaluate the performance of the classifier. In this study, accuracy is used for evaluation of Amharic text news classifier. The following commands are used for testing and evaluating the trained network.

'Y = sim (network, Test);' Test is the test dataset and Y is a matrix like target matrix of training data shown in Table 3.5.

'Yc = vec2ind(Y);' converts the target into value of 0 and 1 to actual category label. In this study, the category labels are 1,2,3,4,5,6,7,8,9.

Ec = Tcs-Yc; E_C is the number of misclassified news, Tcs is category of test dataset and Y_C is the category of the test result. Finally, the percentage of Ec is computed from the total test dataset.

CHAPTER FOUR

EXPERIMENT AND PERFORMANCE EVALUATION

4.1 Introduction

This study is concerned with automatic Amharic text news classification, which is a type of supervised learning. In this case, there are instances which can be used for training or learning and there are instances used for testing the classifier. For both learning and testing, preclassified data is required. Taking this into consideration, precategorized data have been collected and preprocessing is made on the data so as to make it ready for classifier learning. From the preprocessed data, classifier has been constructed and its effectiveness has been evaluated.

The subsequent sections discuss the architecture of this study, the process of making the data ready for classification task and finally, experimentations are conducted on automatic Amharic text news classification using TF and TF*IDF weighting schemes.

4.2 Architecture of Automatic Amharic Text News Classification

The architecture of automatic Amharic text news classifier has three basic components- preprocessing, classifier construction or building and classifier evaluation.

Preprocessing is the process of making the data ready for experiment. Since the collected news items are unstructured and stored in a database in a way suitable for the application of ENASoft, it is converted in such a way that is appropriate for this study. The final goal of this stage is to

convert news collection into a matrix in which index terms constitute rows and news items represent columns with TF or TF*IDF result as weight value of index terms.

After the data is made ready for experimentation, the training dataset has been fed into the learning algorithm so that classifier (model) has been constructed based on learning. Classifier construction is not the final stage; its effectiveness is tested by supplying test dataset in that its potential applicability is seen for researchers and developers alike. Figure 4.1 summarizes the architecture of this study.

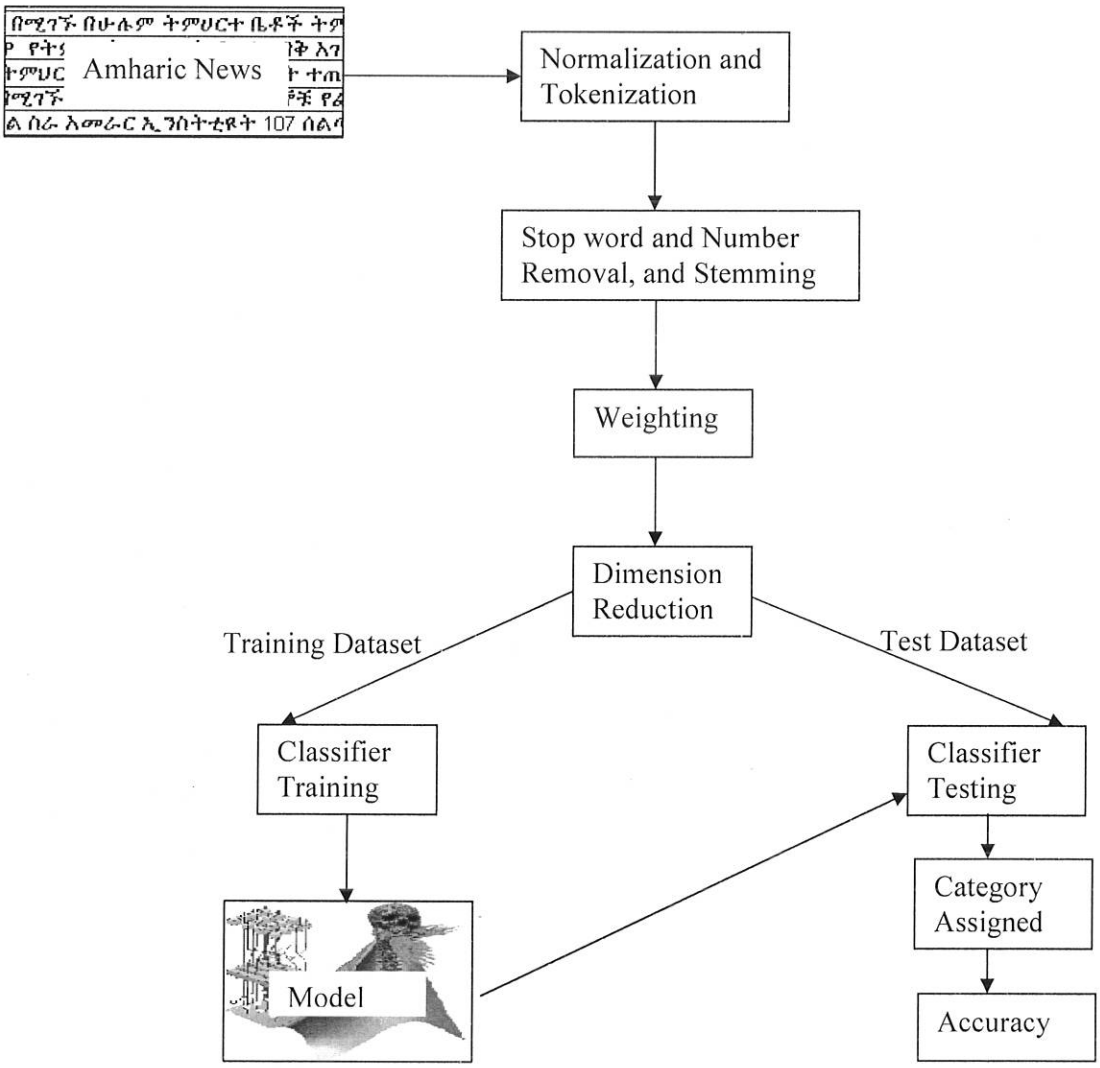

Figure 4. 1: Architecture of Automatic Amharic Text News Classification

In Figure 4.1, Amharic news items are used as an input to the system. Then preprocessing tasks like normalization (changing varying Amharic characters with similar sound to one common form, changing punctuation marks to space), tokenization, stop word and number removal, stemming, weighting terms and dimension reduction are done. After all these preprocesses, datasets are prepared in a matrix form, from which training and test datasets are prepared and used as training and testing purpose respectively. From the training dataset, model (classifier) is constructed. The model is tested using the test dataset. The testing outcome is the assignment of categories for news items that are not encountered during training. Finally, evaluation is made based on the test result using accuracy.

4.3 Data Source

The data source for this study is news of ENA. The news items are stored in SQL server using the format of Visual Ge'ez Unicode. 47, 037 news items are extracted from the server of ENA. The majority (83.6%) of news are from the years 2007 and 2008. Only 16.24%, from the total data collected, are from 2003, 2004, 2005, 2006, 2007 and 2008 years. According to the interview made with ICT coordinator of the agency, the four years news items are not complete due to damage. Table 4.1 shows the number of news items in each year.

No.	Year	Number of news	Percentage
1	2003	39	0.083%
2	2004	98	0.208%
3	2005	2	0.004%
4	2006	7,500	15.945%
5	2007	20,359	43.283%
6	2008	19,039	40.477%
Total		47,037	100%

Table 4. 1: Data Collected From ENA

The data are classified in accordance with major classification and sub classification. Totally, there are 116 classes including the major and sub classes. Among these 13 categories are major categories, and the rest (103) are sub classes under the major classes. The major and sub categories are attached as an Appendix 6.

For the purpose of this study, nine categories are taken into consideration. The nine categories have been selected based on random sampling, which are Bank and insurance, Tourism development, Mines and energy, ICT, Art, Educational coverage, Weather forecast, Religious assemblies and reports, and Creativity work. The number of news items in each category and the total number of news items are shown in Table 4.2.

No.	Category	News No.
1	Bank and insurance	339
2	Tourism development	295
3	Mines and energy	269
4	ICT	205
5	Art	201
6	Educational coverage	147
7	Weather forecast	135
8	Religious assemblies and reports	113
9	Creativity work	58
Total		1, 762

Table 4. 2: Number of News Items for the Nine Categories before Removing Irrelevant Ones

4.4 Removed News

Even though the total number of news items for the nine categories is 1, 762, all of them have not been considered due to various reasons. Which means that there are news removed from the data

and part of news are considered for the construction of the classifier. Some news items disturb the objectivity of the study others are not relevant for the purpose meant for. The removal is made by exporting news from SQL server to Microsoft Excel because Microsoft Excel is found to be very convenient for the researcher and suitable to view Amharic news. 224 news items are removed due to various reasons. Table 4.3 shows the number of news items removed for the various reasons.

Reason for Removal	No. of News
Null Values	53
Misclassified News	36
Meaningless Characters	78
Redundancy Problem	18
Correction of News	26
Language Problem	13
Total	224

Table 4. 3: Number of Removed News Items

The subsequent parts describe the reasons for the removal of news items.

4.4.1 Null Values

The study considers three important parts of news for the purpose of selecting relevant features. It has been believed that news articles can be represented by headline, keyword and slug. That is why the three parts of the news are taken into account for this study. Hence, records are removed if they lack headline, keyword, or slug. Table 4.4 is an example of news with no value in the keyword and slug attribute.

Head Line	Keyword	Slug
የኢትዮጵያን መልካም ገጽታ ለዓለም የሚያስተዋውቅ አንድ የጋዜጠኞች ቡድን አዲስ አበባ ገባ		

Table 4. 4: News Item with Null Value of Keyword and Slug

4.4.2 Misclassified News

Due to error made during data entry in ENA, some news items are entered in a wrong category that they do not belong to. Misclassified news items are identified with the help of reporters of ENA. For example, the following news item with the meaning 'According to consumers and merchants the price of fruit have no change since last week', is categorized as 'Bank and Insurance' while it belongs to the category 'Trade'.

"የአትክልት ምግቦች ዋጋ ላይ ካለፈው ሳምንት ብዙም ለውጥ አለማሳየቱን ሸማቾችና ነጋዴዎች ገለፁ"

4.4.3 Meaningless Characters

Records with erroneously entered characters are removed from the data. If a news item contains characters that do not convey any message, it has been reduced from the data. Actually, care is taken to do that; reporters are asked the meaning associated with such characters but they replied that such scenario happens due to error. Illustration is presented in Table 4.5.

Head Line	Keyword	Slug
??	??	??
ሰሰሰሰፈፈቨቨ	ፈ	ገገ

Table 4. 5: Meaningless Characters Found in News Collection

4.4.4 Redundancy Problem

News items which contain redundant words, phrases or sentences are removed since they disturb the objectivity of news representation using index terms. The following news, meaning: '739 students are graduated', belongs to the category 'Educational Coverage', which is repeated four times after the symbol :: (the final one is not complete).

ኮሌጁ በመደበኛው የትምህርት ፕሮግራም ያሰለጠናቸውን 739 ተማሪዎች ለመጀመሪያ ጊዜ በዲግሬ ዛሬ አስመረቀ:: ኮሌጁ በመደበኛው የትምህርት ፕሮግራም ያሰለጠናቸውን 739 ተማሪዎች ለመጀመሪያ ጊዜ በዲግሬ ዛሬ አስመረቀ:: ኮሌጁ በመደበኛው የትምህርት ፕሮግራም ያሰለጠናቸውን 739 ተማሪዎች ለመጀመሪያ ጊዜ በዲግሬ ዛሬ አስመረቀ:: ኮሌጁ በመደበኛው የትምህርት ፕሮግራም ያሰለጠናቸውን 739 ተማሪዎች

4.4.5 Correction of News

Some news items contain correction of the original news. For such kinds of news items, 'correction' ('ማረሚያ') is written on the headline. Those news items have been removed from the data since they do not contain the sense of the news.

4.4.6 Language Problem

Some Head Lines, Keywords or Slugs are written using English. All these news items are removed from the data. Figure 4.2 illustrates snapshoot of the table in the news database that is written using English, the fifth item.

