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

PHRASE BASED AMHARIC NEWS TEXT CLASSIFICATION

BY

ZELEKE ABEBAW

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

JULY 2010

ACKNOWLEDGEMENT

First of all my blessings shall go to the almighty God, who gave me all the courage, endurance, and forgiveness.

I am deeply indebted to my advisor Ato Wondwosson M. (M.Sc.) for his skilful contribution to this thesis, on going encouragement and all-round support and guidance throughout my thesis work. His warm and kind hospitality together with his constructive criticism that I was rendered in his office gave me power and endurance in this study.

My deepest thanks goes to Dr. Million M. for his genuine cooperation, and sacrifices of time he made for on going discussions during my difficult time of selecting a title for the study. The study on the specified title would have been impossible without his assistance.

I am greatly thankful for Ethiopian News Agency (ENA) for letting me use its data, especially Ato Belsti, News Structuring and Organization Expert, for his valuable support during the data preprocessing stages of the research.

I would like to express my gratitude to my father Abebaw Kassa, sadly died in 1984 E.C, and my mother Fentanesh Kebede who offered me the opportunity of education they never had and who supported me, in one-way or another, throughout my schooling life. My deep gratitude goes to my Sisters Birhan A., Abez A. Yichalem A., Gojjam A, Mahlet M. and to my lost sister Silenat A. who gave me courage on my study in Addis Ababa University. I would also like to express my gratitude to my brother Aschalew T. for his fatherly advice through out my study. I am equally indebted to my best friend Esubalew Y (PhD Student), Addis Ababa University, for his brotherly advices and material supports like Laptop computer almost for the whole semester during my stay at the University.

Finally, I also would like to extend my gratitude to all organizations, friends and relatives who have contributed in many ways through out this study.

Table of Contents

ACKNOWLEDGEMENT……..i
LIST OF TABLES…….....ii
LIST OF FIGURES……..…iii
LIST OF GRAPHS……..…..iv
LIST OF APPENDICES……..…….…....iv
LIST OF ACRONYMS……..v
ABSTRACT…….....vi

CHAPTER ONE 11
INTRODUCTION..... 11
1.1 Background 11
1.2 Statement of the Problem and its Justification 13
1.3 Objectives of the Study..... 18
1.3.1 General Objective 18
1.3.2 Specific Objectives 19
1.4 Significance of the Study 19
1.5 Methodology..... 20
1.5.1 Literature Review 20
1.5.2 Data Source and Data Set Preparation..... 21
1.5.3 Tools and Software used..... 21
1.5.4 Implementation Procedure..... 22
1.6 Scope and Limitation of the Study 24
1.7 Organization of the Thesis 24

CHAPTER TWO 26
TEXT CLASSIFICATION 26
2.1 Introduction..... 26
2.2 Basic Concepts of Text Classification 26
2.3 Text Classification Approaches 28
2.3.1 Manual classification 28
2.3.2 Rule-based 28
2.3.3 Supervised Learning 29

2.3.4 Unsupervised Learning 29
2.4 Basic Concepts of Automatic Classification 30
2.5 Uses of Automatic Text Classification 31
2.5.1 Document Organization 31
2.5.2 Text Filtering 32
2.5.3 Hierarchical Categorization of Web Pages 33
2.6 Steps in Automatic Text Classification 33
2.6.1 Document Analysis 33
2.6.2 Document Representation 34
2.6.2.1 Phrase Based Document Representation in Text classification 35
2.6.3 Feature Selection 38
2.6.4 Term Weighting 41
2.6.5 Text Classifier Learning 42
2.6.6 Evaluation of the Performance of the Classifier 42
2.7 Machine Learning Approach (Support Vector Machine) 44
2.7.1 SVM for Classification 44
2.7.1.1 SVM Model 46
2.7.1.2 SVM kernels 47
2.7.2 Linear SVM: Linearly Separable Case 48
2.7.3 Non Linear SVM: Linearly inseparable Case 54

CHAPTER THREE 57
THE AMHARIC WRITING SYSTEM 57
3.1 Introduction 57
3.2 The Amharic Character Representation 58
3.3 Punctuation Marks 59
3.4 Numbers 59
3.5 Problems in Amharic Writing System 60
3.5 1. Characters (Fidels) with Different Form 60
3.5.2 Transliteration Problems 61
3.5.3 Abbreviations 61
3.6 Amharic Bag-of-Words vs. Bag-of-Phrases in Text Classification 61

CHAPTER FOUR 63
AUTOMATIC AMHARIC NEWS TEXT CLASSIFICATION 63
4.1 Introduction 63
4.2 The Data Source 63
4.3 Data Filtering 64
4.4 The Preprocessing Subsystem 66
4.5 Document Processing 68
4.5.1 Representing Documents by Relevant Phrases 68
4.5.1.1 Feature Extraction 69
4. 5.1.1.1 Character Normalization 69
4.5.1.1.2 Tokenization 70
4.5.1.1.3 Stop word Removal 72
4. 5.1.1.4 Stemming 73
4.5.1.2 Feature Selection 75
4.5.2 Data Transformation 77
4.5.3 Data Conversion 78
4.6 The Experiment 79
4.6.1 Model Building Using Bigrams and Testing Classification Accuracy 80
i. Experiment for the Four News Categories 81
ii. Experiment for Eight News Categories 84
iii. Experiment for Twelve News Categories 86
4.6.2 Model Building using Trigrams and Testing Classification Accuracy 91
i. Experiment with Four News Categories 91
ii. Experiment for Eight News Categories 93
iii. Experiment for Twelve Categories 95
4.6.3 Comparison of Bigrams and Trigrams in Amharic News Text Classification 99

CHAPTER FIVE 103
CONCLUSIONS AND RECOMANDATIONS 103
5.1 Conclusions 103
5.2 Recommendations 106
REFERENCE 107

LIST OF TABLES

Table 1.1 Summary of works done on Amharic Text classification

Table 1.2 Previous Researchers Work Performance Results at Increasing level of News Categories

Table 2.1 Document to Class Matrix

Table 2.2 Confusion Matrix for a Two Class Problem

Table 3.1 Seven forms of Amharic Characters

Table 3.2 Amharic Characters with Different Forms of the Same Sound

Table 4.1 News Documents Considered for the Experiment

Table 4.2 Algorithm for Character (Fidel) Normalization

Table 4.3 An Algorithm for Tokenization

Table 4.4 An Algorithm for Stop Word and Number Removal

Table 4.5 A stemmer Algorithm

Table 4.6 Example Prefix and Suffix Removed

Table 4.7 DF Threshold and Number of Features for Each Category

Table 4.8 Sample Experiment Data format

Table 4.9 Four news Categories with their Number of Instances

Table 4.10 Four Category LibSVM Detailed Accuracy by Class

Table 4.11 Four Categories LibSVM Confusion Matrix

Table 4.12 Eight News Categories with their Number of Instances

Table 4.13 Eight Category LibSVM Detailed Accuracy by Class

Table 4.14 Eight Categories LibSVM Confusion Matrix

Table 4.15 Twelve news categories with their number of instances

Table 4.16 Twelve categories LibSVM Detailed Accuracy by Class

Table 4.17 Twelve Category LibSVM Confusion Matrix

Table 4.18 Summary of Correctly and Incorrectly Classified Instances at Different Number of News Categories using Bigram Phrase Structures

Table 4.19 Four Category LibSVM Detailed Accuracy by Class Using Trigram Phrases

Table 4.20 Four Categories LibSVM Confusion Matrix using Trigram phrases

Table 4.21 Eight Category LibSVM detailed accuracy by Class using Trigram Phrases

Table 4.22 Eight Categories LibSVM Confusion Matrix using Trigram Phrases

Table 4.23 Twelve Categories LibSVM detailed accuracy by class using Trigram Phrases

Table 4.24 Twelve Category LibSVM Confusion Matrix using Trigram phrases

Table 4.25 Summary of Correctly and Incorrectly Classified Instances at Different Number of News Categories using Trigram Phrase Structures

Table 4.26 Summary of Experimental Results using Bigram and Trigram Phrase Structures at Different Category Levels

LIST OF FIGURES

Figure 1.1 Example of SVM Hyperplane PatternFigure 2.1 Possible Decision Boundaries for a Linearly Separable Data Set

Figure 2.2 Margin for a Decision Boundary

Figure.2.3 Text Classification Architecture

Figure 4.1 The General Description of the Data Preprocessing subsystem

LIST OF GRAPHS

Graph 4.1 Correctly and Incorrectly Classified Instances for the Four News Categories using Bigram Phrase Structures

Graph 4.2 Correctly and Incorrectly Classified Instances for Eight News Categories using Bigram Phrase Structures

Graph 4.3 Correctly and Incorrectly Classified Instances for Twelve Categories Using Bigram Phrase Structures

Graph 4.4 Correctly and Incorrectly Classified Instances with Four Categories Using Trigram Phrase Structures

Graph 4.5 Correctly and Incorrectly Classified Instances with Eight Categories using Trigram Phrase Structures

Graph 4.6 Correctly and Incorrectly Classified Instances for Twelve Categories using Trigram Phrase Structures

LIST OF APPENDICES

Appendix 1: Amharic Characters (‘Fidel’)

Appendix 2: Amharic Punctuation Marks Unicode

Appendix 3: Amharic Numbers

Appendix 4: List of Stop words

LIST OF ACRONYMS

CSV	Coma Separated Value
CV	Consonant-Vowel
DR	Dimensionality Reduction
ENA	Ethiopian News Agency
FE	Feature Extraction
FS	Feature Selection
GUI	Graphical User Interface
ICT	Information and Communication Technologies
IDF	Inverse Document Frequency
IR	Information Retrieval
KE	Knowledge Engineering
LibSVM	Library of Support Vector Machine
ML	Machine Learning
MMH	Maximum Marginal Hyperplane
NLP	Natural Language Processing
RBF	Radial Base Function
SMO	Sequential Minimal Optimization
SVM	Support Vector Machines
TC	Text Categorization
TF	Term Frequency
TFIDF	Term Frequency by Inverse Document Frequency
WEKA	Waikato Environment for Knowledge Analysis

ABSTRACT

The recent growth of Information and Communication Technologies (ICT) infrastructure in Ethiopia is resulting in an exponential increase of digital information in local languages including Amharic. Huge and increasing volumes of data are available in Amharic, which is observed on the growing online newspapers, websites, and digital storages of Ethiopian News Agency (ENA).Thus, to tackle the agency's news text management problems, a number of researches have conducted on automatic processing of Amharic news texts using *bag-of-words* feature representation approach.

However, using words as features could result in losing the intended meaning when the concept is created from two or more sequential words. Thus, in order to maintain this concept, a phrase based approach has been proposed and implemented in this research.

Preprocessing, feature representation, and testing were the major steps for the accomplishment of this study. Preprocessing the data (character normalization, stop word removal and stemming) is worked out before the datasets are fed into the classifier. In feature representations, two forms of phrase structures (bigrams and trigrams) have been developed and tested. After features have been represented by these phrase structures and their weights are identified using TFIDF schemes, phrase matrix have been generated and saved as CSV file format. The CSV files have been imported to the LibSVM classifier using the GUI of WEKA application package. Finally, the testing was performed for both bigram and trigram phrase structures at four, eight and twelve news category levels. From this research, using bigram phrase structures, the best accuracy (95.3%) has been obtained at four categories, followed by (81.3%) for eight categories and the least accuracy (72.01%) has been obtained at twelve categories. On the other hand, for trigram phrase structure, the best accuracy was obtained at four categories (72.9 %), followed by 69.7% for eight categories, and the least accuracy has been obtained at twelve categories that accounts to 56.4%. From these results, it can be observed that bigram phrase structures have better performance result (72.01%) than trigram phrase structures (56.4%)for all twelve news categories.

Keywords: Text categorization/classification, Machine Learning, Support Vector Machines, Phrase Based Feature Representations

CHAPTER ONE

INTRODUCTION

1.1 Background

The advancement and proliferation of information technology have fostered rapid creation and dissemination of information on a massive scale. As a result, voluminous information is becoming available at an explosive rate. Despite the growing popularity of multimedia, text remains the dominant form of information, via the Internet (Boley, 1999).

Availability of the vast amounts of textual documents demands appropriate document management solutions to support users search, access, and utilization of the ever-increasing corpora of textual documents. Analysis of prevalent practice suggests the common use of document category by individuals and organizations, thereby sorting documents into different categories. The sheer volume of new documents and the likelihood of their assignments to appropriate categories make manual document-category management approaches prohibitively tedious and ineffective. That means, the method of using domain experts to identify new text documents and allocate them to well-defined categories is time-consuming, subjective, expensive and error prone.

According to Sebastiani (2002), until the late 1980's the most popular approach to text classification at least in the "operational" (i.e. real-world applications) community was knowledge engineering (KE): an expert system, consisted in manually defining set of rules encoding expert knowledge on how to classify documents under the given categories. In the 1990's this approach has increasingly lost because of the popularity of the machine learning (ML) paradigm in the research community.

Machine learning is a general inductive process that automatically builds an automatic text classifier by learning the characteristics of the categories of interest from a set of pre-classified documents. In the areas of machine learning, extensive research has been done to test the possibility of automatic classification of documents. This approach is economically and qualitatively more effective than those achieved by manual classification systems. Moreover, the advantage of machine learning over the knowledge engineering is its effectiveness, with considerable savings in terms of expert labor since no intervention from either knowledge engineers or domain experts is needed for the construction of the classifier or for its porting to a different set of categories (Sebastiani, 2002).

Generally, the identification and categorization of text documents based on their contents are becoming imperative. Hence, automated document-category management represents an appealing alternative and can be greatly supported by appropriate text mining techniques. Of particular importance is text document classification, which partitions a collection of text documents into distinct groups where the documents in each group share great similarity and collectively reveal a specific theme covered in the underling document corpus (Boley, 1999).