ተሰፋዬ እበበ ተጓዥ የትያትር ማደራጃ የተመሠረ	የተሰፋዬ እበበ ትያትር ማደራጃ	ኪነጥበብ
ኢትዮጵያ በሳይንሱ የሰው ልጅ መገኛ ከመሆኗ ባ	የሳቅ ንጉሥ	የሳቅ ንጉሥ
የዘፈንና የፊልም ሕገወጥ ቅጅዎች ሊቃጠሉ ነው	የአዲሺዮዋል እሳታሚዎች ማህበር	ኪነጥበብ
ከ 10 ሚሊዮን ብር በላይ ግምት ያላቸው ህገወጥ	የቅጅ መብት	የቅጅ መብት
Minassie	ደደ	ደደደ
የቅጅ መብት ህግን በማጠናከር ሊከሰት የሚችለው	የቅጅ መብት ህግ	የቅጅ መብት ህግ
የዓለም ወጣቶች ቀንን ምክንያት በማድረግ የሥ	የሥነ ጽሁፍ ምሽት	ኪነጥበብ
የማስታወቂያ ሚኒስቴር እገርአቀፍ የፎቶግራፍ ኤግዚቪሽን በሠመራ ከተማ ተከፈተ		

Figure 4. 2: News Item Written Using English Character

After removing news with the fore mentioned problems a total of 1, 538 news items remain. Table 4.6 shows the categories together with the number of news considered.

No.	Category	News No.
1	Bank and insurance	297
2	Tourism development	253
3	Mines and energy	251
4	ICT	167
5	Art	152
6	Educational coverage	138
7	Weather forecast	132
8	Religious assemblies and reports	103
9	Creativity work	45
Total		1, 538

Table 4. 6 Number of News after Removal of Irrelevant News

News items which contain none of attributes (features) are removed. That is, after news matrix generation, any news item that holds none of the features is removed from the datasets. Table 4.6 does not incorporate such removal. Since the number of removal in such case varies based on the number of categories in each experiment, it is discussed in Section 4.9.2 under three, six and nine categories experiment using TF weight method. For example, from the category 'Art', 4 news items are removed during three and six categories experiment but 3 news items are removed during the nine categories experiment.

4.5 Translating and Exporting Amharic News

Amharic characters are translated into Latin script for processing purpose using SERA, which is a system of converting Amharic characters into ASCII. The character table of SERA in Appendix 5 is modified according to information in Appendix 8. The modification is required in order to change Amharic characters with different symbols but similar sound to one common form. For example, in SERA character table '**ሀ**' is mapped to 'he' and '**ሃ**' is mapped to 'ha' but both '**ሀ**' and '**ሃ**' have similar sound, hence in this study, both are translated to 'he'.

The need to handle various forms of characters with the same sound is, to treat the same word with such different characters as one word. In Amharic, there are various characters with the same sound as discussed in Chapter Two of Section 2.6.4.1. And there is no rule where to use those varying characters. In the news collected, inconsistent usage of such characters is very common. If those characters are not replaced with one common character, the same word is treated differently. As a result, we obtain more than one index term for the same word.

The word '**ትምህርት**' ('education'), for instance, is found in the data written in three forms as '**ትምህርት**', '**ትምሕርት**' and '**ትምኅርት**'; which means, there are three index terms for the word '**ትምህርት**' ('education') even if all the three forms have the same meaning, 'education'. Such kinds of things increase computational complexity (decrease efficiency) and decrease effectiveness (accuracy). Therefore, all words with varying characters but having the same sound are converted to one common form.

The collected news items are stored in SQL server. All the news are exported and saved as text file (txt file format) in order to process the news. From the SQL server, three attributes of news

are taken into consideration. It is believed that the three attributes, headline, keywords and slug, represent the contents of news.

Filenames are generated automatically for each news item during the export. All files are stored in a folder called 'NEWS'. The nine categories are identified by numbers (file names) in a certain range given for each category as shown in Table 4.7.

No.	Category	Filename Range
1	Bank and insurance	0-296
2	Tourism development	297-549
3	Mines and energy	550-800
4	ICT	801-967
5	Art	968-1, 119
6	Educational coverage	1, 120-1, 257
7	Weather forecast	1, 258-1, 389
8	Religious assemblies and reports	1, 390-1, 492
9	Creativity work	1, 493-1, 537

Table 4. 7: Filename Range for each Category

4.6 Stop Word and Number Removal, and Stemming Experiments

Preprocessing done on the data is to generate matrix of terms and their TF and TF*IDF weight values. But all terms are not used as representatives of news due to high dimensionality. Hence forth, after tokenization, words which have less contribution in representing and discriminating news have been eliminated. Such terms are stop words and numbers. As a method of enhancing representation, stemming is also used to bring words with similar concept but varying characters due to grammatical usage of Amharic; that would otherwise be treated differently if stemming is not considered. Table 4.8 shows the results with regard to the number of features, reduction of features in number and percentage (%) after each preprocessing has been made and after all the preprocessing tasks are made together.

Category Name	Tok. N	Number Removal			Stop Word Removal			Stemming			All Preprocessing		
		FN	RN	R%	FN	RN	R%	FN	RN	R%	FN	RN	R %
Bank and insurance	1007	967	40	3.97	974	33	3.28	841	166	16.48	613	394	39.13
Tourism development	1176	1162	14	1.19	1137	39	3.32	926	250	21.26	834	342	29.08
Mines and energy	1303	1267	36	2.76	1264	39	2.99	1026	277	21.26	905	398	30.54
ICT	856	840	16	1.87	827	29	3.39	685	171	19.98	622	234	27.34
Art	743	740	3	0.4	723	20	2.69	627	116	15.61	587	156	21
Educational coverage	794	777	17	2.14	752	42	5.29	628	166	20.91	541	253	31.86
Weather forecast	644	642	2	0.31	619	25	3.88	494	150	23.29	447	197	30.59
Religious assemblies and reports	717	712	5	0.7	684	33	4.6	608	109	15.2	557	160	22.32
Creativity work	340	336	4	1.18	328	12	3.53	296	44	12.94	273	67	19.71
Total	5129	5032	97	1.89	3662	103	2.01	5026	1467	28.6	3258	1881	36.48

Table 4. 8: Stop Word and Number Removal, and Stemming Experiments Results

In Table 4.8, Tok.N is the number of tokens generated by tokenization process, FN is the number of features after each preprocessing, RN is the number of reduction of features from the total tokens generated during tokenization, R% is the reduction of features in percentage.

The 'Total' of the column 'FN', in Table 4.8, is not the sum of all features. There are common features which occur among two or more categories. That is why, the 'Total' of the column 'FN' reduced from the sum of features in each category.

From Table 4.8, it can be observed that stemming reduces features better than other preprocessing methods, which is 28.6% for all categories. Stop word removal reduces features by 2.01% and 1.89% features are reduced from the total features as a result of number removal. When all the preprocessing is done together, 36.48% of features are reduced for all categories.

The amount of features reduced is considerable but greater number of features can be reduced if stop word removal system and stemming systems are complete. If the stop word removal system uses stop word list that incorporates all stop words better reduction can be achieved according to this study. Using the stemmer developed, good result is obtained; if standard stemmer that takes all affixes into consideration is used, again the result of this study indicates that more reduction can be accomplished. Amharic is morphologically reach that needs the development of sophisticated stemmer; for example, the word 'ebyate*' ('አብያተ') is not stemmed to 'bEt' ('ቤት') in this study. If such cases are incorporated, stemming method can brought good reduction in the number of features according to the result of this study.

4.7 Dimensionality Reduction of Features

The number of features is too many (3258) even after undergoing different preprocessing methods. Hence, dimensionality reduction method using Document Frequency (DF) thresholding is carried out to reduce features further. The reduction is required to represent news with the most important features and hence, reduce computational complexity.

DF value of a term is the number of documents which contain that term. Hence, DF is experimented for each category to identify features which can well represent and discriminate each category. In doing so, a semi-automatic technique is employed to identify the most representative terms for each category. This enables to check important terms which do not satisfy the DF threshold and also irrelevant terms which satisfy the DF threshold. Consultation with ENA reporters is fundamental to decide whether a word is keyword or not for that category. For example, the term 'ityoPya' ('ኢትዮጵያ') occurred in 46 news items in the category 'Bank

* Appendix 5 can be referred for the translation of Amharic script to Latin script

and insurance', which satisfy DF threshold. But the term is not keyword for the category; as a result, the term is excluded in the feature list of the category. On the other hand, the term 'inxurans' ('ኢንሹራንስ') has DF of 8, which is below the threshold, but it is the most important keyword in the category 'Bank and insurance'; hence, it is included in the feature list to represent the category. Table 4.9 shows the DF threshold used and the number of features for each category.

Category	DF Threshold	Features No.
Bank and insurance	40	7
Tourism development	24	8
Mines and energy	40	8
ICT	20	9
Art	17	10
Educational coverage	20	15
Weather forecast	30	6
Religious assemblies and reports	12	13
Creativity work	10	4
Total		80

Table 4. 9: DF Threshold and Number of Features for Each Category

For example, the category 'Creativity work' has the features 'feTera' ('ፈጠራ'), 'mrmr' ('ምርምር'), 'elmroewi' ('አእምሮአዊ') and 'patent' ('ፓተንት').

The experiment for this study is carried out on three, six and nine categories; thus, the number of features is not the sum of features for the three (ICT, Art and Educational coverage), six (ICT, Art, Educational coverage, Weather forecast, Religious assemblies and reports and Creativity work) and nine (Bank and insurance, Tourism development, Mines and energy, ICT, Art,

Educational coverage, Weather forecast, Religious assemblies and reports and Creativity work) categories as there are common features which occur across categories. Table 4.10 depicts the feature size for the three, six and nine categories.

Category Number	Feature Number	Feature Sum	Difference
3	32	34	2
6	47	57	10
9	69	80	11

Table 4. 10: Number of Features in Three, Six and Nine Categories Experiments

In Table 4.10, 'Difference' indicates the number of common features. 'Feature Sum' shows, the number of features when number of features in three, six and nine categories are summed according to Table 4.9. 'Feature Number' indicates the actual feature size in the respective experiments; the reduction is due to the common features found in more than one category.

From Table 4.10, we can observe that only 2.12% of the total tokens identified after all preprocessing, are used as features by the application of dimension reduction step.

4.8 Matrix

The input to the learning algorithm is a matrix generated with the value of term weights, TF and TF*IDF. Table 4.11 and Table 4.12 show an example of matrix generated for the nine categories experiment using TF and TF*IDF weight methods respectively; the rows and columns are reduced for viewing purpose.

The algorithm requires only the weight values for training input dataset. That is, during training (learning), terms and class labels are not used. And the class row (last row) is used separately as an output of the input patterns, from which targets are created.

ቱሪዝም 'turizm'	0	0	0	0	0	0	0
ትምህርት 'tmhrt'	0	0	0	0	0	0	0
ባንክ 'bank'	0	1	1	0	3	0	0
ምንዛሪ 'mnzari'	2	1	0	1	1	2	2
ማእድን 'maIdn'	0	0	0	0	0	0	0
Class	1	1	1	1	1	1	1

Table 4. 11: Matrix Using TF Weight Method

In Table 4.11 and Table 4.12, zero indicates that the feature does not occur in that category; otherwise its weight value is used to show its importance.

ቱሪዝም 'turizm'	0	0	0	0	0	0	0
ትምህርት 'tmhrt'	0	0	0	0	0	0	0
ባንክ 'bank'	0	3.46	3.46	0	10.38	0	0
ምንዛሪ 'mnzari'	6.64	3.32	0	3.32	3.32	6.64	6.64
ማእድን 'maIdn'	0	0	0	0	0	0	0
Class	1	1	1	1	1	1	1

Table 4. 12: Matrix Using TF*IDF Weight Method

4.9 Classification Experiment on Amharic Text News

The experiment here is using the matrix generated for both TF and TF*IDF weighting schemes. Using the two weight methods, experiments are carried out for three, six and nine categories respectively.

4.9.1 Experimental Plan

There are 54 experiments carried out totally. For each experiment, nine levels of epochs are used and experimented, which are 100, 500, 1000, 1500, 2000, 2500, 3000, 3500 and 4000 epochs.

Since the objective of the study is to see the trend in performance with the increase in the number of categories and news, the experiments done are on three, six and nine categories with increasing number of categories and news. To select the three groups, categories are arranged in decreasing number of news as can be referred from Table 4.6. And the categories are grouped into three groups with each group having three category members according to Table 4.13.

Group1		Group2		Group3	
Category	**News No.**	**Category**	**News No.**	**Category**	**News No.**
Bank and insurance	297	ICT	167	Weather forecast	132
Tourism development	253	Art	152	Religious assemblies & report	103
Mines and energy	251	Educational coverage	138	Creativity work	45
Total	801	**Total**	457	**Total**	280

Table 4. 13: Categories in Group to Plan the Experiment

Comparing the three groups, 801 is the largest number of news items. One of the aims of this study is to see performance effect at increasing number of news items. Hence, the first group is not taken for the three categories experiment since taking largest number from the beginning is

not plausible. The next group contains medium number of news items; hence, it is sensible to select for the three categories experiment.

For the six categories experiment, if the group with smaller number of news added, the nine categories experiment will have largest enough number of news items that complements with the objective. Hence, Group2 and Group3 are taken for the six categories experiment and Group 1 is left to be added in the nine categories experiment.

Finally, the nine categories experiment contains all the categories or all the Groups. As a result, the experiment at the three levels with increasing number of categories and news items are:

1. Three categories: ICT, Art and Educational coverage.
2. Six categories: ICT, Art, Educational coverage Weather forecast, Religious assemblies and reports, and Creativity work.
3. Nine categories: Bank and insurance, Tourism development, Mines and energy, ICT, Art, Educational coverage, Weather forecast, Religious assemblies and reports, and Creativity work.

Table 4.14 summarizes the training and test datasets for the three, six and nine categories experiments.

Category Level	Training Set	Test Set	Total
Three	271	135	406
Six	454	224	678
Nine	975	488	1463

Table 4. 14: Training and Test Sets at Three, Six and Nine Categories Experiment

4.9.2 Classification Using TF Weighting Scheme

The TF weighting scheme is based on the occurrence of the term in that news item. In a term (rows) by news (columns) matrix, TF represent the cell value as depicted in Table 4.11. The experiments on three, six and nine categories are discussed in Section 4.9.2.1 to 4.9.2.3.

4.9.2.1 Classification for Three Categories

The experiment using three categories has been made considering the categories 'ICT, Art and Educational coverage'. The datasets for training and testing are summarized in Table 4.15.

Class Name	Class Label	Feature No.	News No.	Training set (66.67%)	Test set (33.33%)	Output %
ICT	1	9	122	81	41	0. 299
Art	2	10	148	99	49	0. 365
Educational coverage	3	15	136	91	45	0.336
Total		32	406	271	135	1

Table 4. 15: Datasets Summary for Three Categories

In Table 4.15, Output percentage (%) is one of the parameter required to create LVQ network, which is computed by dividing training set percentage by 100.

Thirty two features have been identified for the three categories as presented in Table 4.10 of Section 4.7. Fifty one news items which do not contain any of the features in the matrix generated have been removed. The reason for removing those records is that, if news item does not contain any of the terms in the matrix; it obviously belong to other category. Because, manual inspection of fifty news items indicates that any news item contains at least one of the features.

So, it is most likely such news items do not belong to the categories considered; which may not be identified during misclassification identification.

The networks have been created and trained at various levels of epochs. Evaluation has been made and accuracy has been recorded for each epoch levels as presented in Table 4.16.

Epoch	100	500	1000	1500	2000	2500	3000	3500	4000
Accuracy	69.63%	68.89%	68.89%	69.63%	94.81%	68.89%	69.63%	68.89%	68.89%

Table 4. 16: Accuracy for Three Categories Using TF Weighting Scheme at Various epoch Levels

Epochs with similar accuracy are discussed together because in these epochs, it has been seen that the misclassifications results and the wrong category assigned are the same.

100, 1500 and 3000 epochs: The same accuracy of 69.63% is obtained. In these tests, it has been observed that the misclassified news items belong to 'ICT' category. When analyzing, 'ICT' category have least number of training data and features as compared to the categories 'Art and Educational coverage'. All the news items of the category of 'Art and Educational coverage' are classified correctly, which have larger representative in the training data.

500, 1000, 2500, 3500 and 4000 epochs: Accuracy of 68.89% is registered. These experiments are similar to the above experiments. The only difference is, one news item that belongs to the category 'Educational coverage' is misclassified to 'Art' category (which contains the highest number of training data). The news item contains the word 'mrqat' ('ምርቃት') and 'temari' ('ተማሪ') that occurs in small number from 91 training dataset (6 and 11 times respectively).

2000 epoch: 94.81% of accuracy is obtained, which is the best result. In this context, most news items are categorized correctly except seven news items which are misclassified to the category

'Art'. As mentioned, the category 'Art' contains the highest number of training instances in the three categories experiment. Misclassified news items and their feature with TF value at 2000 epoch is shown in Table 4.17.

No.	Feature	TF	Class in Test Data	Misclassified To
1	Eyer (አየር)	1	ICT	Art
2	mrmr (ምርምር)	1	ICT	Art
3	Temari (ተማሪ)	3	Educational Coverage	Art
	tmhrt (ትምህርት)	1	Educational Coverage	Art
4	Mrqat (ምርቃት)	2	Educational Coverage	Art
	Temari (ተማሪ)	1	Educational Coverage	Art
5	Temari (ተማሪ)	3	Educational Coverage	Art
	tmhrt (ትምህርት)	1	Educational Coverage	Art
6	tmhrt (ትምህርት)	1	Educational Coverage	Art
7	mrqat (ምርቃት)	2	Educational Coverage	Art
	Temari (ተማሪ)	1	Educational Coverage	Art
	tmhrt (ትምህርት)	1	Educational Coverage	Art

Table 4. 17: Misclassified News for 3 Classes Experiment Using TF Weight Method at 2000 epoch

In the test dataset, the word 'eyer' ('አየር') is found in the news item that belong to the category 'ICT'; but, the word 'eyer' ('አየር') is the keyword for the category 'Weather forecast'. During testing, the news item is classified to the category 'Art', which has the highest number of news, wrongly. The news item in actual context belongs to the category 'Weather forecast'; but it is misplaced in the category 'ICT' due to error. The news item is not identified during misclassification identification at the start.

The second misclassified news item contains the word 'mrmr'('ምርምር'), which is keyword for the category 'Creativity work'. It may be due to this ambiguity that the news item is misclassified to 'Art' category with the larger number of news items.

It is also observed that, news item that contains only the features 'mrqat' ('ምርቃት') and 'temari' ('ተማሪ') in the category 'Educational coverage' are classified wrongly to 'Art' category. The reason may be the features occur in small number of news. Out of 91 training dataset, the word 'mrqat' ('ምርቃት') occurs in 6 news items and the word 'temari' ('ተማሪ') occurs in 11 news items.

The other misclassified news item contains the word 'tmhrt' ('ትምህርት') with TF of 1. The word 'tmhrt' ('ትምህርት') is the most important keyword for the category 'Educational coverage' but news items which contain the feature with frequency of 1 are misclassified to 'Art' category. The feature occurs more than one in all the news items except the misclassified news; this may be the reason of misclassification.

In all the experiments for three categories, when there is misclassification, the wrong category assigned has the highest number of training data. And most misclassifications are due to all news items misclassification in certain categories except 2000 epoch experiment. This occurs when the categories are not well represented by enough number of subclasses (hidden neurons) according to Umer and Khiya (2007).

4.9.2.2 Classification for Six Categories

The six categories considered are ICT, Art, Educational coverage, Weather forecast, Religious assemblies and reports, and Creativity work. From the six categories, 'ICT, Art and Educational coverage' are used in the first experiment. The rest three categories are added for increasing category and news number as the aim of the study is to see the effect on performance when

category and news number are increased. Table 4.18 provides summary statistics of datasets used for training and testing.

Class Name	Class Label	Feature No.	News No.	Training set (66.67%)	Test set (33.33%)	Output %
ICT	1	9	122	84	38	0.185
Art	2	10	148	98	50	0.216
Educational Coverage	3	15	136	90	46	0.198
Weather forecast	4	6	131	88	43	0.194
Religious assemblies and reports	5	13	96	64	32	0.141
Creativity work	6	4	45	30	15	0.066
Total		47	678	454	224	1

Table 4. 18: Datasets Summary for Six Categories

The six categories have forty seven attributes totally. Fifty nine news items which do not contain any of the features have been removed. The results of testing for the six categories are depicted in Table 4.19.

Epoch	100	500	1000	1500	2000	2500	3000	3500	4000
Accuracy	61.61%	61.61%	61.61%	61.61%	61.61%	61.61%	61.61%	61.61%	61.61%

Table 4. 19: Accuracy for Six Categories Using TF Weighting Scheme at Various epoch Levels

In these experiments, similar result, 61.61% of accuracy, is registered for all levels of epochs. Not only the accuracy, but also the misclassification of instances is the same for all the epoch levels. Like the previous results, all the news with small number of news items in the training dataset (ICT, Religious assemblies and reports, and Creativity work) are misclassified into wrong categories with more number of news items and features.

All the news items of 'Educational Coverage and Weather forecast' categories are classified correctly. From the category 'Art', only one instance is misclassified to the category 'Weather forecast'. Observing the test data indicates, the news item contains the words 'ertist' ('አርቲስት') and 'eyer' ('አየር'). The feature 'eyer' ('አየር') is the keyword for the category 'Weather forecast'. This may be the reason that the news item misclassified to the category 'Weather forecast'.

The three categories, 'Art, Educational Coverage and Weather forecast' have more number of news in the training data or more number of features than other categories, which is the reason for the good classification result obtained in the three classes. Whereas, the misclassification of all news items in the categories 'ICT, Religious assemblies and reports, and Creativity work' dictates the lack of enough number of subclasses which can represent each category.

4.9.2.3 Classification for Nine Categories

The final experiments using TF weight method have been made for all the nine categories 'Bank and insurance, Tourism development, Mines and energy, ICT, Art, Educational coverage, Weather forecast, Religious assemblies and reports, and Creativity work'. The statistics of datasets is shown in Table 4.20.

Sixty nine attributes are there for the nine categories. Seventy five news items which contain neither of the attributes have been removed from the total data.

Class Name	Class Label	Feature No.	News No.	Training set (66.67%)	Test set (33.33%)	Output %
Bank and insurance	1	7	294	196	98	0.201
Tourism development	2	8	247	165	82	0.169
Mines and energy	3	8	242	161	81	0.165
ICT	4	9	122	81	41	0.083
Art	5	10	149	99	50	0.102
Educational Coverage	6	15	136	91	45	0.093
Weather forecast	7	6	131	87	44	0.089
Religious assemblies and reports	8	13	97	65	32	0.067
Creativity work	9	4	45	30	15	0.031
Total		69	1463	975	488	1

Table 4. 20: Datasets Summary for Nine Categories

The results of the experiment at the various epoch levels are illustrated in Table 4.21.

Epoch	100	500	1000	1500	2000	2500	3000	3500	4000
Accuracy	62.09%	62.3%	61.48%	70.08%	62.3%	61.07%	61.48%	61.89%	61.68%

Table 4. 21: Accuracy for Nine Categories Using TF Weighting Scheme at Various epoch Levels

Except the 1500 epoch experiment, all the news items which belong to the categories 'ICT, Art, Weather forecast, Religious assemblies and reports, and Creativity work' are not classified correctly. For the 1500 epoch experiment, all the news items of the category 'Weather forecast' are classified correctly. The categories 'ICT, Art, Weather forecast, Religious assemblies and reports, and Creativity work' those their news items are not assigned to correct classes have the least number of news data as compared to other categories considered. This is similar to what has been observed on the three and six categories experiment. The difference here is, the category 'Educational coverage' with less number of news (131 news items) than the misclassified category 'Art' (149 news items) is among the category in which its news items are correctly classified. This may be due to the largest number of features (15) in the category 'Educational coverage' as compared to all the categories.