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 (SVM), Genetic Algorithms, Hidden Markov Models and Classifier committees. In this study Support Vector Machine is considered.

SVM classification algorithms, proposed by Vapnik (1995) to solve two-class problems, are based on finding a separation between hyperplanes defined by classes of data, shown in Figure 1.1. This means that the SVM algorithm can operate even in fairly large feature sets as

the goal is to measure the margin of separation of the data rather than matches on features. Research has shown that SVM scales well and has good performance on large data sets.

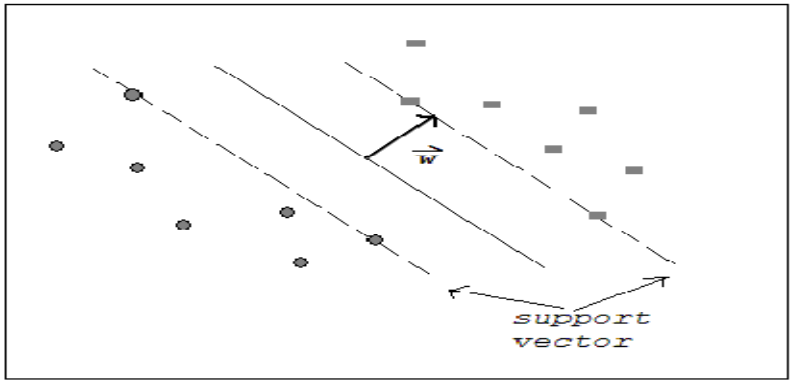


Figure 1.1 Example of SVM Hyperplane Pattern

Therefore, the purpose of this research is to test the effectiveness of Support Vector Machines using phrase-based feature representations in automatic Amharic news text classification.

1.2 Statement of the Problem and its Justification

Developments and application of text processing tools and methods have so far been focused on English and to some extent on European and East Asian languages (Baeza, 1999). However, the recent growth of Information and Communication Technologies (ICT) infrastructure in Ethiopia is resulting in an exponential increase of digital information being produced in local languages including Amharic.

Amharic is one of the major African languages. It is the official language of the Federal Government of Ethiopia and widely spoken throughout the nation. Amharic is an Afro-Asiatic language belonging to the southwest Semitic group with its own unique alphabet (FIDEL) (Atelach, 2005). Huge and increasing volumes of data are available in Amharic, which is observed on the growing online newspapers, websites, and digital storages of Ethiopian News Agency (ENA).

It may not be long before the Agency users are faced with large volumes of Amharic texts in the web and other media. In such scenarios, users will find it very difficult and frustrating to make efficient use of Amharic documents unless they are aided by data processing tools for activities like searching, categorizing and summarization of these documents. To assist this process the agency uses news management software called ENASoft. Though the agency uses ENASoft for its news management, the classification task is done manually. Currently, there are 12 major news categories and 98 sub categories. Using manual classification is challenging for these large number of categories.

Because of the observed non-uniform application of the manual classification system and its impact on efficient service provision, ENA has conducted two in-house studies to identify the problem areas of the system (ENA, 2006).

The major classification problems identified by the study are:

- Problems with the news categories

The first problem is lack of clear distinction between some news categories. For example, confusion is often observed in differentiating between the news items of the **Social** and the **Health** classes, the **Social** and the **Culture and Tourism** classes**,** and the **Politics** and the **Law and Justices** classes.

- Classification errors

There is no strict process to validate the categories given by reporters to the news items they are entering into the system. As a result, classification errors were found in news items that are not difficult to classify. For instance, the **Social** category contains many news items belonging to other classes.

- Entry errors

A look at the representative components of news items (the Headline and the Keywords) shows data entry errors.

Therefore, in order to tackle the above problems and to reduce the burden of human experts, more efforts have been undertaken by researchers and developers in the area of automatic Amharic news text classification. It is through research that one has to determine which techniques, tools, etc. are best for automatic Amharic news text classification. It was with this view that (Zelalem, 2001), (Surafel, 2003), (Yohannes, 2007) and (Worku, 2009) undergone their study using different machine learning approaches. The technique they used and the result they achieved is shown in Table 1.1 below:

Researcher	**No. News Categories Considered**	**Technique Used**	**Accuracy Achieved**
Zelalem Sintayehu (2001)	3	Cosine Similarity	85.05%
Surafel Teklu (2003)	16	Naiv Bayes (NB)	78.48%
		KNearest Neighbor (KNN)	64.4%
Yohannes Afework (2007)	15	Logic Model Tree (LMT)	79.72%
		Library of Support Vector Machine (LibSVM)	81.15%
Worku Kelemework (2009)	9	Artificial Neural Network (ANN)	70.08%

Table 1.1 Summary of works done on Amharic Text classification

Table 1.1 shows the results of automatic news classification done by four researchers using different machine learning approaches and the same document representation technique.

An important issue in text classification is how documents are represented, and how features can be extracted from them, which can be used for classification. A standard document representation is a vector of term occurrences, as used in the information retrieval (IR) field. This kind of frequent technique to represent documents is bag-of-words approach.

In the bag-of-words approach all words from the set of documents are taken and no ordering of words or any structure of text is used (Benbrahim, 2005). Each distinct word corresponds to a feature with a weight as its value that is correlated to the number of times the word occurs in the document. The main advantage of this approach is simplicity and easiness. Observing this simplicity and easiness, the above researchers, i.e. Zelalem (2001), Surafel (2003), Yohannes (2007), and Worku (2009), have implemented the bag-of-words feature representation approach for the development of Amharic news text classification system.

However, the main drawback of the bag-of-words feature representation is that it destructs the semantic relations between words (Benbrahim, 2005). A classical example is “Bill Gates”. Given a bag-of-words of a document in which words “Bill” and “Gates” occur, one can suggest that the document is about “accounting” or “gardening”, but not about computer software. Whereas given a document representation that contains a phrase “bill gates”, the reader will hardly be mistaken about the topic of discussion (Benbrahim, 2005). Moreover, single words are rarely specific enough to support accurate discrimination and their groupings are often accidental. For example, “Apple Computer”, “Apple” and “Computer”, word-based classification cannot distinguish the phrase, though their meanings are quite different. Furthermore, the bag-of-words representation approach increases the dimensionality of feature space which is known to have a negative effect for a classifier. For example, *distal descending colon* corresponds to one feature using bag-of-phrases approach and corresponds to three features in bag-of-words approach. Above all, information is lost due to feature splits.

The other major challenge in the development of automatic news classification system identified by the above researchers is that the decrease in accuracy as the number of news categories increases.

Researches by Surafel (2003) and Yohannes (2007) showed that accuracy decreases as the number of news categories increase as shown in Table 1.2.

Researcher	**News Categories Considered**	**Technique used**	**Accuracy Achieved**
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 1.2 Previous Researchers Work Performance Results at Increasing level of News Categories

As can be seen from Table 1.2 the classification accuracy sharply decreases as the number of news categories increases. This problem was observed as a challenge to implement automatic text classification systems using bag-of-words feature representation. Based on performance

analysis, therefore, Worku (2009) and Yohannes (2007) recommended a phrase-based approach or ontology as feature representations to improve classification accuracy.

Hence, to tackle the aforementioned two challenges faced by researchers (i.e. the problems of bag-of-words feature representations and the decrement in classification performance as news categories increases), this research implemented phrase-based feature representations.

Therefore, this research attempted to investigate the development of phrase-based Amharic news classification system using machine learning approach. To this end, it will try to answer the following research questions:

1. Can we improve classification accuracy using phrase-based feature representations?
2. Does phrase-based feature representation approach improve classification performance at increasing levels of news categories?
3. Does phrase-based feature representation approach solve the problem of bag-of-words (semantic lose of words)?
4. What is the best number of words in a phrase to achieve better accuracy in Amharic news text classification?
5. How to design an efficient phrase-based classification system for Amharic news texts?

1.3 Objectives of the Study

1.3.1 General Objective

The general objective of this research is to investigate the effectiveness of phrase-based feature representations to classify Amharic news texts into their predefined classes using machine learning approaches.

1.3.2 Specific Objectives

The following specific objectives would be met to achieve the aforementioned general objective.

- To review related works and identify techniques of automatic classification of documents.
- To study the Ethiopic writing system, grammatical rules, and its computer representation
- To develop pre-processing programs (codes) for Amharic documents (character controller, stop word and number remover, and stemmer.)
- To develop programs (codes) that generate bigram and trigram phrases from the news documents
- To select classification techniques to build an automatic Amharic news classification system
- To test a classifier using both bigram and trigram phrase structures
- To evaluate the performance of the classifier using representative datasets of Amharic news
- To provide concluding remarks and recommendations for future research directions

1.4 Significance of the Study

In addition to being an academic exercise to fulfill the requirement for the program (Master of Science in Information Science), this research is believed to produce results that can indicate the application of a general automatic Amharic news text classifier based on machine learning techniques. The results of the research can be used as an input for the development of a full-fledged automatic news classifier. The output of this thesis can also be used for further investigations of the development of automatic news text classification system for

those languages which use the Ethiopic scripts like Tigrinya, Guragignya and Geez. In addition, it will be useful as an input to any automatic Amharic text classification system with little modification.

1.5 Methodology

The following methods were employed to achieve the above stated objectives.

1.5.1 Literature Review

To get an understanding on various issues of automatic document classification systems, relevant materials such as books, documents, articles and journals were reviewed particularly on:

- the methods of effective information access techniques in general and automatic document classification in particular
- techniques like word-based and phrase-based feature representations
- basic characteristics of text classifier algorithms and their performances using different feature representations
- developed tools and techniques for Amharic text classification to study how they were developed
- language processing softwares particularly for Amharic like python
- Amharic language and its writing system

In order to further understand the problems of the news classification system implemented at ENA, the current classification scheme and the software (ENASoft) they are using were explored. In addition, interviews and discussions were conducted with appropriate staff of the agency.

1.5.2 Data Source and Data Set Preparation

The data source for the study is nearly a five year's collection of Amharic news articles from the Ethiopian News Agency (ENA), which has more than 16,000 news items in "html" format. One of the reasons for the selection of ENA's news database as the data source is the availability of large collection of news items most of which are labeled through an established manual categorization scheme.

For data preparation, since the Amharic documents were collected in their pre-defined categories in "html" format, the whole data was changed to "text" file formats by saving the "html" files as ".txt" file format for further pre-processing.

The source data were cleaned using manual inspection and automatic methods. Automated data cleaning and other pre-processing tasks were carried out using language processing software. In pre-processing the news documents, it is assumed that a document is fully represented by the phrases in it. The data pre-processing carried out in this research involved:

- data cleaning which involves removal of repeated news items, removal of entry errors, etc.
- normalizing different letters of the Amharic script that have the same sound
- identifying and removing stop words
- removing prefix and suffixes of words
- creating bigram and trigram phrase structures from the stemmed words
- selection of relevant attributes (features) of news documents

1.5.3 Tools and Software used

In this study the WEKA (Waikato Environment for Knowledge Analysis) application package is used as a tool to classify Amharic news documents. WEKA is an open source data mining software developed at the University of WAIKATO in New Zealand. The whole

package is written in Java, so it can be run on any platform. The package offers three different interfaces.

- A command line interface
- An Explorer GUI (Graphical User Interface) interface: which allows different types of data preparation, and modeling algorithms on a dataset
- An Experimenter GUI interface: which allow running different algorithms in batch and comparing the results

WEKA is used in this research because the WEKA package provides several classifiers for automatic classification of preprocessed datasets. The one which is implemented in this research is the Library of Support Vector Machine (LibSVM). Moreover, it is more convenient and the researcher's familiarity with the tool and its additional facility like data preparation using the Explorer GUI interface. Hence, particularly the Explorer GUI is used for the experimental processes.

Regarding the software, Python is used to process Amharic texts to prepare representative phrases. This software is selected because it is robust to handle and process Amharic characters (Fidels) and its familiarity to the researcher. Therefore, all those text format of Amharic news documents which were converted from the "html" to "text" file format have been processed using this software.

1.5.4 Implementation Procedure

Automatic document classifications often utilize different methods of feature representations like bag-of-words and bag-of-phrases. For this research bag-of-phrases (bigrams and trigrams) were used to represent features.

Automatic classification involves identification of keyphrases called feature phrases or attributes for representing documents. Huge text datasets can have millions of phrases

representing the document collection in the dataset. However, not all phrases in the dataset are useful for automatic classification. Classification systems use stop list to avoid stop words from document attributes. Stop words - like **ስለ፣ እና፣ እንደ፣ወደ ፣ አስታውቀ፣** and **ገለፀ** are considered irrelevant for classification. Moreover, some attributes may be irrelevant for a given classification task. For example, if the task is to classify news items to the major news categories, attributes such as the news creation date are likely to be irrelevant.

After preprocessing, the experimental dataset is rearranged to the format suitable for the WEKA package that is used for the automatic classification. WEKA requires the input data to be rearranged in an attribute weight matrix before it is converted to CSV file format.

The preprocessed dataset is rearranged in an attribute weight matrix by:

- ✓ considering the whole dataset as a relation with phrase attributes
- ✓ considering three labels such as ‘PhraseFeatures’, ‘TFIDF Values’, and ‘NewsCategory’ in the dataset as fields (column) of the relation
- ✓ taking each document instances as a separate record or instance (row) of the relation
- ✓ using the weight of attributes in a document as the value of the fields for the instance the document represents
- ✓ considering class labels as nominal attributes in the dataset

After the arrangement, the experimental dataset is converted to CSV file format, which is suitable for applying classifier algorithms provided by WEKA i.e. Library of Support Vector Machine (LibSVM).