Most of the news items which belong to categories with larger number of news items are classified correctly, only some news items are misclassified from these categories. These misclassifications at the various levels of epochs are analyzed and discussed below by observing the data.

100 epoch: 62.09% of accuracy is registered. At this experiment, all the news items of the category 'Tourism development' are classified correctly. One news item is misclassified to 'Tourism development' in the category 'Bank and insurance'; the news item contains the features 'mereja' ('መረጃ') and 'faynans' ('ፋይናንስ'). The word 'mereja' ('መረጃ') is keyword for the category 'ICT'; this ambiguity may be the reason of misclassification to category with the higher number of features, 'Tourism development'. And also one news item in the category 'Mines and energy' is classified to 'Bank and insurance' wrongly. When the data is observed, the news item contains the features 'heyl' ('ሀይል') with TF of 1, 'bdr' ('ብድር') with TF of 3, 'ElEktrik' ('ኤሌክትሪክ') with TF of 1 and 'bank' ('ባንክ') with TF of 1. The words 'bdr' ('ብድር') and 'bank' ('ባንክ') are the keywords to the category 'Bank and insurance' with higher TF sum value; which may be the cause of misclassification.

One news item from the category 'Educational coverage' is misclassified to the category 'Tourism development', which is the second in the number of news it contains (but exceeds the first in the number of features). The news item contain 'mrqat' ('ምርቃት') and 'temari' ('ተማሪ'), which are words occur in few (6 and 11 respectively) number of news as described earlier.

500 and 2000 epochs: at these epoch levels, the same accuracy of 62.3% is resulted. The results of the experiments are similar like the 100 epoch experiment. The difference is, the misclassified

news item in the category 'Bank and insurance' is correctly classified; in this case, the consideration of the word 'faynans' ('ፋይናንስ') may be the reason for the correct classification.

1000 and 3000 epochs: 61.48% of accuracy is registered. Like 100 epoch experiment, the same result is encountered except the misclassification of additional three news items. These news items belong to the category 'Bank and insurance' but wrongly classified to 'Tourism development' category. When the news items are observed, all the news items contain the word 'faynans' ('ፋይናንስ') only. The presence of only one feature may be the cause of misclassification.

2500 epoch: at this experiment, 61.07% of accuracy is registered. The result is similar to the experiments at 1000 and 3000 epochs but two additional misclassifications. These news items with the word 'qrs' ('ቅርስ'), belong to the category 'Tourism Development' but misclassified to the category 'Mines and energy'. The news items contain only one feature 'qrs' ('ቅርስ'); hence, it may be the motive to be misclassified into the wrong category. The other difference with the 1000 and 3000 epochs experiments is, the three news with the word 'faynans' ('ፋይናንስ') in the category 'Bank and insurance' are misclassified to the category 'Tourism development' in the 1000 and 3000 epochs experiment; but, in this case, they all are misclassified as 'Mines and energy' category, which is the third largest in the number of news items that it contains.

3500 epoch: 61.89% of accuracy is resulted at this experiment. Two news items with the word 'qrs' ('ቅርስ'), like 2500 epoch experiment, are misclassified; but, in this experiment they are misclassified to the category 'Bank and insurance', which have the largest number of news items. As of experiment 100 epoch, the news item with the features 'heyl' ('ሀይል') with TF of 1, 'bdr'

('ብድር') with TF of 3, 'ElEktrik' ('ኤሌክትሪክ') with TF of 1 and 'bank' ('ባንክ') with TF of 1, is misclassified to 'Bank and insurance' while it belongs to the category 'Mines and energy'. And similar to the 100 epoch experiment, one news item in the category 'Educational coverage' misclassified to the category 'Tourism development'.

4000 epoch: 61.68% of accuracy is achieved for this experiment. One additional news item is misclassified compared to the above experiment, 3500 epoch. This news item contains the features 'mereja' ('መረጃ') and 'faynans' ('ፋይናንስ') belonging to the category 'Bank and insurance' but misclassified to 'Mines and energy' category, which has the third highest number of news items. The other misclassifications are similar to the above experiment except the news with the word 'qrs" ('ቅርስ') are misclassified to the category 'Mines and energy' like the 2500 epoch experiment.

1500 epoch: In the nine categories experiment, the best accuracy, 70.08%, is registered at this epoch. In this context, the improvement in accuracy is due to the correct classification of all the news items of 'Weather forecast' category. Except this difference, the same thing is resulted in this experiment like the experiment at 2500 epoch.

Like three and six categories experiments, wrong categories assigned have larger number of news items or larger number of features. And most misclassifications are due to all news items misclassification in certain categories; this situation is due to lack of enough number of subclasses in categories as described in three and six categories experiments.

4.9.3 Classification Using TF*IDF Weighting Scheme

After TF weight method is experimented for the three, six and nine categories, the effect of TF*IDF weight method is seen for same numbers of categories.

Like TF weighting scheme experiments, using the same datasets and categories, experimentation is held for TF*IDF weighting scheme for three, six and nine categories. Table 4.22 shows accuracies registered at various levels of epochs experiments.

Epoch	100	500	1000	1500	2000	2500	3000	3500	4000
3 Classes	69.63%	62.22%	65.19%	69.63%	65.19%	65.19%	65.19%	69.63%	62.22%
6 Classes	57.78%	70.22%	78.22%	57.78%	57.78%	58.22%	72.89%	57.78%	57.78%
9 Classes	62.09%	62.3%	61.48%	68.03%	62.3%	62.3%	62.3%	54.51%	62.3%

Table 4. 22: Accuracy Using TF*IDF Weighting Scheme at 3, 6 and 9 Classes at various epoch Levels

As TF weight method, TF*IDF discussion is based on epochs with similar accuracy result. Experiments which resulted similar accuracies have the same instances of news misclassified and of course the wrong category assigned is the same for all the experiments with similar accuracies.

The analysis for the TF*IDF weight scheme is presented for the three, six and nine categories experiments as follows.

4.9.3.1 Classification for Three Categories

In these experiments, it has been observed that all the news items in one of the three categories are misclassified. In 100, 1500 and 3500 epochs experiments, all news items of 'ICT' category are not assigned correctly; in 500 and 4000 epochs experiments all news items of category 'Art' are not classified correctly; and all the news of the category 'Educational coverage' are not

assigned to correct category in 1000, 2000, 2500 and 3000 epochs experiments. Most of the news items in two categories under each experiment are classified correctly except some, which are discussed below.

500 and 4000 epochs: An accuracy of 62.22% is resulted at these experiments. Here a different result is obtained. All the news items of the category 'Art' are misclassified. But it is the category with the largest number of news items. In addition to 'Art' category news items, there are two 'ICT' category news items which are classified wrongly. One contains the word 'eyer' ('አየር') and classified to the category 'Educational coverage'. As discussed in three categories experiment using TF weight method, the word is the keyword for the category 'Weather forecast'. But the news item is categorized to 'ICT' category in the test data due to error. Hence, the news item is misclassified to 'Educational coverage' category with the largest number of features during testing. The other news item contains the word 'mrmr' ('ምርምር'), which is keyword for the category 'Creativity work'. It may be due to this ambiguity that the news item is misclassified to the category 'Educational coverage', which has largest number of features.

1000, 2000, 2500 and 3000 epochs: An accuracy of 65.19% is registered at these experiments. All the news items of 'Educational coverage' category, the second larger in the number of news items, are misclassified. In these experiments, the news items which contain the word 'eyer' ('አየር') and 'mrmr' ('ምርምር') are misclassified like experiments in 500 and 4000 epochs. The only difference here is the category assigned is 'Art', which has the largest number of news items.

100, 1500 and 3500 epochs: 69.63% of accuracy is registered, which is the best result for the three categories experiment using TF*IDF weight method. In this experiment, all the news items

which belong to 'ICT' category are misclassified. This category is with the least number of news items than other categories. All news items in the categories 'Art and Educational coverage' are classified correctly, which are the first and the second in the number of news items they contain.

For two epoch level experiments, all the news items in the category 'Art' with the largest number of news items are misclassified. And for three epochs level experiments, all news items in category 'ICT' with the least number of news items are misclassified. And four level epochs experiments resulted in the misclassification of all news items in the category 'Educational coverage' with the second larger in the number of news items it contains. In all the cases, all news items are misclassified; which indicates the categories are not well represented by enough number of subclasses like all the above TF weight method experiments.

4.9.3.2 Classification for Six Categories

The experimentation for the six categories experiment using TF*IDF weight method results the following.

100, 1500, 2000, 3500 and 4000 epochs: for these epoch levels, least accuracy of 57.78% is registered. Except news items of 'ICT' category, all the news items of categories with the least number of news items are misclassified. These are Educational coverage, Religious assemblies and reports, and Creativity work. Additionally, there are two news items from the category 'ICT' which are misclassified to 'Weather forecast and Art' categories. The first news item contains the feature 'eyer' ('አየር'), which is misclassified to 'Weather forecast' category. The feature 'eyer' ('አየር') is the keyword for the category 'Weather forecast', it may be due to this fact that the misclassification occurs. As described in three categories experiments (500 and 4000 epochs),

this news item is contained in the test data of 'ICT' category due to error, which is not identified in misclassification identification of news during initial preprocessing of news items.

The second news item with the word 'mrmr' ('ምርምር') is misclassified to 'Weather forecast' category in 100 and 1500 epochs experiments and to 'Art' category in 2000, 3500 and 4000 epochs experiments. The category 'Art', with largest number of news items, is assigned wrongly three times in 2000, 3500 and 4000 epochs but the category 'Weather forecast' assigned wrongly two times in 100 and 1500 epochs.

500 epoch: An accuracy of 70.22% is resulted for this experiment. All the news items of the category 'Art' (largest number of news items) and 'Creativity work' category (least number of news items), are classified to wrong categories. Additionally, there are two news items which are misclassified like the above experiments, specifically 100 and 1500 epochs experiments.

2500 epoch: 58.22% of accuracy is recorded at this experiment. The same result is obtained like the 100, 1500, 2000, 3500 and 4000 epochs experiments except, the misclassified news item in the category 'ICT' with the word 'mrmr' ('ምርምር') is classified correctly.

3000 epoch: 72.89% of accuracy is registered here. All the news items in the category 'Creativity work' with the smallest number of news and 'Educational coverage' category with the third smallest number of news items are misclassified. But all the news items in the category of 'ICT, Art, Weather forecast, and Religious assemblies and reports', are classified correctly.

1000 epoch: The best accuracy of 78.22% for the TF*IDF weight method in the six categories experiment is registered at this epoch. Here, all the news items in the categories 'Religious assemblies and reports, and Creativity work' with the least number of news items are

misclassified. From the categories 'ICT, Art, Educational Coverage and Weather forecast', which have the largest number of news items, only two news items are misclassified. The two news items belong to the category 'ICT'. The first one contains the feature 'eyer' ('አየር') and the second contains the feature 'mrmr' ('ምርምር') as described in 100, 1500, 2000, 3500 and 4000 epochs experiments. In this experiment, they are misclassified to the categories 'Weather forecast and Educational coverage' (largest number of features) respectively.

The misclassification of all news items due to lack of enough number of subclasses in certain categories also appear in the six categories experiments using TF*IDF weight method. The misclassified news items assigned to categories with largest number of news items or features in most cases like the TF weight method experiments.

4.9.3.3 Classification for Nine Categories

The following parts discuss the results for the nine categories experimentation.

100 epoch: The same accuracy of 62.09% is resulted like the nine categories experiment using TF weight method at the 100 epoch. The difference here is, the misclassified news (all news) items are in the categories of 'ICT, Art, Educational coverage, Religious assemblies and reports, and Creativity work'. In the TF experiment, all news items of 'Weather forecast' category are misclassified instead of the category 'Educational coverage'. Only two news items are misclassified in the categories 'Bank and insurance, Tourism development, Mines and energy, and Weather forecast' which contain the largest number of news items with the exception of 'Weather forecast' category.

In both TF and TF*IDF experiments for the nine categories, there are two similar misclassified news items in the categories 'Bank and insurance, and Mines and energy'. The first news item is with the features 'mereja' ('መረጃ') and 'faynans' ('ፋይናንስ') and the second one is with the features 'heyl' ('ሀይል'), 'bdr' ('ብድር'), 'ElEktrik' ('ኤሌክትሪክ') and 'bank' ('ባንክ'). The first one, in this case, is misclassified to the category 'Mines and energy', third largest in the number of news items; in the TF experiment it is misclassified to 'Tourism development' category, second largest in the number of news items. The second one is classified wrongly to the category 'Bank and insurance' like the TF experiment. The possible reason of the two news items misclassification is discussed at 100 epoch experiment for the nine categories using TF weight method.

500, 2000, 2500, 3000 and 4000 epochs: as of the nine categories experiment in TF weight method at 500 and 2000 epochs, the same accuracy of 62.3% is registered at these experiments. In these experiments, similar result is registered like the above experimentation except the misclassified news item in the category 'Bank and insurance' is correctly classified as mentioned in the nine categories experiment at 500 and 2000 epochs using TF weight method.

1000 epoch: here again similar result with accuracy of 61.48% is recorded like 1000 and 3000 epochs experiment for the nine categories experiment using TF weight method.

Similar result is seen by observing the test data like the 100 epochs experiment of this section. But, three additional news items belonging to the category 'Bank and insurance', are misclassified to the category 'Mines and energy', third largest in the number of news items it contains. These three news items are the same with 1000 and 3000 epochs experiment for the

nine categories using TF weight method except, in this case they are misclassified to the category 'Tourism development', second largest in the number of news items it contains.

3500 epoch: Least accuracy of 54.51% is registered. All the news items of the categories 'Tourism development, ICT, Art, Religious assemblies and reports, and Creativity work' are classified incorrectly. With the exception of 'Tourism development' category, all the categories contain less number of news items. From the other categories, 'Educational coverage, Weather forecast, Mines and energy, and Bank and insurance', only two news items are misclassified. One news item belongs to the category 'Bank and insurance' but misclassified to 'Weather forecast' category. The other one is in the 'Mines and energy' category but 'Bank and insurance' category is assigned wrongly. The first news item is with the features 'mereja' ('መረጃ') and 'faynans' ('ፋይናንስ'); and the second news item contains the features 'mereja' ('መረጃ'), 'heyl' ('ሀይል'), 'bdr' ('ብድር') and 'ElEktrik' ('ኤሌክትሪክ').The discussion in 100 epochs nine categories TF weight method experiment, can be applied here for the two news items.

1500 epoch: Best result with an accuracy of 68.03% is achieved in the nine categories experiment using TF*IDF weight method. All the news items of the categories 'Art, Educational coverage and Creativity work' are misclassified. And 46 news items are misclassified from the categories 'Bank and insurance, Tourism development, Mines and energy, ICT, Weather forecast, and Religious assemblies and reports'. Observation of the data reveals that the possible reasons of misclassification are similar to the above experiments.

Like all the experiments, the nine categories experiments are the same in that all news items of certain categories are misclassified whereas very few or none misclassifications occur in the other

classes. The all news items misclassification in certain categories resulted due to lack of enough number of subclasses for those categories as mentioned earlier.

4.9.4 Comparison of TF and TF*IDF Weight Schemes Results

The results of experiments using TF and TF*IDF weight schemes show the reasons of misclassifications are similar in most cases. It is found that news items which are not classified correctly in TF weight method also found to be misclassified in the TF*IDF weight method. For example, in nine categories experiment at 100 epoch, in both TF and TF*IDF weight methods, two misclassified news are exactly the same. The first news item contains the features 'mereja' ('መረጃ') and 'faynans' ('ፋይናንስ') and the second news item contains the features 'heyl' ('ሀይል'), 'bdr' ('ብድር'), 'ElEktrik' ('ኤሌክትሪክ') and 'bank' ('ባንክ') in both TF and TF*IDF weight methods.

Though there exists the fore mentioned similarity, one method register better accuracy than the other by varying epoch. The difference in accuracy is due to the switching of the correctly classified and the incorrectly classified news items of categories. For instance, in six categories experiment, 61.61% of accuracy is recorded, which is the least accuracy for TF weight method; the misclassified news items belong to the categories 'ICT, Religious assemblies and reports, and Creativity work'. In contrary, for the TF*IDF weight method, best accuracy of 78.22% is recorded at this experiment. The difference here is, the news items in the category 'ICT' are among the correctly classified news items except two news items.

Table 4.23 shows the best accuracy scored in three, six and nine categories experiments and the average accuracy using the TF and TF*IDF weight methods.

Classes	News No.	Accuracy	
		TF	TF*IDF
Three	406	94.81%	69.63%
Six	678	61.61%	78.22%
Nine	1463	70.08%	68.03%
Average		75.5%	71.96%

Table 4. 23: Accuracy at Increasing No. of Classes and News Using TF and TF*IDF Weight Schemes

From Table 4.23, it can be observed that in the three categories experiment, TF weight method is better than TF*IDF weight method by 25.18%. But for the six categories experiment TF*IDF weight method is better than TF weight method by 16.61% than TF weight method. The result of nine categories experiment testifies that TF weight method scored better accuracy than TF*IDF weight method by 2.05%. The average of all the experiments indicates that TF weight method registered better accuracy by 3.54% than TF*IDF weight method.

The main performance difference between the two weighting schemes happens because of the range of values in the weighting schemes. In the datasets, TF weight value is between 0 and 5 and TF*IDF weight value is in the range of 0 and 45. This affects the classifier accuracy. Because, according to Thulasiraman (2005), it is recommended to have maximum value of 1 and minimum value of -1 for the input pattern of LVQ algorithm. This seems plausible for the greater accuracy result of TF weight method than TF*IDF weight method.

4.9.5 Performance at Increasing Number of Categories and News

As depicted in Table 4.23, using TF weight method the best accuracy is registered at three categories experiment. The least accuracy is recorded at the six categories experiment. The nine categories experiment, resulted accuracy lower than the three categories but higher than the six

categories experiment. Hence, we can say that the increase in the number of categories and news are not the determinant factor for the decrease of performance with regard to the LVQ algorithm.

Based on Table 4.23, least accuracy for the TF*IDF weight method is scored in the nine categories experiment with less (1.6%) difference in the three categories experiment. The best accuracy for this weighting method is recorded in the six categories experiment. The three categories experiment resulted in the second best accuracy. Like the TF weight method, it can be said that the increase in the number of categories and news items are not the major factors in the reduction of performance using LVQ algorithm for the Amharic text classifier.

The above description is the effectiveness aspect of performance based on the accuracy registered. When performance is seen from efficiency perspective, there is an impact as the size increases both in category and number of news items. During training, it has been observed that the time required for getting the output of the training increases as the number of news items, categories and features increases.

4.10 Discussion

The text news classifier for Amharic is based on preprocessing made for selecting features which can represent news. Tokenization produces the tokens generated from each news item. But the number of tokens is too huge, 5129, to be used as features, which is computationally intensive. As a result, preprocessing methods such as stop word and number removal, and stemming are made which together helped to reduce the number of tokens by 1881 or 36.48%. Even if substantial amount of reduction is achieved by the preprocessing, still the number of tokens is too large, about 3258. Hence, dimension reduction using Document Frequency (DF) thresholding is

applied and final tokens which can be used as features are selected. Using dimension reduction technique 80 features are identified when the numbers of features are summed; but that becomes 69 when the features are merged to form matrix, the reduction is due to common features. Here, the matrix generated is using the features and, TF and TF*IDF value of each feature that is fed to LVQ learning algorithm.

The important thing that is noticed in this study is unavailability of standard tools which can be used to preprocess Amharic documents. If standardized stemmer and stop word list are ready, the result of this study testifies that considerable reduction of features can be reached.

After matrix is generated, experiment is carried out using TF and TF*IDF weight method. The results of the experiments witness that TF weight method is better than TF*IDF weight method by 3.54% when the average of the three, six and nine categories experiments is taken. As indicated, LVQ works better in a situation when input patters are normalized between -1 and 1 according to Thulasiraman (2005); which may be the reason why TF outperform than TF*IDF. The normalized data has to be tried after the raw data is tried, which is the view followed in this study by considering Russell, Eberhart and Shi (2007) suggestion. The raw data is tried and time limitation inhibits to try the normalized one.

As indicated in the analysis of all experiments, most of the misclassifications occur in all news items which belong to a certain category. To clarify, an example is taken from the nine categories experiment using TF*IDF weight method at 100 epoch. All the news items belonging to the categories 'ICT, Art, Educational coverage, Religious assemblies and reports, and Creativity work', are misclassified. Where as only two news items are misclassified in categories of 'Bank and insurance, and Mines and energy'. And all the news items of the categories 'Tourism

development and Weather forecast' are classified correctly. The misclassification of all news items in certain categories and the correct classification of all or most news items in certain other categories is the problem of assigning enough number of hidden neurons or subclasses for categories that their news items are misclassified.

The hidden neurons determine the number of subclasses; subclasses are the classes which transform input data to target. If each category is not well represented by enough number of subclasses, its news items are misclassified to a category with the highest number of instances or with the highest number of features according to this study. The categories those their news items are not totally classified correctly are, categories that are not well represented by subclasses or may not have subclass representation at all. Hence, in classification of Amharic news using LVQ algorithm, not representing each class with enough number of subclasses is the major reason for misclassification rather than the number of categories and news items.

Self Organizing Map (SOM) learning method may improve the problem of hidden neurons (subclasses) number according to Khalifa and et. al. (2004).Both SOM and LVQ are first created by Kohonon, SOM is used in unsupervised manner to determine clusters. Whereas, LVQ works in a supervised manner; after subclasses are determined, which transform input data to the defined targets. Hence, the clusters obtained by SOM in unsupervised manner can be used as an input in creating LVQ network for the determination of the number of hidden neurons (subclasses).

In this study, it is found that keywords in one category may be found in another category; which is one of the reasons for misclassifications. For example, the term 'mrmr' ('ምርምር') is the keyword for the category 'Creativity work' but is found in the categories 'ICT and Educational

coverage'. If feature representation is based on phrase or ontology, each category may be represented and discriminated well than the word level features. Phrases can discriminate categories than mere words are used. And use of ontology enables to define concepts and relations representing knowledge about a particular document in domain specific terms according to Paralic and Kostial (2004).

CHAPERT FIVE

CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

The size of electronic documents is increasing from time to time at an alarming rate. The issue in this digital age is, how can we make use of the best from the massive available information? Various researches have been conducted in different contexts to devise techniques which can enable to change threats into opportunities, in this case for wise use of information instead of information overload.

News items are among the information produced and dispatched electronically; growing in size as time goes on around the globe. Ethiopia is not exceptional; in that, there are lots of electronic news items created and distributed. As the size of news items increase, utilization is affected badly unless countermeasure is taken to maximize utilization. Classification is one of the methods that can be employed to organize information for effective and efficient use. Manual classification is hardly possible with the incredible increase in the volume of data; as a result, automatic classification is very important area of research specifically for Amharic since the language is technologically unfavored. This research work contributes on this regard.

Preprocessing, classifier construction or building and testing are the major steps for the accomplishment of this study. Preprocessing the data is worked out before the datasets are fed into the learning algorithm, Learning Vector Quantization. The final goal of preprocessing is to produce a term by news matrix.

of each category with enough number of neurons is the major reason for the decrease in accuracy. On the contrary, previous researches on Amharic text classification, shows that continuous decrease in performance as the number of news and categories increase.