1.6 Scope and Limitation of the Study

The research focused for Amharic news items using phrase-based (bigram[1] and trigram[2]) feature representations. Moreover, not all components of the news items are taken as news documents for the experiment. Observation of the structure of ENA's news items shows that a Headline contains a concise summary of a news item and one can get the essence of the whole news from the Headline. In addition, the other news components, the Keywords, provide representative words. Thus, only these major elements, i.e., the Headline and Keywords are taken as representatives of news items.

Regarding the news categories, twelve Amharic news categories were considered. These are Accident, Culture and Tourism, Economy, Education, Politics, Social, Defense, Weather, Sport, Law and Justice, Science and Technology, and Health. All of these news categories have been used for the experiment in three groups. The first group contains four news categories, the second group contains eight news categories, and the third group contains twelve news categories.

1.7 Organization of the Thesis

This thesis is organized into five chapters. The first chapter is introductory, in which the background of the research is described. This chapter also presents statement of the problem, objective of the study and the methods followed.

The second part of the thesis is chapter two. In chapter two the techniques available in the area of automatic classification are reviewed, and the techniques followed in this research are described in detail.

[1] Two words
[2] Three words

Sine the research is done for Amharic news documents, chapter three reviewed language aspects that should be considered in the development of text analysis.

Chapter four discusses the experimental part of the research, which is the main concern of the research. The results obtained from the experiment are presented in this chapter.

Finally, the conclusions drawn from the study and the recommendations forwarded for future work are presented in chapter five.

CHAPTER TWO

TEXT CLASSIFICATION

2.1 Introduction

We live in the world where information has a great value and the amount of available information, mostly on the internet, has been exponentially growing. There is so much information around us, that it becomes a problem to find those that are relevant for us. Factors that contribute to the accelerated growth of information include the computerization of businesses, scientific and government transactions as well as advances in data collection tools ranging from written texts to satellite remote sensing systems. In addition, popular use of the World Wide Web as a global information system has made access to huge amount of data and information possible (Han and Kamber, 2006). Because of these, there are many databases and catalogues of information divided into categories, helping the user to navigate to the information one would like to obtain. Most of theses information are texts and here the text classification comes to the scene.

2.2 Basic Concepts of Text Classification

The concept of text classification is defined by number of authors in similar ways. For example, Blumberg and Atre (2003); Giorgino (2004); Klein (2004); Liao, Alpha and Dixon (2003); Sebastiani (2005); Skarmeta, Bensaid and Tazi (2000) and Wang and et. el. (2005) defined text classification as the task of automatically assigning a set of documents into 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 classification as a mapping of text documents to classes. To clarify, if C= {c_1, c_2…cm} is a set of classes and D= {d_1, d_2,…,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.

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

Table 2.1 Document to Class Matrix

In Table 2.1 $d_1 \ldots d_n$ refers set of documents, $c_1 \ldots c_m$ refers set of classes 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**.

Therefore, text classification is the process of matching document representatives with class representatives based on the similarity that exist between documents and classes.

According to 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 class, or even under neither, depending on the subjective judgment of the expert.

2.3 Text Classification Approaches

Depending on the context of their application, there are four approaches to text classification. These are Manual, Rule-based, Supervised Learning and Unsupervised Learning (automatic classification.)

2.3.1 Manual classification

Manual classification is often used in library and technical collections as well as in call centers and forms-processing environments. Manual classification requires individuals to assign each document to one or more categories. These individuals are usually domain experts who are thoroughly versed in the category structure or taxonomy being used. Manual classification can achieve a higher degree of accuracy –although even domain experts will occasionally disagree on how to categorize a document (Han and Kamber, 2006). However, manual classification is more labor-intensive and therefore more costly than automatic techniques.

2.3.2 Rule-based

In this form of classification, keywords or Boolean expressions are used to categorize a document. This is typically used when a few words can adequately describe a category. For example, if a collection of medical papers is to be classified according to a disease, then a medical thesaurus that lists each disease together with its scientific, common and alternative names can be used to define the keywords for each category.

While rule-based systems are effective for carefully tuning a limited number of categories, the expense of defining and maintaining categories is generally prohibitive for large scale classification systems (Blumberg and Atre, 2003).

2.3.3 Supervised Learning

Most approaches to automatic text classification require a human subject expert to initiate the learning process by manually classifying or assigning a number of “training documents” to each category. This classification system first analysis the statistical occurrences of each concept in the example documents and then constructs a model or “classifier” for each category that is used to classify subsequent documents automatically. The system refines its model, in a sense “learning” the categories as new documents are processed (Blumberg and Atre, 2003).

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, may not found categories which are intuitive to humans (Blumberg and Atre, 2003)

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).

Having discussed the different approaches, let us now justify one of the approaches which is best to text classification.

Text classification poses many challenges for indicative learning methods since there can be millions of word features. Automatic classifiers, however, have many advantages because:

- they are easy to construct and update

- they depend only on information that is easy for people to provide, information like examples of items that are in or out of categories
- they can be customized to specific categories of interest to individuals
- they allow users to smoothly trade-off precision and recall depending on their task

Hence, due to the above reasons and their remarkable results, a growing number of statistical classification and machine learning approaches have been applied to text classification including Support Vector Machines (SVM) which is discussed in section 2.7.

In this thesis, therefore, automatic text classification (supervised learning approach) supported by limited Natural Language Processing (NLP) technique is used for the Amharic news text classification. As a result, the basic concept of automatic text classification is discussed briefly in the following section.

2.4 Basic Concepts of Automatic Classification

It is obvious that the goal of any classification process is to group similar documents together. This implies, the procedure of classifying documents into groups require a quantitative measure of the "likeness" of the document for a given class, and the separation of unlike ones. In other words, it involves measuring similarity of a document with different classes.

Whenever we talk of similarity of items, we are talking of their similarity usually with their attributes. The attributes can be anything that characterizes the bases of the classes for the purpose of the classification process. Therefore, the first thing in the process of automatic classification is to identify the attributes of documents and classes so that the matching of the two becomes simple or at least possible.

In short, automatic classification is the process of matching document representatives with class representatives to automatically assign classification codes to documents based on the similarity that exist between documents and classes. The codes assigned to a document should be the code of a class that has the maximum similarity value with the given document. The following section discusses the uses of automatic classification.

2.5 Uses of Automatic Text Classification

Text Classification has been used for a number of different applications: automatic indexing for Boolean information retrieval systems, document organization, text filtering, hierarchical categorization of web pages, prepositional phrase attachment and word choice selection in machine translation and data analysis/text data mining (Sebstiani, 2002).

2.5.1 Document Organization

Indexing with a controlled vocabulary is an instance of the general problem of document base organization. In general, many other issues pertaining to document organization and filing, be it for purposes of personal organization or structuring of a corporate document base, may be addressed by Text Categorization (TC) techniques. For instance, at the offices of a newspaper incoming “classified” ads must be, prior to publication, categorized under categories such as Personals, politics, economy, etc. Newspapers dealing with a high volume of classified ads would benefit from an automatic system that chooses the most suitable category for a given ads. Other possible applications are the organization of patents into categories for making their search easier, the automatic filing of newspaper or news stories under the appropriate sections (e.g., Politics, Home News, Lifestyles, etc.), or the automatic grouping of conference papers into sessions or case summaries may be put based on a sort of case classification (Sebastiani, 2002). Topic spotting for newswire stories is one of the most commonly investigated applications domains of TC (Yang, 1999).

2.5.2 Text Filtering

Text filtering is the activity of classifying a stream of incoming documents dispatched in an asynchronous way by an information producer to an information consumer (Sebstiani, 2002). A typical case is a news feed, where the producer is a news agency and the consumer is a newspaper reader (Sebastiani, 2002). In this case, the filtering system should block the delivery of the documents the consumer is likely not interested in (e.g., all news not concerning sports, in the case of a sports newspaper). Filtering can be seen as a case of single-label TC, that is, the classification of incoming documents into two disjoint categories, the relevant and the irrelevant. Additionally, a filtering system may also further classify the documents deemed relevant to the consumer into thematic categories; in the example above, all articles about sports should be further classified according to which sport they deal with, so as to allow journalists specialized in individual sports to access only documents of prospective interest for them.

Similarly, an e-mail filter might be trained to discard “junk” mail and further classify non junk mail into topical categories of interest to the user. A filtering system may be installed at the producer end, in which case it must route the documents to the interested consumers only, or at the consumer end, in which case it must block the delivery of documents deemed uninteresting to the consumer. In the former case, the system builds and updates a profile for each consumer, while in the latter case, which is the most common, a single profile is needed. A profile may be initially specified by the user, thereby resembling a standing Information Retrieval (IR) query, and is updated by the system by using feedback information provided (either implicitly or explicitly) by the user on the relevance or non relevance of the delivered messages (Androutsopoulos et al., 2000).

2.5.3 Hierarchical Categorization of Web Pages

Text categorization is crucial to find interesting information on the World Wide Web, and to guide users search through hypertext (Joachim, 1996). Hierarchical categorization of web pages is decomposing the classification problem into a number of smaller classification problems, each corresponding to a branching decision at an internal node. TC has recently aroused a lot of interest also for its possible application to automatically classifying Web pages, or sites, under the hierarchical catalogues hosted by popular Internet portals. When Web documents are catalogued in this way, rather than issuing a query to a general purpose Web search engine a searcher may find it easier to first navigate in the hierarchy of categories and then restrict the search to a particular category of interest. Classifying Web pages automatically has obvious advantages, since the manual categorization of a large enough subset of the Web is infeasible.

2.6 Steps in Automatic Text Classification

2.6.1 Document Analysis

Document analysis is the process of analyzing the text of a document to find document representatives out of it. As discussed by Cheng and Wu (1995) classification requires document analysis, which is heavily dependent on the representation of the document. In fact, document analysis is very important especially when there are a huge number of electronic documents. The reason is manipulating this huge collection for whatever purpose will be very difficult in terms of storage space and processing time.

Generally speaking, document analysis is the analysis of documents to find efficient document representatives for the purpose of storage and retrieval, which is also called indexing. The purpose of indexing is to obtain a number of descriptors, which act as surrogates for the document. This means, given a written text in natural language, it is

essential to represent the information contained in the text by one or more entries, variously known as indexes, keywords, key terms. In fact, the classic models in IR consider that each document is described by a set of representative index terms. These descriptors, or keywords, can be obtained manually or automatically by computer analysis of the document file, abstract or text. The problem is to choose "good" terms, which collectively reflect the information content as accurately as possible.

The representation can be by analyzing the whole document or only part of the document. For instance, Losee and Haas (1995) have tried to use the titles of the documents for representation purpose while Kwok (1975) used title and cited titles to represent documents. On the other hand, Enser (1985) used back of book indexes to classify books. In fact, as discussed by Enser (1985) documents can be suitably classified by examining only limited portions of the document, which can save considerable time and money. Therefore, the first step in document analysis is deciding on the representation of documents.

2.6.2 Document Representation

From the beginning, documents can be considered as a stream of characters. However, for the problem of automatic classification these streams should be transformed into representatives like single words or sequence of words (phrases) which are suitable for the process of classification.

Due to the drawback of the bag-of-words feature representation which destructs the semantic relations between words (refer to section 1.2), in this research a phrase based document representation is used. Because, as discussed earlier using phrases[3] as features incorporates word sequence information.

[3] a phrase is a textual unit usually larger than a word but lesser than a full sentence

2.6.2.1 Phrase Based Document Representation in Text classification

Observing those limitations of bag-of-words feature representations, IR researchers have expressed their dissatisfaction on this approach, and have tried to use notions of features that are at the same time semantically richer and technically feasible. In particular, a number of authors Mladenic and Grobelnik(1998), Fürnkranz (1998), Schütze (1995), Schapire (1998), Lewis (1992), Dumais(1998), Scott and Matwin (1999), Zhixu Li1 et al(2009), Crawford (2006),etc. have investigated phrase based features.

There are two kinds of phrases. These are syntactical and statistical phrases. Syntactic phrases denote any phrase that is such according to a grammar of the language under consideration whereas statistical phrase is any sequence of words that occur contiguously in a text. According to Caropreso, Matwin and Sebastiani (2001) using syntactic phrases in indexing seems an interesting idea, in that:

- syntactic phrases come closer than individual words or their stems to expressing structured concepts;
- syntactic phrases have a smaller degree of ambiguity than their constituent words
- By using syntactic phrases as index terms, a document that contains a phrase that occurs in the request would be ranked higher than a document that just contains its constituent words in unrelated contexts;

Unfortunately, a number of researches that have investigated the usefulness of indexing with syntactic phrases in IR have obtained discouraging results. The likely reason for this is that, although indexing languages based on syntactic phrases have superior semantic qualities, they have inferior statistical qualities with respect to indexing languages based on single words Fagan (1987) and Lewis (1992).

For instance, the phrase *nuclear waste disposal* definitely denotes an interesting, articulated concept, but unless it occurs frequently enough in the document collection it is unlikely to make an impact in terms of effectiveness. This situation is worsened by the fact that the same concept may be triggered by related but linguistically different units (such as *disposing of nuclear waste*, *Dispose of your nuclear waste*, etc.), each of which is usually considered, from the standpoint of frequency, a different unit, unless the similarity of the underlying concepts is recognized.

Also, not every syntactic phrase denotes an interesting concept: "*associate professor"* does, but "*tall professor"* does not, and telling a phrase that does from one that does not is difficult.

Lewis (1992) has been the first to study the effects of syntactic phrase indexing in a text classification context. He reported that, in the context of a Na¨ıve Bayes classifier, this yields significantly lower effectiveness.

Dumais et al. (1998) reported no benefit at all from the use of syntactic phrases with a variety of text classifiers in the context of Reuters-21578 experimentation.