- This study shows that Learning Vector Quantization can be employed to automatic Amharic text classification but an integration of standard preprocessing techniques is crucial. Besides, clustering datasets before submitting to classifier helps to define optimum number of subclasses (hidden neurons) for each category.

- Confusing words that occurs in more than one category affect classification performance negatively.

- In the course of training using LVQ algorithm, it is found that computational time increases as the number of news items, features and categories increases.

5.2 Recommendations

This study aims to see the potential application of Amharic text news classification using Learning Vector Quantization and it is possible to construct classifier using LVQ-neural network algorithm. But the accuracy obtained has to be improved. The recommendations given revolve around improving accuracy, facilitating Amharic text classification, or untouched problems related to text classification for Amharic context and recommendations for the agency, ENA.

Stemmer: Standard stemmer that can be applied on Amharic is vital to decrease feature size. The result obtained on this study is encouraging in reducing the size of features by applying stemmer. If standard stemmer is available, it can play great role in the reduction of features as a result

Subclass Determination: The determination of subclasses is not done using clustering algorithms in this study. Hence, clustering algorithms like Self organizing Map (SOM) can be used to find the number of subclasses (hidden neurons) that can be used as an input in the parameter setup for creating LVQ network. The use of clustering algorithms enables to use appropriate number for subclasses so that each category is represented with enough number of neurons for better classification.

Corpus Preparation: In other languages, it is very common to prepare corpus for research purpose; unfortunately, we are not lucky for not having standard corpus for Amharic text classification, as to the researcher's knowledge. Researchers can devote much time on their work if standard corpus is prepared for Amharic classification experiments like 'Reuters-21578' for English.

Feature Preparation: The features that can represent categories are selected using words in this study. But confusion occurred when the words are common across categories that resulted in misclassifications. Hence, there is a need to undergo research on text classification that considers features selected based on phrases or using ontology.

Classification Types: The data on ENA SQL server exhibit hierarchical in nature. As far as the researcher's knowledge, this problem is not researched for Amharic. So, this is potential area of research. And some news items in ENA reveal the characteristics of more than one class. But ENA uses only single-label classification scheme. So, it is recommended for ENA to start implementation of multi-label classification scheme so that the true characteristics of news items are exhibited. This also helps researchers to undergo study on multi-label classification of Amharic news.

ENA: Manual classification is used in ENA till now. The results of Amharic text news classification researches are promising. Hence, the company shall start to think the implementation on automatic classification for Amharic news.

Other Domains: As to the knowledge of the researcher, Amharic text classification is still tried on news items text only. Other areas, like classifying 'research papers', can be researched for Amharic documents.

REFERENCE

1. Anderson, J. (2006). An Introduction to Neural Networks. Prentice-Hall: New Delhi.
2. Atelach, A. (2002). Automatic Sentence Parsing for Amharic Text an Experiment Using Probabilistic Context Free Grammars. Thesis. Addis Ababa University: Addis Ababa.
3. Baeza-Yates, R. and Ribeiro-Neto, B. (1999). Modern Information Retrieval. Addison-Wesley: New York.
4. Bender, M.; Bowen, J.; Cooper, R. and Ferguson, C. (1976). Language in Ethiopia. Oxford University Press: London.
5. Bi, Y.; Murtagh, S. and Anderson, T. (1999).Text Passage Classification Using Supervised Learning. Retrieved on August 25, 2008.Web site: http://ir.dcs.gla.ac.uk/lumis99/papers/bi.pdf.
6. Blumberg, R. and Atre, S. (2003).Automatic Classification Moving to the Mainstream. Retrieved on September 25, 2008.Web site: www.soquelgroup.com/Articles/dmreview_0403_classification.pdf.
7. Calvo, R.; Lee, J.; and Li, X. (2004). Managing Content with Automatic Document Classification. Journal of Digital Information, 5(2), np.
8. Crammer, K.; Gilad-Bachrach, R. and Navot, A.(2003). Margin Analysis of the LVQ Algorithm. Retrieved on October 24, 2008.Web site: www.books.nips.cc/papers/files/nips15/LT19.pdf.
9. Curtis, A. (2002). Neural Network Architectures. Retrieved on February 24, 2009.Web site: http://www1.pacific.edu/~acurtis/architectures.htm.
10. Daniel, Y. (1996). Frequently Asked Questions about SERA. Retrieved on October 23, 2008. Web site: http://www.abyssiniacybergateway.net/fidel/sera-faq.txt.

11. Demuth, H. and Beale, M. (2004). Neural Network Toolbox: For Use with Matllab.

12. Ethiopia, T. (2002). Application of Case-Based Reasoning for Amharic Legal Precedent retrieval: a Case Study with the Ethiopian Labor Law. Thesis. Addis Ababa University: Addis Ababa.

13. Encyclopedia Britannica (1992). The New Encyclopedia Britannica. Vol. 4. Chicago: Encyclopedia Britannica.

14. Frakes, W. and Baeza-Yates, R. (2002). Information Retrieval: Data Structures and Algorithms. Prentice Hall.

15. Ghosh, A. (2007). Robustness of Shape Descriptors and Dynamics of Learning Vector Quantization. Academic Press Europe: Groningen.

16. Giorgino, T. (2004). An Introduction to Text Classification. Retrieved on October 13, 2008.Web site: www.cs.cornell.edu/people/tj/publications/joachims_98a.pdf

17. Goyal, N. (2009). Classifier Accuracy. Retrieved on October 13, 2008.Web site: http://csis.bitspilani.ac.in/faculty/goel/Data%20Mining/Lecture%20Slides/Classifier%20Accuracy.ppt.

18. Haykin, S. (1999). Neural Networks: A Comprehensive Foundation. Prentice-Hall: New Jersey.

19. Heaton, J. (2005). Introduction to Neural Networks with Java. Heaton Research: USA.

20. Hooper, R. and Paice, C. (2005). What is Porter Stemming? Retrieved on October 15, 2008. Web site: http://www.comp.lancs.ac.uk/computing/research/stemming/general/porter.htm.

21. Ifrim, G.; Theobald, M. and Weikum, G. (2005). Learning Word-to-Concept Mappings for Automatic Text Classification. Retrieved on October 13, 2008.Web site: www.infolab.stanford.edu/~theobald/pub/icml-lws05.pdf.

22. Khalifa, K.; Bedoui, M.; Dougui, M. and Alexandre, F. (2004). Alertness States Classification by SOM and LVQ Neural Networks. World Academy of Science, Engineering and Technology, 3 (2). np.

23. Klein, B. (2004).Text Categorization or Classification. Retrieved on October 12, 2008. Web site: http://www.bklein.de/text_classification.php.

24. Krishnakumar, A. (2006). Text Categorization: Building a KNN Classifier for the Reuters-21578 Collection. Retrieved on October 12, 2008. Web site: http://en.scientificcommons.org/42606011.

25. Lavesson, N. (2003). Evaluation of Classifier Performance and the Impact of Learning Algorithm Parameters. Thesis. Blekinge Institute of Technology: Sweden.

26. Liao, C.; Alpha, S. and Dixon, P. (2003). Feature Preparation in Text Categorization. Retrieved on October 12, 2008.Web site: http//www.oracle.com/technology/products/text/pdf/feature_preparation.pdf.

27. Maly, K.; Zeil, S.; Zubair, M. and Ratkal, N. (2007).A Machine Learning Approach for Automatic Text Categorization. Retrieved September 24, 2008.Web site: www.cs.odu.edu/~extract/publications/icsiitk05-25a.doc.

28. Manning, C.; Raghavan, P. and Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press: Cambridge.

29. Martín-Valdivia, M.; Ureña-López, L. and García-Vega, M. (2007). The Learning Vector Quantization Algorithm Applied to Automatic Text Classification Tasks. Science Direct, 20(6), pp. 748-756.

30. Math Works (2009). Neural Networks Key Features. Retrieved on September 27, 2008.Web site: http://www.mathworks.com/products/neuralnet/description1.html.

31. Michie, D.; Spiegelhalter, D. and Taylor, C. (1994). Machine Learning, Neural and Statistical Classification. Retrieved on October 13, 2008.Web site: http://www.amsta.leeds.ac.uk/~charles/statlog/whole.pdf.

32. Monica Ines, M. (2001).Research Methodology. Retrieved on October 14, 2008. Web site:http://scholar.lib.vt.edu/theses/available/etd10292001151807/unrestricted/22chaptIV methods.pdf.

33. Nega, A. and Willett (2002). Stemming of Amharic Words for Information Retrieval. Literary and Linguistic Computing, 17 (1), pp. 1-17.

34. Novovicová, J. (2005).Text Document Classification. ERCIM News, No. 62. Retrieved on September 24, 2008. Web site: www.ercim.org/publication/Ercim_News/enw62/novovicova.html.

35. Paralic, J. and Kostial, I. (2004). Ontology-based Information Retrieval. Retrieved on May 23, 2009. Web site: http://people.tuke.sk/jan.paralic/papers/IIS03.pdf.

36. Russell, C; Eberhart and Shi, Y. (2007). Computational intelligence: Concepts to Implementation. Morgan Kaufmann Publishers: San Fransisco.

37. Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. ACM Computing Surveys, 34(1), 1-47.

38. Sebastiani, F. (2005). Text categorization. In Zanasi, A. (ed.), Text Mining and its Applications. WIT Press: Southampton, pp. 109-129.

39. Sebastiani, F. (2006). Classification of Text, Automatic. The Encyclopedia of Language and Linguistics(Vol. 14, pp. 457-462). Elsevier Science Publishers: Amsterdam.

40. Skarmeta, A.; Bensaid, A. and´ Tazi, N.(2000).Data Mining for Text Categorization with Semi-Supervised Agglomerative: Hierarchical Clustering. International Journal of Intelligent systems, 15, pp.633-646.

41. Solomon, T. and Menzel, W. (2007). Syllable-Based Speech Recognition for Amharic. Proceedings of the 5th Workshop on Important Unresolved Matters, pp. 33–40.

42. Stergiou, C. and Siganos, D. (1997).Neural Networks. Retrieved on September 24, 2008. Web site: http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html#Contents.

43. Surafel, T. (2003). Automatic Categorization of Amharic News Text: A Machine Learning Approach. Thesis. Addis Ababa University: Addis Ababa.

44. Tewodros, H. (2003). Amharic Text Retrieval: an Experiment Using Latent Semantic Indexing (LSI) with Singular value Decomposition (SVD). Thesis. Addis Ababa University: Addis Ababa.

45. Thulasiraman, P. (2005). Semantic Classification of Rural and Urban Images Using Learning Vector Quantization. Thesis. Madras University: India.

46. Umer, M. and Khiya, S. (2007). Classification of Textual Documents Using Learning Vector Quantization. Information Technology Journal, 6(1), pp.154-159, 2007.

47. Wang, Y, Zang, H., Spencer, B. and Yan, Y. (2005). A Text Categorization Approach for Match-Making in Online Business Tendering. Journal of business and technology, 1(1), np.

48. Wapedia (2009). አማርኛ. Retrieved on April 24, 2009. Web site: http://wapedia.mobi/am/.

49. Yi, K. and Beheshti, J. (2004).A Comparative Study on Feature Selection of Text Categorization for Hidden Markov Models. Retrieved on September 24, 2008.Web site: http://www.jsbi.org/journal/GIW02/GIW02F006.pdf.

50. Yohannes, A. (2007). Automatic Amharic Text Categorization. Thesis. Addis Ababa University: Addis Ababa.

51. Zelalem, S. (2001). Automatic Classification of Amharic News Items: The case of Ethiopian News Agency. Thesis. Addis Ababa University: Addis Ababa.

APPENDIX

Appendix 1: Interview with the ICT Coordinator of ENA

1. How news are created in ENA and distributed for readers?
2. In what manner the data are stored? What amount of data is available? If anything else related to the data, would you mined to elaborate?
3. Currently, most organizations are becoming familiar with information technology. What kind of technology do ENA use for managing news? For what purpose the use is/are?