As a result, some researchers have attempted to find a way out of these problems by understanding the notion of phrases in a statistical sense, rather than syntactically, and do so in a statistically interesting way. According to Caropreso et.al (2001) statistical phrases have a number of advantages over syntactic ones:

- they may be recognized by means of more robust and less computationally demanding algorithms;
- the effect of irrelevant syntactic variants can be factored out; and
- uninteresting phrases (e.g. *tall professor*) tend to be filtered out from interesting ones (e.g. *associate professor*).

Accordingly, Mladenic and Grobelnik (1998) have used statistical phrases by extracting n-grams[4] of length up to 5 by means of a fast algorithm that relies on document frequency as a statistical filter. On a Na¨ıve Bayes classifier applied to a corpus of Web pages, they have found that n-grams of length up to 4 give significant benefits with respect to the single words case, while 5-grams do not provide additional benefit.

Similarly, Fürnkranz (1998) uses an algorithm similar to that of Mladenic and Grobelnik (1998) to extract n-grams of length up to 5 to test the performance of statistical phrases on different classifier called Ripper[5]. On Reuters-21578[6] he has found that Ripper has a significant improvement in performance when n-grams of length up to 2 were used, but that longer n-grams reduce classification performance; on another dataset of Usenet[7] newsgroup articles he instead found also 3-grams to have some utility, whereas the negative contribution of larger n-grams was confirmed.

Moreover, statistical phrase-based approach has been tried by other researchers for some Asian languages like China, and significant improvements have been achieved using Support Vector Machine (SVM) (Zhixu Li1et al, 2009).

SVMs are a set of related supervised learning methods used for classification (see section 2.7 for detail). Therefore, this research deals with assessing the values of statistical phrases for document indexing in the context of text categorization for Amharic news texts; the activity of inductively learning to classify natural language texts with topical categories from a pre-specified set.

---

4 n-grams are sequence of words(where ***n*** stands for two, three , four etc consecutive words)

5 Text classification algorithm used by the researcher

6 News repositories of the Reuter in which the total number of the documents are 21,578 used by researchers for document classification

7 Collection of news data used by researchers for text classification experiment

In general, when we talk of statistical phrase[8]based document representation (bigrams and trigrams phrases), a document that describes a certain concept is more likely to have phrases from that domain. In other words, a document about "*Economy*" will have a bigram phrase "*market price*" or its synonyms in it. A document that contains the bigram phrases "*teaching learning*", "*school director*", "*student center*", "*ministry education*", "*reading assignment*" etc. more likely deals with the "*Education*" news category. Similarly, the trigram phrases such as "*market price analysis", "oil demand supply" etc* are more likely to deal with *"Economy"* news classes. Similarly, documents that contain trigram phrases like "*teaching learning process*", "*Higher Education Institution*" etc are likely to be under *"Education"* news categories.

The problem, however, is that all of the *phrases (bigrams or trigrams)* in the documents cannot be considered as features of documents. Rather only the good descriptors should be taken to minimize the problem of storage and computation time. In fact, in the field of text classification, it has been seen that maximum performance is often not achieved by using all available features, rather by using only a "good" subset of those (Joachim, 1996). The problem of finding these "good" subset features is called feature selection.

2.6.3 Feature Selection

Feature selection is often an essential step in text classification as text collections can have more than 100,000 unique terms (words or phrases). Removing less informative and noisy terms reduces the computational cost and often improves classifier generalization.

In order to select features a variety of ranking criteria have been used in text classification with varying degrees of success. Statistical techniques to feature selection are widely used in the area of automatic text classification. Statistical techniques to text analysis and feature

8 Hereafter phrase mean statistical phrase

selection are based on term frequency[9]. The pioneering work of Luhn (1958) showed statistical analysis of the words or phrases in a document will provide some clues as to its content. The idea behind this principle is, a term (phrase) that is frequently present in a document is useful to represent the document. The problem, however, is how frequent should a term (phrase) be found in a document to be accepted as an index term.

One technique proposed by Luhn (1958) says both high-frequency words and rare-frequency words are unlikely to be able to represent documents and should not be considered (Cheng and Wu, 1995). The former are discarded because they occur too often to indicate the subject matter and the later because they are too rare. Therefore, the remaining intermediate frequency terms are assumed important.

It has also been proved that, grammatical function words such as "the", "and", "of" and "to" exhibit approximately equal frequencies of occurrence in all the document collection and should not be considered in the selection process. The indexing technique, therefore, should consider frequency of the content bearing words only. The frequency of occurrence of non function (terms that relate to the subject of a given document) words may actually be used to indicate term importance for content representation.

Salton (1989) suggested an index extraction method that can be done using the following three steps:-

1. Eliminate common function words from the document by consulting a special dictionary, called negative dictionary or stop word list, which contains a list of high frequency words.

[9] term frequency is the number of times a term occurs in a given document.

2. Compute the term frequency tf_{ij} for all remaining terms t_j in each document d_i, specifying the number of occurrences of t_j in d_i.

3. Choose a threshold frequency T, and assign to each document d_i all terms t_j for which $tf_{ij} > T$.

The main task in the above procedure is to find the threshold frequency value that actually affects significantly the result of the classification. A slightly low threshold will lead to close classification, while a high threshold will lead to a broad classification[10] (Cheng and Wu, 1995).

In fact, a high frequency term is acceptable for indexing purpose only if its occurrence frequency is not equally high in all the documents of a collection. That is, it must have a discrimination power between documents (classes). This implies, some terms (phrases) have high discrimination power than others.

Salton (1989) suggests a formula to evaluate a term discriminating power by an inverse function. That is, in a collection of N documents, if a term t_j occurs in df_i documents, then the discrimination value of the term is N/df_i .

In other words, even after we identified the index terms, not all the terms are equally important for classification. Some are more important than others. The technique of assigning importance value for terms is term weighting. In IR weighting is done to rank results for a given query. Term weighting assigns indications of importance to terms. As a basic step in automatic classification, the following section reviews the term weighting techniques.

10 Close classification is classifying each subject as completely or as fully as possible, and Broad classification is classifying the material only in main divisions and subdivisions without using the minute breakdown of individual categories.

2.6.4 Term Weighting

Term weighting assigns values to terms (phrases) that indicate their level of importance. The importance of an index term (phrase) to a document is shown by using weight (Giorgino, 2004). 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 or phrase (Baeza-Yates and Ribeiro-Neto, 1999).

$$Term\ Weighting = \frac{Frqeuency\ in\ the\ document}{Frequency\ in\ all\ documents}$$

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

$TF = FREQ_{ik}$……. *Equation 2.1 Term Frequency*

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

$IDF = log_2^{\frac{N}{dk}}$ …… *Equation 2.2 Inverse Document Frequency*

In *Equation 2.2*, *N* is the total number of documents in the collection, *dk* 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.

Equation 2.3 shows $TF * IDF$.

$$TF * IDF = FREQ_{ik} * log_2^{\frac{N}{dk}} \quad \text{…..……. Equation 2.3 } TF * IDF \text{ weight of term } k$$

In *Equation 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, dk the number of documents in which term k occurs.

2.6.5 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).

2.6.6 Evaluation of the Performance of the Classifier

According to (Tan.et.al, 2006) evaluation of the performance of a classification model is based on the counts of test records correctly and incorrectly predicted by the model. These counts are tabulated in a table known as a confusion matrix. Table 2.2 shows a confusion matrix for a binary classification problem. Each entry f_{ij} in this table denotes the number of records from class i predicted to be of class j. For instance, f_{01} is the number of records from $class\ 0$ incorrectly predicted as $class\ 1$. Based on the entries in the confusion matrix, the total number of correct prediction made by the model is $(f11 + f00)$ and the total number of incorrect prediction is $(f10 + f01)$.

		Predicted class	
		$class = 1$	$class = 0$
Actual class	$class = 1$	f_{11}	f_{10}
	$class = 0$	f_{01}	f_{00}

Table 2.2 Confusion Matrix for a Two Class Problem

Although a confusion matrix provides the information needed to determine how well a classification model performs, summarizing this information with a single number would make it more convenient to compare the performance of different models. This can be done using a performance metric such as accuracy, which is defined as follows.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ Number\ of\ predictions} = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}$$ *...Formula 2.4*

Equivalently, the performance of a model can be expressed in terms of its error rate, which is given by the following equation:

$$Error\ rate = \frac{Number\ of\ \text{wrong}\ predictions}{Total\ Number\ of\ predictions} = \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}}$$ *…Formula 2.5*

Most classification algorithms seek models that attain the highest accuracy, or equivalently, the lowest error rate when applied to the test set. In this research, accuracy is used for evaluation of Amharic news text classifier.

2.7 Machine Learning Approach (Support Vector Machine)

Machine Learning is considered as a subfield of Artificial Intelligence and it is concerned with the development of techniques and methods which enable the computer to learn. In simple terms, it is the development of algorithms which enable the machine to learn and perform tasks and activities. Over the period of time many techniques and methodologies were developed for machine learning tasks.

Support Vector Machine (SVM) is one of the machine learning techniques. It was first heard in 1992, introduced by Boser, Guyon, and Vapnik in COLT-92. SVMs are a set of related supervised learning methods used for classification and regression (P.Lewis, 2004). They belong to a family of generalized linear classifiers. SVM is a classification and regression prediction technique that uses machine learning theory to maximize predictive accuracy while automatically avoids overfitting (i.e. to avoid incorporating particular characteristics of the training data that do not represent the whole dataset) to the data. Support Vector Machines can be defined as systems which use hypothesis space of a linear functions in a high dimensional feature space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory. SVM is being used for many applications, such as hand writing analysis, face recognition, text classification and so forth.

2.7.1 SVM for Classification

SVM is a useful technique for data classification. A classification task usually involves with training and testing data which consist of some data instances (Duda R, 1973). Each instance in the training set contains one target value and several attributes. The goal of SVM is to

produce a model which predicts target value of data instances in the testing set where given only the attribute values (Nello, 2000).

Classification in SVM is an example of Supervised Learning. Known labels help indicate whether the system is performing in a right way or not. This information points to a desired response, validating the accuracy of the system, or be used to help the system learn to act correctly. A step in SVM classification involves identification of features as which are intimately connected to the known classes. This is called feature selection. Feature selection and SVM classification together have a use even when prediction of unknown samples is not necessary. They can be used to identify key sets which are involved in whatever processes distinguish the classes (Nello, 2000).

Other classification approaches perform poorly when working directly because of the high dimensionality of the data, but Support Vector Machines can avoid the pitfalls of very high dimensional representations (P.Lewis, 2004). One remarkable property of SVMs is that their ability to learn is independent of the dimensionality of the feature space. SVMs measure the complexity of hypotheses based on the margin with which they separate the data, not the whole number of features. This means that we can generalize even in the presence of very large features, if our data is separable with a wide margin using functions from the hypothesis space. The following two arguments also explain the reason in using SVMs for Text classification.

- **High dimensional input space**: When learning text classifiers, one has to deal with very many (more than 10000) features. Since SVMs use overfitting protection, which does not necessarily depend on the number of features, they have the potential to handle these large feature spaces (Joachim, 1998).

- **Most text classification problems are linearly separable**: All Ohsumed[11] categories are linearly separable and so are many of the Reuters tasks. The idea of SVMs is to find such linear (or polynomial, RBF-Radial Base Function, etc.) separators (see 2.7.11 below).

These arguments give theoretical evidence that SVMs should perform well for text categorization (Joachim, 1998).

2.7.1.1 SVM Model

SVM models for classification can be divided into two distinct groups based on the error function:

- Classification SVM Type 1 (also known as C-SVM classification)
- Classification SVM Type 2 (also known as nu-SVM classification)

- **Classification SVM Type 1(C-SVM classification)**

For this type of SVM, training involves the minimization of the error function:

$$\frac{1}{2}w^T w + C\sum_{i=1}^{N}\xi_i$$

subject to the constraints:

$$y_i\left(w^T\phi(x_i)+b\right)\geq 1-\xi_i \text{ and } \xi_i \geq 0,\ i=1,\ldots,N$$

where C is the capacity constant, w is the vector of coefficients, b a constant and  are parameters for handling nonseparable data (inputs). The index i labels the N training cases. Note that $y \in \pm 1$ is the class labels and xi is the independent variables. The kernel ϕ is used to transform data from the input (independent) to the feature space. It should be noted that the larger the C, the more the error is penalized.

[11] clinically-oriented datasets

- **Classification SVM Type 2 (nu-SVM classification)**

In contrast to Classification SVM Type 1, the Classification SVM Type 2 model minimizes the error function:

$$\frac{1}{2}w^T w - \nu\rho + \frac{1}{N}\sum_{i=1}^{N}\xi_i$$

subject to the constraints:

$$y_i\left(w^T\phi(x_i)+b\right) \geq \rho - \xi_i,\ \xi_i \geq 0,\ \ i = 1,\ldots,N \text{ and } \rho \geq 0$$

Compared to regular C-SVM, the formulation of nu-SVM is more complicated, so up to now there have been no effective methods for solving large-scale nu-SVM (Joachims, 1998).Thus, C-SVM classification is used in this research.

2.7.1.2 SVM kernels

Kernels are classes of algorithms whose task is to detect and exploit complex patterns in data (e.g.by clustering, classifying, ranking, cleaning, etc. the data). Typical problems are: how to represent complex patterns; and how to exclude spurious (unstable) patterns (over fitting). The first is a computational problem; the second is a statistical problem.

The class of kernel methods implicitly defines the class of possible patterns by introducing a notion of similarity between data, for example, similarity between documents by length, topic, language, etc. Kernel methods exploit information about the inner products between data items. Many standard algorithms can be rewritten so that they only require inner products between data (inputs). When a kernel is given there is no need to specify what features of the data are being used.

In SVM there are four common kernels:

- linear kernel K(xi,xj) = $x^{T}_{i}\, x_j$
- Polynomial kernel of degree h: $k(X_i, X_j) = (X_{i,..}\, X_j + 1)^h$
- Radial Base Function (RBF) kernel: $k(X_i, X_j) = \exp(-\gamma \| X_i - X_j \|^2)$
- Sigmoid kernel: $k(X_i, X_j) = \boldsymbol{tanh}(k X_i \,.\, X_j - \delta)$

In general RBF is a reasonable first choice because the kernel matrix using sigmoid may not be positive definite and in general it's accuracy is not better than RBF, linear is a special case of RBF, and polynomial may have numerical difficulties if a high degree is used. In this research RBF kernel is implemented whose kernel function is:

$$\exp(-\gamma \| X_i - X_j \|^2)$$

Thus, the RBF is by far the most popular choice of kernel types used in Support Vector Machines. This is mainly because of their localized and finite responses across the entire range of the real x-axis.

The following two classification problems provide insight on how SVM works: the case when the data are linearly separable and the case when the data are linearly inseparable.

2.7.2 Linear SVM: Linearly Separable Case

Considering the simplest case of a two-class problem where the classes are linearly separable (meaning that we can draw a line on a graph of A1 Vs A2 separating the two classes),

Let the dataset D be given as $(x_1, y_1), (x_2, y_2), \ldots, (x_{|d|}, y_{|d|})$

Where, xi is the set of training instances with associated class labels, yi.

Each y_i can take one of two values either $+1$ or -1, corresponding to the two classes: class-1 (A1) and class-2(A2) respectively. i.e. $y_i \in \{+1, -1\}$

This can be more elaborated using Figure 2.1 below based on two input attributes A1 (squares) and A2 (circle).

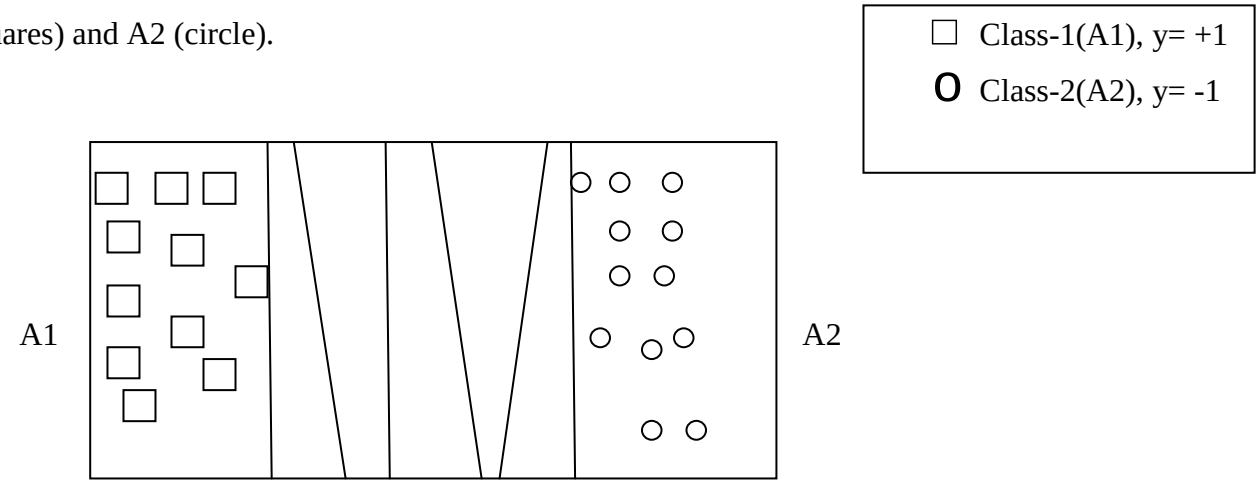


Figure 2.1 Possible Decision Boundaries for a Linearly Separable Data Set

Figure 2.1 shows a plot of data set containing examples that belong to two different classes representing as squares and circles. The data set is linearly separable; i.e. we can find a hyperplane such that all the squares reside on one side of the hyperplane and all the circles reside on the other side. However, as shown in Figure 2.1, there are infinitely many such hyperplanes possible. Although there training errors are zero, there is no guarantee that the hyperplanes will perform equally well on previously unseen examples. The classifiers must choose one of these hyperplanes to represent its decision boundary, based on how well they are expected to perform on test examples.

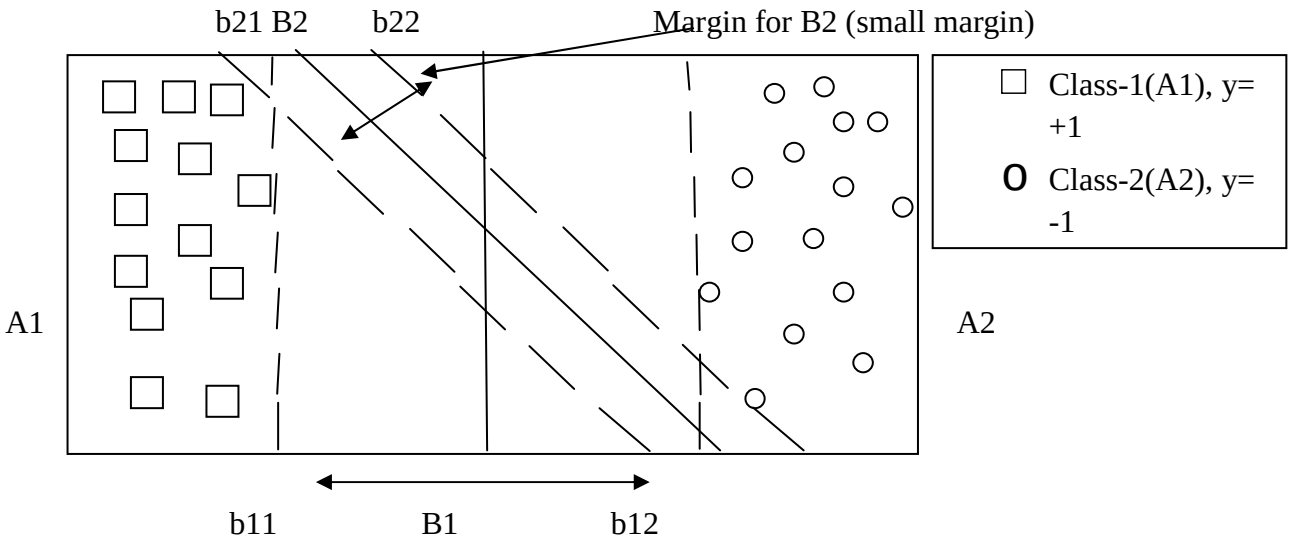


Figure 2.2 Margin for a Decision Boundary

Where,

B1 - the first hyperplane that divides class A1 and class A2 with large margin

B2 - the second hyperplane that divides class A1 and class A2 with small margin

b11 - in the matrix the first number 1 indicates the first hyperplane (B1) and the second number 1 indicate class-1, i.e. A1

b12 - in the matrix the first number 1 indicates the first hyperplane (B1) and the second number 2 indicate class-2, i.e. A2

b21 - in the matrix the first number 2 indicates the second hyperplane (B2) and the second number 1 indicate class-1, i.e. A1

b22 - in the matrix the first number 2 indicate the second hyperplane (B2) and the second number 2 indicate class-2, i.e. A2

From Figure 2.2 it can be seen that the 2-D data are linearly separable and there are an infinite number of separating lines that could be drawn. The problem is to find the best line that will have the minimum classification error on previously unseen instances. Note that for data with three attributes (3-D data) the problem would be finding the best separating *plane*. Therefore, in general for n-dimensions the problem would be to find the best *hyperplane.*

SVM approaches this problem by searching for the maximum marginal hyperplane. Consider Figure 2.2, which shows two possible separating hyperplanes and their associated margin. The figure reveals that both hyperplanes can correctly classify all the given data instances. However, decision boundaries with large margins (b11, b12) tend to have better generalization error than those with small margins (b21, b22). Intuitively, if the margin is small, then any slight perturbations to the decision boundary can have quite a significant impact on its classification. Classifications that produce decision boundaries with small margin are therefore more susceptible to model overfitting and tend to generalize poorly on previously unseen examples (Tan et.al, 2006). This is why, during the learning phase, the SVM searches for a hyperplane with the largest margin - the Maximum Marginal Hyperplane (MMH).

A separating hyperplane can be written as

$$W.X + b = 0 \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots Equation\ (2.6)$$

Where, W is a weight vector, namely, $W = \{ w_{1,}\ w_{2,}, w_{3,}, \ldots w_{n,}\}$ n is the number of attributes, and b is a scalar referred as a bias.

Consider two input attributes, $A_1$and A_2 as in Figure 2.2. The training instances are 2-D, like $X = (x_1, x_2)$, where x_1 and x_2 are the values of the attributes A_1 and A_2, respectively, for X.

Taking b as an additional weight, w_0, the separating hyperplane in $Equation\ (2.6)$ can be rewritten as:

$$w_0 + w_1 x_1 + w_2 x_2 = 0 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2.7)$$

Any point that lies above the separating hyperplane thus satisfies the equation:

$$w_0 + w_1 x_1 + w_2 x_2 > 0 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2.8)$$

Similarly, any point that lies below the separating hyperplane satisfies:

$$w_0 + w_1 x_1 + w_2 x_2 < 0 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2.9)$$

The weights can be adjusted so that the hyperplanes defining the two sides of the margin can be written as:

$$H_1 = w_0 + w_1 x_1 + w_2 x_2 \geq 1 \text{ for } y_i = +1 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2.10)$$

$$H_2 = w_0 + w_1 x_1 + w_2 x_2 \leq 1 \text{ for } y_i = -1 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2.11)$$

This means any instance that falls on or above H_1 belongs to *class-1* and any instance that falls on or below H_2 belongs to *class-2*.

Combining $Equation$ (2.10) and $Equation$ (2.11) one can write:

$$y_i(w_0 + w_1 x_1 + w_2 x_2) \geq 1, \quad \forall_i \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2.12)$$

Any training instances that fall on hyperplanes and satisfy $equation$ (2.12) are called **support vectors**. They are equally close to the separating maximum marginal hyperplane (MMH) (Han and Kamber, 2006).

Using a Lagrangian[12] formulation and solving for the solution using Karush-Kuhn-Tucker (KKT)[13] condition, $Equation\ (2.12)$ can be rewritten as a constrained convex quadratic optimization problem.

Solving the constrained convex quadratic problem is required to find the support vectors and MMH and thus train the support vector machine. Such trained SVM, are called *linear* SVMs, since the MMH is a linear class. Thus, the MMH can be written as a decision boundary, based on the Lagrangian formulation

$$d(x^{\mathrm{T}}) = \sum_i^l y_i\, a_i x_i x^{\mathrm{T}} + b_0 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.13)$$

where y_i is the class label of support vector x_i

x^{T} - is the test instance

b_0 - numeric parameters determined automatically by the SVM algorithm

a_i - are lagrangian multipliers and

l - the number of support vectors

Using the test instances x^{T} in $equation\ (2.13)$ is how classification is done by SVMs. If the sign of the result is positive, then x^{T} falls on or above the MMH, and SVM predicts that x^{T} belongs to *class-1*. If the sign is negative, then x^{T} falls on or below the MMH and the prediction is for *class-2* (Han and Kamber, 2006).

The compact prediction model of SVM comes from the fact that the learned classifier is characterized by the number of support vectors rather than the dimensionality of the data as we discussed earlier. Hence, SVMs tend to be less prone to overfitting than some other

12 In mathematical optimization, the method of **Lagrange multipliers** (named after Joseph Louis Lagrange) provides a strategy for finding the maxima and minima of a function subject to constraints.

13 the **Karush–Kuhn–Tucker** conditions (also known as the **Kuhn-Tucker** or **KKT** conditions) are necessary for a solution in nonlinear programming to be optimal, provided that some regularity conditions are satisfied.

methods. An SVM with a small number of support vectors can have good generalization, even for a high dimensional data.

2.7.3 Non Linear SVM: Linearly inseparable Case

When the data classes are not linearly separable, the approach used for linear SVM can be extended to create *nonlinear* SVMs for the classification of nonlinear separable data. Such SVMs are capable of finding nonlinear decision boundaries (i.e. non linear hypersurfaces) in input space.

Nonlinear SVM extends the approach for linear SVM using two main steps:

a) Transforming the original input data into higher dimensional space using a nonlinear mapping and then

b) Searching for a linear separating hyperplane in the new space. Thus getting a quadratic optimization problem that can be solved using the linear SVM formulation

The maximal marginal hyperplane found in the new space corresponds to a nonlinear separating hypersurface in the original space.

Considering the following example of transformation of an input data into a higher dimensional space, a 3-D input vector $X = (x_{1,}x_2, x_3)$ is mapped to a 6-D space Z, using mappings:

$\Phi_1(X) = x_1$, $\Phi_2(X) = x_2$,$\Phi_3(X) = x_3$,$\Phi_4(X) = (x_1)^2$, $\Phi_5(X) = x_1\ x_2$,and $\Phi_6(X) = x_1\ x_3$

The decision hyperplane in the new space is linear and given as

$d(Z) = WZ + b$, where z are vectors

Solving the above equation involves choosing a nonlinear mapping to a higher dimensional space and a subsequent costly calculation for the classification of test instant x^{T} (refer to $Equation$ 2.13). However, there is a way of avoiding both.

When searching for linear SVM in the new higher dimensional space, the training instances appear only in the form of dot products (Han and Kamber, 2006).

$$\Phi(X_i) \,.\, \Phi(X_j),$$

Where, $\Phi(X)$ is the nonlinear mapping function applied to transform the training instances.