Appendix 2: Amharic Characters ('Fidel')-ፊደል (Zelalem, 2001)

Order							Labialized				
1st	2nd	3rd	4th	5th	6th	7th					
ሀ ha	ሁ hu	ሂ hi	ሃ ha	ሄ he	ህ h	ሆ ho					
ለ lä	ሉ lu	ሊ li	ላ la	ሌ le	ል l	ሎ lo			ሏ $l^{w}a$		
ሐ ha	ሑ hu	ሒ hi	ሓ ha	ሔ he	ሕ h	ሖ ho					
መ mä	ሙ mu	ሚ mi	ማ ma	ሜ me	ም m	ሞ mo			ሟ $m^{w}a$		
ሠ sä	ሡ su	ሢ si	ሣ sa	ሤ se	ሥ s	ሦ so					
ረ rä	ሩ ru	ሪ ri	ራ ra	ሬ re	ር r	ሮ ro			ሯ $r^{w}a$		
ሰ sä	ሱ su	ሲ si	ሳ sa	ሴ se	ስ s	ሶ so			ሷ $s^{w}a$		
ሸ šä	ሹ šu	ሺ ši	ሻ ša	ሼ še	ሽ š	ሾ šo			ሿ $š^{w}a$		
ቀ qä	ቁ qu	ቂ qi	ቃ qa	ቄ qe	ቅ q	ቆ qo	ቈ $q^{w}ä$	ቍ $q^{w}i$	ቋ $q^{w}a$	ቌ $q^{w}e$	ቍ $q^{w}ə$
በ bä	ቡ bu	ቢ bi	ባ ba	ቤ be	ብ b	ቦ bo			ቧ $b^{w}a$		
ተ tä	ቱ tu	ቲ ti	ታ ta	ቴ te	ት t	ቶ to			ቷ $t^{w}a$		
ቸ čä	ቹču	ቺ či	ቻ ča	ቼ če	ች č	ቾ čo			ቿ $č^{w}a$		
ኀ hä	ኁ hu	ኂ hi	ኃ ha	ኄ he	ኅ h	ኆ ho	ኈ $h^{w}ä$	ኊ $h^{w}i$	ኋ $h^{w}a$	ኌ $h^{w}e$	ኍ $h^{w}ə$
ነ nä	ኑ nu	ኒ ni	ና na	ኔ ne	ን n	ኖ no			ኗ $n^{w}a$		
ኘ ñä	ኙ ñu	ኚ ñi	ኛ ña	ኜ ñe	ኝ ñ	ኞ ño			ኟ $ñ^{w}a$		
አ a	ኡ u	ኢ i	ኣ a	ኤ e	እ ə	ኦ o					
ወ wä	ዉ wu	ዊ wi	ዋ wa	ዌ we	ው w	ዎ wo					
ዐ a	ዑ u	ዒ i	ዓ a	ዔ e	ዕ ə	ዖ o					
ከ kä	ኩku	ኪ ki	ካ ka	ኬ ke	ክ k	ኮ ko	ኰ $k^{w}ä$	ኲ $k^{w}i$	ኳ $k^{w}a$	ኴ $k^{w}e$	ኵ $k^{w}ə$
ኸ hä	ኹhu	ኺhi	ኻha	ኼhe	ኽ h	ኾ ho					
ዘ zä	ዙ zu	ዚ zi	ዛ za	ዜ ze	ዝ z	ዞ zo			ዟ $z^{w}a$		
ዠ žä	ዡžu	ዢži	ዣža	ዤže	ዥ ž	ዦ žo					
የ yä	ዩ yu	ዪyi	ያ ya	ዬ ye	ይ y	ዮ yo					
ገ gä	ጉ gu	ጊ gi	ጋ ga	ጌ ge	ግ g	ጎ go	ጐ $g^{w}ä$	ጒ $g^{w}i$	ጓ $g^{w}a$	ጔ $g^{w}e$	ጕ $g^{w}ə$
ደ dä	ዱdu	ዲ di	ዳ da	ዴde	ድ d	ዶ do			ዷ $d^{w}a$		
ጀ ǧä	ጁǧu	ጂǧi	ጃǧa	ጄǧe	ጅǧ	ጆǧo					
ጠ ṭä	ጡṭu	ጢṭi	ጣṭa	ጤṭe	ጥṭ	ጦṭo			ጧ $ṭ^{w}a$		
ጨ ć ä	ጩć u	ጪć i	ጫć a	ጬć e	ጭ ć	ጮć o			ጯ $ć^{w}a$		
ጸ şä	ጹ şu	ጺ şi	ጻ şa	ጼ şe	ጽ ş	ጾ şo			ጿ $ş^{w}a$		
ፀ ṣä	ፁ ṣu	ፂ ṣi	ፃ ṣa	ፄ ṣe	ፅ ṣ	ፆ ṣo					
ፈ fä	ፉ fu	ፊ fi	ፋ fa	ፌ fe	ፍ f	ፎ fo			ፏ $f^{w}a$		
ፐ pä	ፑ pu	ፒ pi	ፓ pa	ፔ pe	ፕ p	ፖ po					
ቨ vä	ቩ vu	ቪ vi	ቫ va	ቬ ve	ቭ v	ቮ vo					

Appendix 3: Amharic Punctuation Marks (Atelach, 2002)

No.	Punctuation mark	Symbol	Purpose
1	The four dots or double colon	::	Mark end of a sentence
2	Colon	:	Separate words in a sentence: not common
3	White space		Separate words in a sentence: current practice
4	Question mark	?	Placed at the end of questions
5	Exclamation mark	!	Used at the end of sentences that show exclamation
6	Comma	፣	Used like comma
7	Semi-colon	፤	Used like semi-column
8	Three dots	...	For deliberate omission of words, phrases, or sentences
9	Quotation marks	«»	Used at the beginning and at the end of quoted word, phrase, etc.
10	Parenthesis	()	To enclose elaboration
11	Stroke	/	Separate date, month, etc.
12	Mocking mark	፧	Placed at the end of mocking sentence

Appendix 4: Amharic Numbers (Zelalem, 2001)

1	፩	6	፮	20	፳	70	፸
2	፪	7	፯	30	፴	80	፹
3	፫	8	፰	40	፵	90	፺
4	፬	9	፱	50	፶	100	፻
5	፭	10	፲	60	፷	1000	፼

Appendix 5: Amharic Script Translation to Latin Script for Preprocessing Purpose

ሀ	he	ሩ	ru	ቨ	vi	ኘ	Na
ለ	le	ሱ	su	ቲ	ti	አ	a
ሐ	He	ሹ	xu	ቺ	ci	ካ	ka
መ	me	ቁ	qu	ኀ	`hi	ኻ	Ha
ሠ	`se	ቡ	bu	ኒ	ni	ዋ	wa
ረ	re	ቩ	vu	ኚ	Ni	ዓ	`a
ሰ	se	ቱ	tu	ኢ	i	ዛ	za
ሸ	xe	ቹ	cu	ኪ	ki	ዣ	Za
ቀ	qe	ኁ	`hu	ኺ	Hi	ያ	ya
በ	be	ኑ	nu	ዊ	wi	ዳ	da
ቨ	ve	ኙ	Nu	ዒ	`i	ጃ	ja
ተ	te	ኡ	u	ዚ	zi	ጋ	ga
ቸ	ce	ኩ	ku	ዠ	Zi	ጣ	Ta
ኀ	`he	ኹ	Hu	ዪ	yi	ጫ	Ca
ነ	ne	ዉ	wu	ዲ	di	ጳ	Pa
ኘ	Ne	ዐ	`u	ጂ	ji	ጻ	Sa
አ	e	ዙ	zu	ጊ	gi	ፁ	`Sa
ከ	ke	ዠ	Zu	ጢ	Ti	ፋ	fa
ኸ	He	ዩ	yu	ጪ	Ci	ፓ	pa
ወ	we	ዱ	du	ጲ	Pi	ፄ	hE
ዐ	`e	ጁ	ju	ጺ	Si	ሌ	lE
ዘ	ze	ጉ	gu	ፂ	`Si	ሔ	HE
ዠ	Ze	ጡ	Tu	ፊ	fi	ሜ	mE
የ	ye	ጩ	Cu	ፒ	pi	ሤ	`sE
ደ	de	ጱ	Pu	ሃ	ha	ሬ	rE
ጀ	je	ጹ	Su	ላ	la	ሴ	sE
ገ	ge	ፀ	`Su	ሐ	Ha	ሼ	xE
ጠ	Te	ፉ	fu	ማ	ma	ቄ	qE
ፘ	Ce	ፑ	pu	ሣ	`sa	ቤ	bE
ጰ	Pe	ሂ	hi	ራ	ra	ቬ	vE
ጸ	Se	ሊ	li	ሳ	sa	ቴ	tE
ፀ	`Se	ሒ	Hi	ሻ	xa	ቼ	cE
ፈ	fe	ሚ	mi	ቃ	qa	ኄ	`hE
ፐ	pe	ሢ	`si	ባ	ba	ኔ	nE
ሁ	hu	ሪ	ri	ቫ	va	ኼ	NE
ሉ	lu	ሲ	si	ታ	ta	ኤ	E
ሑ	Hu	ሺ	xi	ቻ	ca	ኬ	kE
ሙ	mu	ቂ	qi	ኃ	`ha	ኼ	HE
ሡ	`su	ቢ	bi	ና	na	ዌ	wE

ዔ	`E
ዜ	zE
ዠ	ZE
ዬ	yE
ዴ	dE
ጀ	jE
ገ	gE
ጠ	TE
ጨ	CE
ጰ	PE
ጸ	SE
ፀ	`SE
ፈ	fE
ፐ	pE
ሀ	h
ለ	l
ሐ	H
መ	m
ሠ	`s
ረ	r
ሰ	s
ሸ	x
ቀ	q
በ	b
ቨ	v
ተ	t
ቸ	c
ኀ	`h
ነ	n
ኘ	N
አ	I
ከ	k
ኸ	H
ወ	w
ዐ	`I
ዘ	z
ዠ	Z
ይ	y
ደ	d
ጀ	j
ገ	g
ጠ	T
ፉ	C

ጰ	P
ጸ	S
ፀ	`S
ፍ	f
ፕ	p
ሆ	ho
ሎ	lo
ሐ	Ho
ሞ	mo
ሦ	`so
ሮ	ro
ሶ	so
ሾ	xo
ቆ	qo
ቦ	bo
ቮ	vo
ቶ	to
ቾ	co
ኆ	`ho
ኖ	no
ኞ	No
ኦ	o
ኮ	ko
ኽ	Ho
ዎ	wo
ዖ	`o
ዘ	zo
ዠ	Zo
ዮ	yo
ዶ	do
ጆ	jo
ጎ	go
ጠ	To
ፅ	Co
ጰ	Po
ጸ	So
ፀ	`So
ፍ	fo
ፖ	po

ሯ	rWa
ሷ	sWa
ⶓ	xWa
ⵇ	qWe
ⵈ	bWa
ⵔ	vWa
ⵜ	tWa
ⵛ	cWa
ⵀ	hWe
ⵏ	nWa
ⵏ	NWa
ⵅ	kWe
ⵣ	zWa
ⵥ	ZWa
ⴷ	dWa
ⴵ	jWa
ⴳ	gWe
ⵟ	TWa
ⵚ	CWa
ⵖ	PWa
ⵙ	SWa
ⶒ	fWa
ⵒ	pWa
ⵇ	qWu
ⵀ	hWu
ⵅ	kWu
ⴳ	gWu
ⵇ	qWi
ⵀ	hWi
ⵅ	kWi
ⴳ	gWi
ⵇ	qWa
ⵃ	hWa
ⵄ	kWa
ⵂ	gWa
ⵇ	qWE
ⵃ	hWE
ⵄ	kWE
ⵂ	gWE
ⵍ	ea
፡	.
፣	,
፤	;

÷	:
፥	:-
፧	?
?	?
/	/
:	:
.	.
"	"
-	-
#	#
$	$
%	%
*	*
(	(
)	)
_	_
+	+
=	=
\	\
{	{
}	}
<	<
>	>
~	~
0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9

Appendix 6: News Items Major and Sub Categories in ENA

ID.	Category Code in Amharic	Category Description	Category Description in Amharic	Parent category
2	ስፖ	Sport	ስፖርት	No
3	ጤ	Health	ጤና	No
4	ት	Education	ትምህርት	No
5	ፖ	Politics	ፖለቲካ	No
6	ፖ-ብ	National Politics	ብሔራዊ ፖለቲካ	5
7	ፖ-ኢ	International Politics	አለም አቀፋዊ ፖለቲካ	5
8	ፖ	Test	አንደኛ ደረጃ	4
10	ሁ	Secondary	ሁለተኛ ደረጃ	4
15	ኢኮ	Economy	ኢኮኖሚ	No
16	የ	Foreign Relation, defense and security	የውጭ ግንኙነት መከላከያና ደህንነት	No
17	ማህ	Social	ማህበራዊ	No
18	ባ	Culture and Tourism	ባህልና ቱሪዝም	No
19	ህ	Law and Justice	ህግና ፍትህ	No
20	ሳይ	Science and Technology	ሳይንስና ቴክኖሎጂ	No
21	አካ	Environment Preservation and wether condition	የአካባቢ ጥበቃና የአየር ሁኔታ	No
22	አ	Accident	አደጋ	No
23	ሰላ	Peace and Stability	ሰላምና መረጋጋት	5
24	ዲ	Democracy and Good Governance	ዲሞክራሲና መልካም አስተዳደር	5
25	ሰብ	Humaniterian and Democratic Rights	ሰብአዊና ዲሞክራሲያዊ መብቶች	5
26	ምር	Election	ምርጫ	5
27	ውይ	Discussions, Rules and Regulations	ውይይቶች ውሳኔዎችና አዋጆች	5
28	ፖለ	Political Parties	የፖለቲካ ፓርቲዎች	5
29	ፖሹ	Political Appointment	የፖለቲካ ሹመት	5
30	ግብ	Agriculture and Rural Development	የግብርናና ገጠር ልማት	15
31	ኢድ	Industrial Development	የኢንዱስትሪ ልማት	15
32	መሰ	Infrastructure Development	የመሰረተ ልማት	15
33	ንግ	Trade	ንግድ	15
34	ኢቨ	Investment	ኢንቨስትመንት	15
35	ማኢ	Micro Enterprise	ማይክሮ ኢንተርፕራይዝ	15
36	እር	Donation and Development	እርዳታና የልማት ትብብር	15

		Agreement		
37	ማዕ	Mines and Energy	ማዕድንና ኢነርጂ	15
38	ውሃ	Water Resource	የውሃ ሃብት	15
39	ባን	Bank and Insurance	ባንክና ኢንሹራንስ	15
40	አ.ዕ	Overall Economic Growth	አጠቃላይ የኢኮኖሚ ዕድገት	15
41	ዲግ	Diplomatic Relation	ዲፕሎማሲያዊ ግንኙነት	16
42	ወተ	Military Mission	ወታደራዊ ተልዕኮ	16
43	ሃ	Nation Stability	የሃገር ድህንነትና ለዓላዊነት	16
44	ወታ	Military Training	ወታደራዊ ስልጠናና የማዕረግ እድገት	16
45	ሽብ	Terrorism	ሽብርተኝነት	16
46	ውግ	Foreign Conflicts	የውጭ ግጭቶችና ውይየቶች	16
47	ዜስ	Citizenship and Immigrants	ዜግነትና ስደተኞች	16
48	ዓለ	World wide and Continent Activities	ዓለም አቀፍና አህጉራዊ ክንዋኔዎች	16
49	በመ	Disease Protection	በሽታን መከላከል	3
50	ወባ	TB,HIV Aids	ወባ ቲቢና ኤች አይ ቪ ኤይድስ	3
51	ሌበ	Other Diseases	ሌሎች በሽታዎች	3
52	ባህ	Cultural Treatment	ባህላዊ ህክምና	3
53	ህእ	Children and Mothers Health	የህፃናትና እናቶች ጤና	3
54	ጤአ	Health Service	የጤና አገልግሎት	3
55	ጤተ	Health Institutions Construction	የጤና ተቋማት ግንባታ	3
56	ጤባ	Health Professionals	የጤና ባለሙያዎች	3
57	መአ	Medicines and Narcotic Drugs	መድሃኒቶችና አደገኛ ዕፆች	3
58	ህመ	Health Tools	የህክምና መሳሪያዎች	3
59	መህ	Kindergarten	መዋዕለ ህፃናት	4
60	አደ	Elementary	አንደኛ ደረጃ ትምህርት	4
61	ሁደ	Secondary School	ሁለተኛ ደረጃ ትምህርት	4
62	ቴክ	Technical	የቴክኒክና ሙያ ትምህርት	4
63	ከት	Higher Education	ከፍተኛ ትምህርት	4
64	ተት	Distance Learning	ተከታታይና የርቀት ትምህርት	4
65	መያ	Non Regular Education	መደበኛ ያልሆነ ትምህርት	4
66	ትሸ	Educational Coverage	የትምህርት ሽፋን	4
67	ነፃ	Free Education	የነፃ ትምህርት እድል	4
68	መጉ	Teachers and Students Affairs	የመምህራንና የተማሪዎች ጉዳይ	4
69	መዘ	Learning Methods	የትምህርት መገናኛ	4

			ዘዴዎች	
70	ትመ	Learning Tools	የትምህርት መሳሪያዎች	4
71	ትግ	Educational Institutions Construction	የትምህርት ተቋማት ግንባታ	4
72	ሴት	Women and Education	ሴቶችና ትምህርት	4
73	ሴጉ	Women Affair	ሴቶች ጉዳይ	17
74	ህወ	Youth and Young Affair	የህፃናት ወጣቶች ጉዳይ	17
75	አጉ	Physical Disabilities	የአካል ጉዳተኞች	17
76	አረ	Elders	አረጋዊያን	17
77	ሙህ	Professional and Social Associations	የሙያና ህዝባዊ ማህበራት	17
78	እድ	Edroch	እድሮች	17
79	ጋፍ	Wedding and Divorce	ጋብቻና ፍቺ	17
80	ልዕ	Birth and Death	ልደትና ዜና ዕረፍት	17
81	ሰእ	Humanitarian Donation	ሰብአዊ እርዳታ	17
82	አሰ	Employer and Employee	አሰሪና ሰራተኛ	17
83	ስአ	Unemployment	ስራ አጥነት	17
84	ሰያ	Sex Discipline	ስርአተ ያታ	17
85	ብብ	Nations and Nationalities Culture	የብሄር ብሄረሰቦችና ህዝቦች ባህል	18
86	ጎጂ	Bad culture	ጎጂ ልማዳዊ ድርጊቶች	18
87	ቅር	Antiques	ቅርሶች	18
88	ጎብ	Tourists	ጎብኚዎች	18
89	ቱል	Tourism Development	የቱሪዝም ልማት	18
90	ሃብ	Religious and National Holidays	ሃይማኖታዊና ብሄራዊ በዓላት	18
91	ኪነ	Art	ኪነጥበብ	18
92	ታሪ	History	ታሪክ	18
93	ሃጉ	Religious Assemblies and Reports	ሃይማኖታዊ ጉባኤዎችና መግለጫዎች	18
94	ፍብ	Justice Affairs	የፍትህ ብሄር ጉዳዮች	19
95	ወጉ	Crime Affairs	የወንጀል ጉዳዮች	19
96	ህጉ	Constitutional Affairs	ህገመንግስታዊ ጉዳዮች	19
97	ሙስ	Corruption	ሙስና	19
98	ዘር	Genocide	ዘር ማጥፋት	19
99	ፍአ	Judiciary Bodies	የፍትህ አካላት	19
100	ኢቴ	Information and Communication Technology	ኢንፎርሜሽንና ኮሙኒኬሽን ቴክኖሎጂ	20
101	መብ	Broadcasting Agencies	መገናኛ ብዙሃን	20
102	ምጥ	Research and Study	ምርምርና ጥናት	20
103	ፈስ	Creativity Work	የፈጠራ ስራዎች	20
104	የት	Weather Forecast	የአየር ትንበያ	21

105	በር	Desertification	በረሃማነትና የሙቀት መጨመር	21
106	ደን	Forestry	ደን ልማት	21
107	ዱር	Wild Animals Protection	የዱር እንስሳት ጥበቃ	21
108	አብ	Environmental Pollution	የአካባቢ ብክለት	21
109	አፅ	Environmental Greenery	የአካባቢ ፅዳትና ውበት	21
110	አት	Athletics	አትሌቲክስ	2
111	እኳ	Foot Ball	እግር ኳስ	2
112	ባስ	Traditional Sport	ባህላዊ ስፖርት	2
113	ሌዘ	Other Modern Sport Types	ሌሎች ዘመናዊ የስፖርት ዓይነቶች	2
114	ፌዴ	Federations and Clubs	ፌዴሬሽኖችና ክለቦች	2
115	ተአ	Natural Disaster	የተፈጥሮ አደጋ	22
117	ሰሰአ	Man-made Disaster	ሰው ሰራሽ አደጋ	22
118	አመ	Disaster Prevention	አደጋን መከላከል	22
121	ቦክስ	Box	ቦክስ	2
124	ጥቆማ	Information	ጥቆማ	No
125	አስቸ	Urgent	አስቸኳይ	124
126	የሚጠ	Expected	የሚጠበቅ	124
127	መደበ	Regular	መደበኛ	124

Source: SQL server of ENA

Appendix 7: List of Affixes Removed from a Word

Suffices	Prefixes
ም	ለ
ምና	ስለ
ና	በ
ን	በየ
ንም	እንደ
ንና	እንደየ
እና	እየ
ኡ	ከ
አች	ወደ
አችም	ወደየ
አችን	የ
አቻችን	
አቻችንም	
ዉ	
ዎቻቸው	
ዎቻቸዉን	
ዎቻቸዉንም	
ዎች	
ዎቻችን	
ዎችን	

Appendix 8: Amharic Various Characters with the Same Sound and their Translation to one Common Form

Characters with the Same Sound	Translated to
ሀ, ሃ, ሐ, ሓ, ኀ and ኃ	he
ሁ, ሑ and ኁ	hu
ሂ, ሒ and ኂ	hi
ሄ, ሔ and ኄ	hE
ህ, ሕ and ኅ	h
ሆ, ሖ and ኆ	ho
ሠ and ሰ	se
ሡ and ሱ	su
ሢ and ሲ	si
ሣ and ሳ	sa
ሤ and ሴ	sE
ሥ and ስ	s
ሦ and ሶ	so
ሸ and ሼ	xe
ሺ and ሽ	xi
ኘ and ኜ	Ne
ኚ and ኝ	Ni
አ, ኣ, ዐ and ዓ	e
ኡ and ዑ	u
ኢ and ዒ	i
ኤ and ዔ	E
እ and ዕ	I
ኦ and ዖ	o
ወ and ዎ	we
ዉ and ዑ	w
ዠ and ዤ	Ze
ዢ and ዣ	Zi
የ and ዬ	ye
ዩ and ይ	y
ጀ and ጄ	Je
ጂ and ጅ	ji
ጎ and ጐ	go
ፈጨ and ፈጭ	Ce
ፈጪ and ፊጮ	Ci
ጸ and ፀ	Se
ጹ and ፁ	Su
ጺ and ፂ	Si
ጻ and ፃ	Sa
ጼ and ፄ	SE
ጽ and ፅ	S
ጾ and ፆ	So

DECLARATION

THIS THESIS IS MY ORGINAL WORK AND HAS NOT BEEN SUBMITTED FOR DEGREE IN ANY OTHER UNIVERSITY

WORKU KELEMEWORK BIRHANIE

THE THESIS HAS BEEN SUBMITTED FOR EXAMINATION WITH OUR APPROVAL AS UNIVERSITY ADVISOR

MILLION MESHESHA (PhD)

LEMLEM HAGOS (M.I.Sc.)