Moreover, applying a *kernel function* $k(X_i, X_j)$ is found to be equivalent to computing the dot product on the transformed data instances, i.e.

$$k(X_i, X_j) = \Phi(X_i) \,.\, \Phi(X_j) \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.14)$$

$Equation$ (2.14) shows how both nonlinear mapping and calculation on transformed data can be avoided. Afterwards the maximal separating hyperplane can be found in a process similar to linear SVM, though the non-linear SVM involves placing a user-specified upper bound, C, on the Lagrange multipliers α_i . This upper bound is best determined experimentally.

Some of the kernel functions that can be used to replace the dot product (See $Equation$ 2.14) include.

Polynomial kernel of degree h: $k(X_i, X_j) = (X_i, X_j + 1)^h$ ..... (2.15)

Radial Base Function kernel: $k(X_i, X_j) = \exp(-\gamma \| X_i - X_j \|^2)$ ..... (2.16)

Sigmoid kernel: $k(X_i, X_j) = \boldsymbol{tanh}(kX_i \,.\, X_j - \delta)$ ..... (2.17)

In summary, a machine learning approach to text classification passes some steps as discussed in the previous sections of this chapter. Given a pre-classified documents, the words in a document are assembled into a dictionary and represented in a vector of terms. After feature reduction from the vocabulary, the documents are trained by the learning algorithm. Then a new instance is to be classified based on the contents of the training data. The following figure is a summary of machine learning approach to text classification.

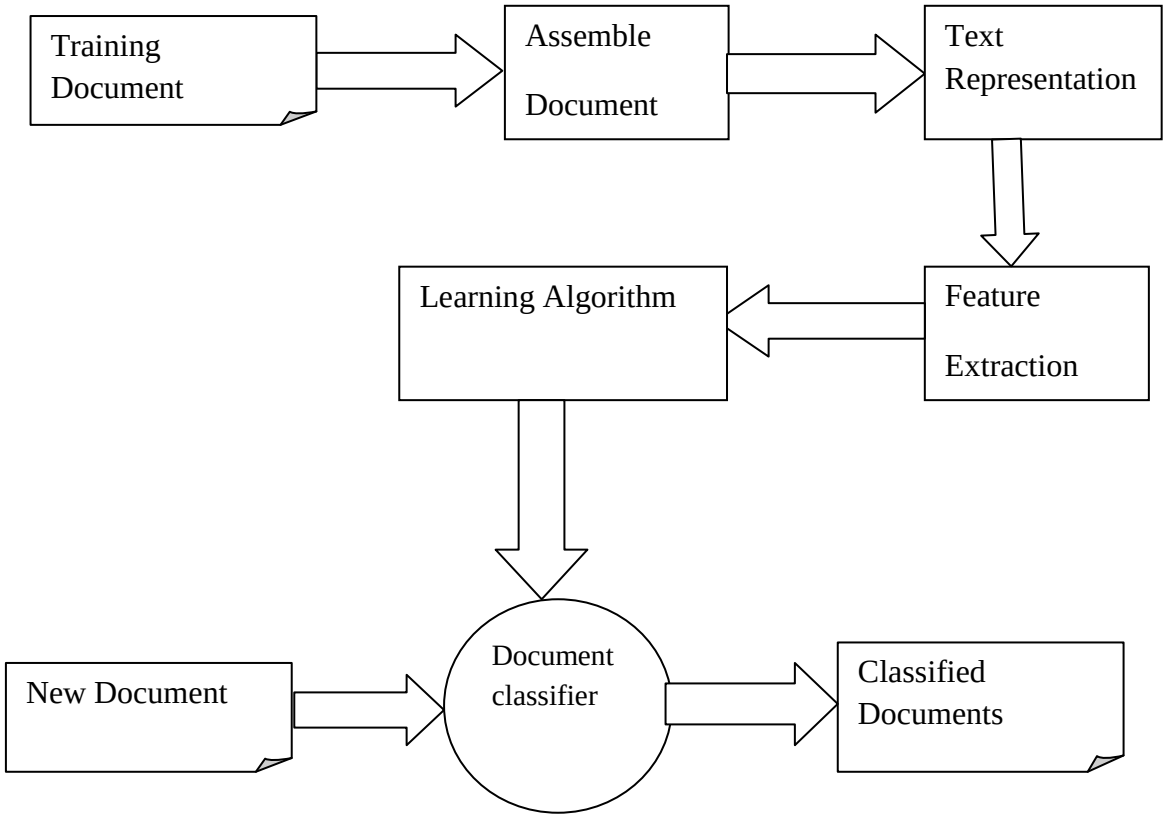


Figure.2.3 Text Classification Architecture

CHAPTER THREE

THE AMHARIC WRITING SYSTEM

3.1 Introduction

Amharic is the official working language of the Federal Democratic Republic of Ethiopia and spoken by over 20 million people as a first or second language. It is the second most spoken Semitic language in the world (after Arabic) and closely related to Tigrinya (Atelach, 2006). It is probably the second largest language in Ethiopia (after Oromo, a Cushitic language) and possibly one of the five largest languages on the African continent. Following the Constitution drafted in 1993, Ethiopia is divided into nine fairly independent regions, each with its own nationality language. However, Amharic is the language for nation-wide communication and was also for a long period the principal literal language and medium of instruction in primary and secondary schools of the country, while higher education is carried out in English (Atelach, 2006)

In spite of the relatively large number of speakers, Amharic is still a language for which very few computational linguistic resources have been developed. Very little has been done in terms of making useful higher level Internet or computer based applications available to those who only speak Amharic (Atelach, 2006). It is this fact that instigated researchers in processing the Amharic language. Observing the gap between the language and its limited application in information technology, researchers have been trying to contribute to the advancement of the language and its application in the digital age. Among those researchers Zelalem (2001), Surafeal (2003), Yohaness (2007) and Worku (2009) are some who had done their study on Amharic texts, particularly on news texts using different machine learning

approaches and bag-of-words feature representation. This research as well is to contribute to the effort that is undertaken to further the application of automatic Amharic language processing in the area of document classification, particularly news classification using phrase based feature representations.

Hence, this section briefly discusses about Amharic characters, punctuation marks, numerals being used, and the Amharic writing problems and the effect of Amharic bag-of-phrase and bag-of-words feature representation.

3.2 The Amharic Character Representation

Geez has been used as a language of literature in Ethiopia and is now used for the liturgy of the Ethiopian Orthodox Church. Written Geez can be traced back to at least the 4th century A.D. The first versions of the Geez script included only consonants while the characters in the later versions represent consonant-vowel (CV) phoneme pairs (Yohannes, 2007).

Amharic has borrowed most of its characters from Geez. Like Geez, the Amharic writing uses characters created by a CV fusion. Seven vowels are used in Amharic each of which comes in seven different forms (orders) reflecting the seven vowel sounds (ኧ ኡ ኢ ኣ ኤ እ ኦ). That is each of the 33 Amharic characters has seven forms representing a consonant and a vowel at the same time which makes the Amharic script syllabic. The first order is the basic form and there are 33 basic forms with six derivations for each giving 231 characters (Getachew, 1966). As an example, the symbolic representations of the seven forms of the Amharic characters ሀ (ha), ለ (le), መ (me) are shown in Table 3.1

1st order	2nd order	3rd order	4th order	5th order	6th order	7th 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 3.1 Seven forms of Amharic Characters

According to Bender and et. al.(1976) in addition to these, there are four so – called labio–vellars, which have five orders (e.g. ቈ ቍ ቊ ቋ ቌ) and eighteen additional labialized consonants (e.g. ኍ ሟ ሯ ሷ ሿ ቇ ኋ). A listing of the whole Amharic character set, also called fidel (ፊደል) is attached in Appendix 1.

Totally, the Amharic alphabet has around 290 letters. The alphabet does not have any distinction between capital and lower case letters.

3.3 Punctuation Marks

Amharic has its own punctuation system where the symbol ። (*hulet neteb*) is used to separate words, the symbol ፣ (*netela serez)* is used as a comma, the symbol ፤ (*dereb serez*) is used as a semicolon, and the symbol ።(*arat neteb*) is used as a full stop. The question and exclamation marks have recently been included in the Amharic writing system (see Appendix 2 for the list).

3.4 Numbers

Numbers in Amharic consist of single characters for one to ten, for multiples of ten (twenty to ninety), hundred, and thousand (see Appendix 3 for the list). According to Bender et al.

(1976), these characters are derived from Greek letters, and some were modified to look like Amharic Fidel. Each of the symbols has a horizontal stroke above and below. There is no symbol for zero in the Amharic script. As a result, people generally use the Hindu-Arabic numerals. Ethiopic numbers are used mostly in writing dates and page numbers in text (Bender et al., 1976).

3.5 Problems in Amharic Writing System

A number of problems have been observed regarding the writing system of the Amharic language. These problems are discussed below.

3.5 1. Characters (Fidels) with Different Form

Sometimes more than one character is used for similar sound in Amharic language. 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 3.2 illustrates the different forms of Amharic characters with similar sound.

Characters (Fidels)	**Other forms**
ሀ (hä)	**ሐ , ኀ , ኻ**
ሠ (sä)	**ሰ**
አ (ä)	**ዐ**
ጸ (tsä)	**ፀ**

Table 3.2 Amharic Characters with Different Forms of the Same Sound

These different representations of the same sound in different forms pose a problem for the classifier during feature preparation for text classification. For example, '**አለም አቀፍ ዜና**' may be written in different forms across different documents like '**ዓለም ዓቀፍ ዜና**' or '**ዐለም**

ዐቀፍ ዜና'. This variation of letters in each phrase has a negative impact for the machine learning process.

3.5.2 Transliteration Problems

Transliteration from foreign words to Amharic words is also another problem. The problems resulted from the use of loan words that are borrowed from other languages and that do not possess their own translation in Amharic. The word "Oxford" may be transliterated as **ኦክስፈርድ** (oxferd) or **ኦክስፎርድ** (oxford) (Bender et al., 1976). In addition, the word "Television" could be transliterated in different forms of the Amharic characters like **ቴሌቪዥን**, **ቴሌብዥን**, or **ተለብዥን**. Again such transliteration problems could negatively affect the classifier.

3.5.3 Abbreviations

No consistency is kept in writing Amharic abbreviations. For example, the phrase '**ዓመተ ምህረት**, meaning 'AD', can be abbreviated as **ዓም**, **ዓ.ም** or **ዓ/ም**. This poses challenges since the same word is treated in different forms in the process of feature preparation for text classifier. This inconsistency creates another problem for automatic classification.

3.6 Amharic Bag-of-Words vs. Bag-of-Phrases in Text Classification

As discussed previously (in section1.2), the main drawback of the b*ag-of-words* feature representation is that it destructs the semantic relations between words by using words in a phrase separately. For example, the phrase "**እግር ኳስ**" which means "football" intuitively is classified under the domain "sport". However, given a *bag- of-words* of a document in which words "**እግር**" (foot) and "**ኳስ**" (ball) occur, one can suggest that the document is talking about an *animal* or a *furniture leg*, or about a *ball* which is not specific to football. Whereas given a document representation that contains a phrase "**እግር ኳስ**", the reader will not be

mistaken about the topic of discussion. Moreover, single words are rarely specific enough to support accurate discrimination. For example, "**ግብርና ምርምር**"(agricultural research) if the single term "**ምርምር**`" (research) occur more number of times in a document about medical science than a document about agriculture, word-based classification cannot distinguish the phrase **ግብርና ምርመር** (agricultural research) (Worku, 2009). Consequently misclassification problem could occur, i.e. a document which talks about agriculture will be classified under Health news group due to the single term "**ምርምር**" (research). However, if the phrase "**ግብርና ምርመር**" (agricultural research), is taken together this misclassification problem might not occur.

Furthermore, the *bag-of-words* representation approach increases the dimensionality of feature space which is known to have a negative effect on the classification performance. For example, **ፌዴሬሽን ምክር ቤት** (Federation Meker Bet) corresponds to one feature using *bag-of-phrases (trigram)* approach and corresponds to three features in *bag-of-words* approach. Above all, information is lost due to feature splits which affects achieving higher classification performance.

Therefore, using phrases as features is one solution for incorporating word sequence information into the Amharic news text classification.

CHAPTER FOUR

AUTOMATIC AMHARIC NEWS TEXT CLASSIFICATION

4.1 Introduction

In this chapter the data source and data filtering mechanisms carried out will be explained briefly. Moreover, the preprocessing algorisms (i.e. all major preprocessing algorithms developed to process Amharic news texts like character (Fidel) normalization, stemming, stop word and number removal) will be discussed. The processes involved in selecting class representative phrases out of the news items of a category will also be explained. In addition, the steps followed in the conversion of the preprocessed data to the Coma Separated Value (CSV) file format which is appropriate for use in WEKA application package to classify the Amharic news texts will be discussed. Furthermore, the experimentation on training and testing the classifier, and the evaluation procedures followed for automatic classification of Amharic news texts will be detailed. Finally, the performances of the WEKA classifier (LibSVM) for two phrase structures (bigrams and trigram) at different number of news category levels will be compared and the results will be discussed.

4.2 The Data Source

As discussed in section 1.5.2 the data for this study is collected from Ethiopian News Agency (ENA). ENA news items are manually classified into twelve major and ninety eight sub categories. All in all 16,075Amharic news items in *"html"* format are collected from 2006-2010.

However, not all of these news items are useful for the classification experiment because of errors made during data entry and manual classification. Therefore, applying data filtering was the right approach in order to take relevant news documents.

4.3 Data Filtering

Due to the problems of ENA's news classification system discussed in section 1.2 not all news items were taken for this research. Hence, the following filtering procedures were followed.

As discussed earlier, news items were considered for the study only if they have data for the Headline and for the Keywords. If data is missing in one of the two news sections, the news item is dropped. A lot of attention was given for the identification and clearing of the misclassification errors using manual scanning of the news items, a process which took a lot of time and effort. As a result, the following classification errors were identified.

- Entering the same news items to more than one category
- Entering a news item two or more times into one category
- Entering news items to categories that have no relation to the news

After the identification of the misclassified news items, necessary corrective actions were taken during the sampling of news items for the experiment, i.e., duplicate news items were not considered for the experiment.

For example, in the Social category, 2440 data entry errors were identified. Out of which 350 are repeated news entries, and 2090 are entries that do not belong to the Social category (i.e. classification errors). Out of the 2090 classification errors, 1076 belong to the Education, the Economy, and the Health categories, and the rest 1014 belong to the Politics and Culture and Tourism news categories. Similarly, a total of 4074 records of the Sport, the Accident, the Weather, the Defense, and Science and Technology categories were identified as repeated

entry and misclassification errors. All these activities were performed with the help of an expert from the news agency.

Table 4.1 shows filtered news documents for the study after correcting the identified classification and data entry errors.

No.	News Categories	Total No.of News Items Collected	No. of news Items considered	No. of News items Dropped due to Errors
1	Culture and Tourism (ባህልና ቱሪዝም)	1110	720	390
2	Economy (ኢኮኖሚ)	1734	1481	253
3	Education (ትምህርት)	917	670	247
4	Health (ጤና)	1160	584	576
5	Law and Justice (ህግና ፍትህ)	1289	820	469
6	Politics (ፖለቲካ)	1786	1200	586
7	Social (ማህበራዊ)	4764	2324	2440
8	Sport (ስፖርት)	809	480	329
9	Accident (አደጋ)	1102	850	252
10	Weather (የአየር ሁኔታ)	1241	325	916
11	Relations, Defense, and Security (ግንኙነት፣ መከላከያና ደህንነት)	1915	200	1715
12	Science and Technology (ሳይንስና ቴክኖሎጂ)	982	120	862
	Total	16,075	9,775	6,300

Table 4.1 News Documents Considered for the Experiment

Hence, from the total 16,075 news documents collected from ENA after filtering was done only 9,775 news items were selected for the Experiment. These 9,775 documents were taken

for further preprocessing. The following section briefly shows the preprocessing subsystem applied to the Amharic news documents, the detailed explanation of each preprocessing subsystem (algorithm) is discussed in 4.5.1.1

4.4 The Preprocessing Subsystem

The preprocessing step of the experiment involves phrase-level processing of the source dataset with the ultimate aim of identifying feature phrases that are representatives of the documents in the dataset. This step also involves the conversion of the preprocessed data to **CSV**, which is suitable for the WEKA package used for automatic classification. Figure 4.1 shows the components of the design of the preprocessing subsystem.

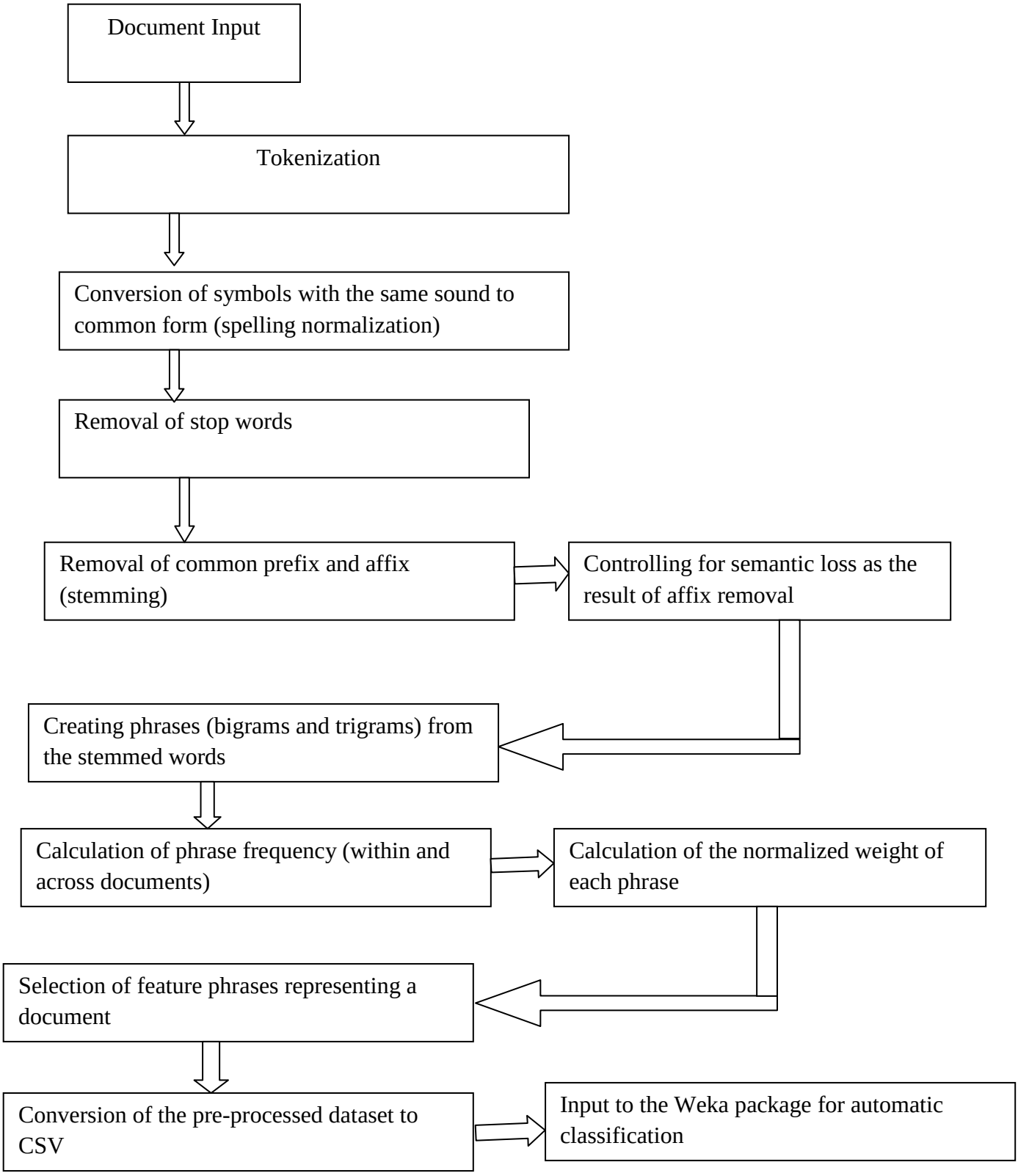


Figure 4.1 The General Description of the Data Preprocessing subsystem

4.5 Document Processing

Section 4.4 depicted the graphical representations of data preprocessing steps which should be taken for Amharic news document preparation for classification. This section goes in detail with document processing that are important for the efficiency and accuracy of automatic classification, i.e., relevance analysis, data reduction, data transformation, and data conversion.

4.5.1 Representing Documents by Relevant Phrases

Since documents are considered as a collection of phrases for the purpose of classification, identifying phrases that adequately represent a document is important. A document can have many phrases but not all phrases of a document are equally important to uniquely identify documents. We need to have a mechanism to represent documents.

A text document is typically represented as a vector of phrase weights (phrase features) from a set of phrases (dictionary), where each phrase occurs at least once in a certain minimum number of document. A major characteristic of the text classification problem is the extremely high dimensionality of text data. The number of potential features often exceeds the number of training documents.

Document preprocessing[14] and dimensionality reduction (DR) allows an efficient data manipulation and representation. DR is the exclusion of a large number of keywords, based preferably on a statistical process, to create a low dimension vector. Effective dimension reduction makes the learning task more efficient and save more storage space. DR is a very important step in text classification, because irrelevant and redundant features often degrade the performance of classification algorithms both in speed and classification accuracy and

[14] All the preprocessing activities like character normalization, stop word removal, stemming etc, was done with Alemu Kumilachew as group project

also its tendency to reduce overfitting. DR techniques can be classified into Feature Extraction (FE) and Feature Selection (FS) approaches, as discussed below.

4.5.1.1 Feature Extraction

The process of preprocessing is to make clear the border of each language structure and to eliminate as much as possible the language dependent factors, tokenization, stop words removal, and stemming. FE is the first step of preprocessing which is used to present the text document into clear phrase format. So character normalization, tokenization, removing stop words, and stemming words are the pre-processing tasks. Hence, the steps which took place for feature extraction for text classification are discussed below.

4. 5.1.1.1 Character Normalization

In Amharic language there are different variants of characters (Fidels) which have the same sound but different forms (shapes). Among these characters, for example, ሀ (ha) has different forms such as "ሀ" ፣ "ሃ" ፣ "ሐ" ፣ "ኀ". All these different forms of characters (Fidels) represent the same sound. Therefore, these variants of characters (Fidels) need to be converted to the same or common character form (shape) in order to avoid representing the same phrase using different letters having the same sound which will negatively affect the frequency of terms with in a document and/or across documents. Therefore, by representing a common letter (Fidel) for all characters which have the same sound as done in *Fikir Eskemekabir* (Haddis Alemayehu, 1958), we can control the effect of letters having the same sound and increase the occurrence of terms in a given document. The algorithm of normalizing characters (Fidels) which is implemented in the process of Amharic news text classification is shown in Table 4.2 below.

```
Read the content of a file
Do
        If word exist in variant list then
            Replace with common character
        End If
        Assign the content to a variable
While End Of File
```

Table 4.2 Algorithm for Character (Fidel) Normalization

The algorithm implemented (Table 4.2) was effective in that all variants of characters (Fidels) were replaced to their common forms.

4.5.1.1.2 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, phrases (bigrams and trigrams) are taken as tokens. All punctuation marks except end of a sentence (።) are converted to space and space is used as a word demarcation. The exclusion of the punctuation mark (።) from the removal list is to use it as control of end of a sentence while creating phrases i.e. to avoid combination of words from different sentences. A phrase should be created from its own sentence. The last word of a sentence and the first word of the next sentence should not be combined together to create two or three consequent word phrases. If done, this may create a phrase which has different context in meaning from a given category of news items. For example, if the first sentence of a news item which talks about *sport category* ends with ‘sport federation’ (**ስፖ'ርት ፌዴሬሽን**) after

stop word removal and the next sentence talks about the participation of youths in football games and starts with the word '*wotatoch*' (**ወጣቶች**), i.e. the first and the next sentence is written as *("…..**ስፖርት ፌዴሬሽን አስታወቀ ::**" " **ወጣቶች ባደረጉት ተሳትፎ**….."*), combining these words (the last word of the first sentence and the first word of the next sentence) will create a phrase which has different context. Taking those words will create a phrase which belongs to a different news category. This phrase is '***ፌዴሬሽን ወጣቶች*** which is under *politics news category (**አስታወቀ*** is removed as stop word). Thus, while removing punctuation during the tokenization process we need not avoid end of sentence punctuation mark (**::**). This punctuation mark declares the end of a sentence. Therefore, by using this punctuation mark we can control reading sentences and creating phrases only from each sentence and we can avoid combining two words which may have extremely different context from a given news category. The following table 4.3 shows the algorithm for the tokenization process.

```
Read content of file
Do
        If the content has punctuation marks except end of a sentence
                Substitute it with blank space
Split each word and assign to a variable
While End Of File
```

Table 4.3 An Algorithm for Tokenization

The next step in the process of feature extraction is stop word removal which is discussed below.

4.5.1.1.3 Stop word Removal

As discussed in before, stop words are non content bearing words, which are less discriminating among documents since they appear in almost all documents. Using this concept, all high frequency words that exist in all documents were retrieved and reviewed to identify stop words of two types. These are news specific stop words and common stop words.

a) News Specific Stop Words

Reporters and journalists most of the time report an incident to the public. As a result, they use vocabularies peculiar for this purpose. An example of such words, which they use very frequently, is "notify" (**አስታወቀ፣ ገለፀ፣ አመለከተ፣ ጠቆመ፣ ዘገበ**). Basically, they use this word when reporting about an official or organizational press release. In fact, these words are pure verbs and are usually found at the end of a sentence. Therefore, the identification of these words was done in two steps:

- List of high frequency terms from all news categories were selected.
- Then, common terms in all categories were selected, and the result was reviewed manually. The assessment was made with an expert from the Agency.

b) Common Stop Words

Like other languages, some words in Amharic are used very frequently in the normal usage of the language. There are common stop words which are used for grammatical purposes like **ነው፣ ነበር፣ ሆኖም ፣ እና ፣ ነገር ግን**, etc, which are not informative to identify one document from the other. Because of the unavailability of comprehensive standard stop word lists done by previous researchers, the researcher of this study is required to develop stop word lists.

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 word

list is prepared after identifying stop words; the list that contains, words which have to be removed from tokens (keyphrases) generated during the tokenization process. The need of manual inspection is because of frequently occurring keyphrases. Using the two techniques mentioned above a total of 984 stop words were produced, and the implementation of this step reduced the feature size on average by 20%. The stop word lists are listed in appendix 4. Table 4.4 demonstrates the algorithm in removing stop words and numbers from token list.

```
Read each token from token list
If token in Stop List then
        Remove from token list
end if
If token is number then
        Remove from token list
end if
```

Table 4.4 An Algorithm for Stop Word and Number Removal

4. 5.1.1.4 Stemming

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

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 '**ን**' ('ne') from the word '**ክርሰቲያን**' ('kersetiyane') would give '**ክርሰቲያ**' ('keresetiya'), which is meaningless; and removing '**በ**' ('be') from '**በርበሬ** ('berbri') gives '**ርበሬ**' ('rbri'),which does not represent the original meaning. Hence, two exception lists were prepared for which affix removal rules do not apply.

- List of words that prefix removal rule does not hold true and
- List of words from which suffix removal rule does 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. The following Table 4.6 shows the stemmer algorithm.

```
Read tokens
For each token in list
If token starts with prefix
        If token not in prefix-exceptional list then
        Remove prefix
End If
If token ends with suffix
        If token not in suffix-exceptional list then
        Remove suffix
End if
Update token list
```

Table 4.5 A **s**temmer Algorithm

An example of words with prefix and suffix removed is shown in Table 4.6 below.

Original word	**Prefix**	**Suffix**	**Stemmed(root) word**
የተማሪዎችን	**የ**	**ዎችን**	**ተማሪ**
ተማሪዎች		**ዎች**	**ተማሪ**

Table 4.6 Example Prefix and Suffix Removed

The second pre-processing step in the processes of document representation is feature selection which is discussed below.

4.5.1.2 Feature Selection

After feature extraction, the other important step in the preprocessing task of text classification to construct vector space which improves the scalability, efficiency and accuracy of a text classifier is feature selection (FS). The main idea of feature selection is to select subset of features from the original documents.

Feature selection is performed by keeping the phrases with highest score according to predetermined measure of the importance of the phrase. The selected features retain original meaning and provide a better understanding for the data and learning process. For text classification a major problem is the high dimensionality of the feature space. Almost every text domain has much number of features, most of these features are not relevant and beneficial for text classification task, and even some noise features may sharply reduce the classification accuracy. Hence, FS is commonly used in text classification to reduce the dimensionality of feature space and improve the efficiency and accuracy of classifiers.

There are mainly two types of FS methods in machine learning; wrappers and filters. Wrappers use the classification accuracy of some learning algorithms as their evaluation function. Since wrappers have to train a classifier for each feature subset to be evaluated, they are usually much more time consuming especially when the number of features is high. So wrappers are generally not suitable for text classification and are not implemented in this research. As opposed to wrappers, filters perform FS independently of the learning algorithm that will use the selected features. In order to evaluate a feature, filters use an evaluation metric that measures the ability of the feature to differentiate each category. In text classification, a text document may partially match many categories. We need to find the best matching category for the text document. The term (phrase) frequency (TF) approach is commonly used to weight each phrase in the text document according to how unique it is. In

other words, the TF approach captures the relevancy among phrases, text documents and particular categories.

Using the TF approach those features (phrases) with good discrimination power have been selected for the Amharic news text classification tasks. Hence, dimensionality reduction method using Document Frequency (DF) thresholding is carried out to reduce features further (see Table 4.7). The reduction is required to represent news with the most important features and hence, reduce computational complexity. DF value of a phrase is the number of documents which contain that phrase. Therefore, 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 phrases for each category. This enables to check important phrases which do not satisfy the DF threshold and also irrelevant phrases which satisfy the DF threshold.

For example, the phrase "**አየር መንገድ**" ("Airline") occurred in 37 news items in the category "Accident" which satisfy DF threshold. But the term is not keyword for the category according to the experts' judgment; as a result, the term is excluded from the features list of the category. On the other hand, the phrase "**አውሮፕላን አደጋ**" has DF of 26, which is below the threshold, but it is the most important keyphrase in the category "Accident"; hence, it is included in the feature list to represent the category.

No.	Major News Categories	Total No. of Documents	DF Threshold	No. of Features Selected
1	Culture and Tourism (ባህልና ቱሪዝም)	720	20	127
2	Economy (ኢኮኖሚ)	1481	30	570
3	Education (ትምህርት)	670	30	366
4	Health (ጤና)	584	30	458
5	Law and Justice (ህግና ፍትህ)	820	30	246
6	Politics (ፖለቲካ)	1200	30	658
7	Social (ማህበራዊ)	2324	40	934
8	Sport (ስፖርት)	480	20	180
9	Accident (አደጋ)	850	35	301
10	Weather (የአየር ሁኔታ)	325	15	87
11	Relations, Defense, and Security (ግንኙነት፣ መከላከያና ደህንነት)	200	10	37
12	Science and Technology (ሳይንስና ቴክኖሎጂ)	120	10	41
	Total			4,005

Table 4.7 DF Threshold and Number of Features for Each Category

4.5.2 Data Transformation

After feature selection, a document is treated as collection of the representative feature phrases while the whole dataset is a collection of documents. Feature phrases representing a document are therefore attributes of the document that are given some value to reflect the degree to which they represent a document.

In this study document attribute values are calculated using the frequency of feature phrases in a document. The frequency is then normalized as shown in Equation 2.1. The attribute

values are scaled by the inverse document frequency factor (using the relation in Equation 2.3) to reflect how uniquely an attribute represents a document with respect to other documents of the same category. Moreover, an attribute is taken as document representative only if its frequency in the document is greater than one (before it is normalized).

For example, consider the two phrases **ሰንደቅ አላማ** (Flag) and **ብሄር ብሄረሰብ** (Nations and Nationalities) that are found in two different documents that belong to the Culture and Truism news category. Both phrases are taken as attributes of their respective news documents because **ሰንደቅ አላማ** has a frequency of 7 and is found in 4 Culture and Tourism documents while **ብሄር ብሄረሰብ** has frequency of 5 and is found in 35 Culture and Truism documents. The phrase **ሰንደቅ አላማ**, however, has much bigger attribute value (5.12330530552) than the phrase **ብሄር ብሄረሰብ** with an attribute value of (2.98859476155), hence it is more discriminative.

4.5.3 Data Conversion

After feature selection is completed and number of feature phrases per category is known, the next step is converting the dataset to a format appropriate for automatic classification. As discussed in section 1.7, WEKA application package is used to classify Amharic news documents.

WEKA expects the source data for classification to be in comma separated value (CSV) format. After finding the TFIDF value (phrase weight) of each feature per category the result is exported to Microsoft Excel application and saved with coma delimited (CSV) file format to make it ready for classification. A sample of the processed data is shown in Table 4.8.

No	News Category/class	Features(phrases)	Phrase Weight
1	አደጋ/Accident	አውሮፕላን አደጋ	3.32192809489
2	አደጋ/Accident	ህንፃ ፍንዳታ	3.11010011111
3	አደጋ/Accident	አውሎ ንፋስ	3.11010011111
4	አደጋ/Accident	ሰደድ እሳት	3.50154101222
6	ትምህርት/Education	መማር ማስተማር	2.01267542121
7	ትምህርት/Education	ድህረ ምረቃ	2.03521013111
8	ኢኮኖሚ/Economy	ውጭ ምንዛሬ	2.09120313122
9	ስፖርት/sport	እግር ኳስ	0.29836212312

Table 4.8 Sample Experiment Data format

4.6 The Experiment

In order to deal with the effect of two forms of phrase structures (i.e. bigrams and trigrams) for Amharic news text classification system, this experiment was performed in two phases. The first phase was done using bigram phrase structures for different numbers of news categories. Bigrams phrases are those phrases which are made up of two sequential words after preprocessing is performed on the news documents. The second phase of the experiment was performed by using trigram phrase structures which constitute of three sequential words. The only difference between the two experiments is in representing features.

In addition, to see the performance of the classifier at different category levels for both phrase structures, the experiment was done on various numbers of news categories. The experiment started with four categories, and then the number of categories is increased to eight and finally all twelve categories were tested. Finally, the results of the classifier on both phrase

structures (bigrams and trigrams) are compared and analyzed as a better phrase based feature representation for automatic Amharic news text classification.

For the reasons discussed in Chapter 2, for both experiments Support Vector Machine (SVM) was used. WEKA came about through the perceived need for a unified workbench that would allow researchers easy access to state-of the-art techniques in machine learning. The WEKA version 3.6 used for the experiment comes with the following SVM classifiers: GaussianProcesses and Sequential Minimal Optimization (SMO). Moreover, WEKA supports an add-on SVM function Library of SVM (LibSVM). The LibSVM is implemented in WEKA by the classifier class: weka.classifiers.functions.LibSVM.

The whole experiment dataset has two types of attributes: numeric attributes for the weight of the feature phrases and nominal attributes for the category labels .Since the GaussianProcesses function cannot handle nominal classes, only the performances of the LibSVM and the Sequential Minimal Optimization (SMO) classifiers were tested on the dataset. After testing the dataset with both the SOM and the LibSVM classifiers, the LibSVM classifier was found to be better. Hence, LibSVM is used for the experiment.

4.6.1 Model Building Using Bigrams and Testing Classification Accuracy

Classifiers rely on being trained before they can reliably be used on new data. Of course, it stands to reason that the more instances the classifier is exposed to during the training phase, the more reliable it will be as it has more experience.

However, once trained, we would like to test the classifier too, so that we are confident that it works successfully. For this, yet more unseen instances are required.

A problem which often occurs is the lack of readily available training/test data. These instances must be pre-classified which is typically time-consuming (hence the reason to

automate it with a software classifier). A nice method to circumvent this issue is known as cross-validation. It works as follows:

1. Separate data into fixed number of partitions (or folds)
2. Select the first fold for testing, whilst the remaining folds are used for training.
3. Perform classification and obtain performance metrics.
4. Select the next partition as testing and use the rest as training data.
5. Repeat classification until each partition has been used as the test set.
6. Calculate an average performance from the individual experiments.

Therefore, this experiment was carried out using 10-fold cross validation.

The experiment began by loading the data into WEKA. Next, the LibSVM classifier was selected from the lists of available classifiers. Then, the various parameters for the classifier (like SVMType: C-SVM, and kernelType: RBF) were specified. Finally, the 10-fold cross-validation was selected, this is necessary to get a reasonable idea of accuracy of the generated model.

The following section discusses the results obtained for the first phase of the experiment (bigram phrase structures) at different category levels.

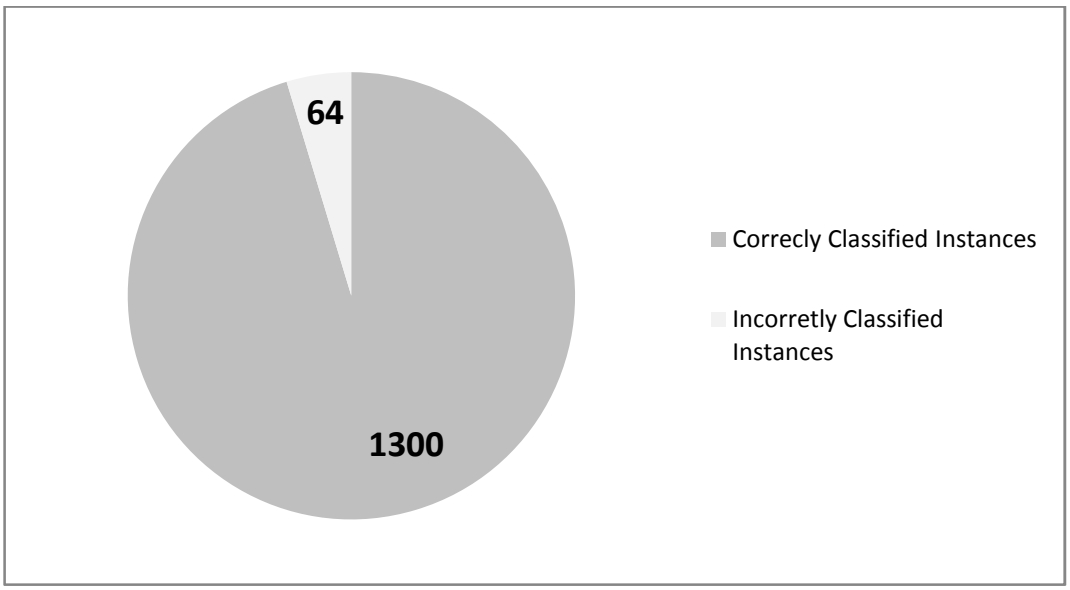
i. Experiment for the Four News Categories

The experiment started with four categories: Accident (**አደጋ**),Culture and Tourism (**ባህልና ቱሪዝም**), Economy (**ኢኮኖሚ**) and Education (**ትምህርት**). These categories have a total of 1,364 instances (see Table 4.9).

No.	Category	No. of Instances
1	Accident (አደጋ)	301
2	Culture and Tourism (ባህልና ቱሪዝም)	127
3	Economy (ኢኮኖሚ)	366
4	Education (ትምህርት)	570

Table 4.9 Four news Categories with their Number of Instances

The following Graph 4.1 shows summary statistics of instances which are correctly classified and which are not.



Graph 4.1 Correctly and Incorrectly Classified Instances for the Four News Categories using Bigram Phrase Structures

The LibSvm classifier was tested on the dataset of four categories. The experiment was performed using 10-fold stratified cross-validation as stated above. The overall accuracy of the LibSVM classifier was found to be 1300 (95.3079%) correctly classified instances and 64 (4.6921%) incorrectly classified instances out of the total of 1,364 instances.

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class/ category
1	0	1	1	1	1	**ACCIDENT**
0.874	0.002	0.982	0.874	0.925	0.936	**CULTURE**
0.996	0.078	0.902	0.996	0.947	0.959	**ECONOMY**
0.874	0	1	0.874	0.933	0.937	**EDUCATION**

Table 4.10 Four Category LibSVM Detailed Accuracy by Class

Observation of the detailed accuracy for each category in table 4.10 reveals that the SVM classifier performs better for the Accident classes followed by the Economy and Education classes which have large number of instances. A relatively less performance result is observed on the Culture category which has less number of instances.

=== Confusion Matrix ===

ACCIDENT	CULTURE	ECONOMY	EDUCATION	Classified as
301	0	0	0	**ACCIDENT**
0	111	16	0	**CULTURE**
0	2	568	0	**ECONOMY**
0	0	46	320	**EDUCATION**

Table 4.11 Four Categories LibSVM Confusion Matrix

The confusion matrix shown in Table 4.11 confirms the detail class statistical values used to determine the class level accuracy of the classifier. Hence, in Table 4.11, out of 301 instances of the Accident categories all of them are classified as Accident. No instances are incorrectly classified to other categories. However, out of the 127 instances of the Culture category 111 instances are classed as Culture, and 16 instances incorrectly classified as Economy class. Likewise, 2 instances of the Economy categories are incorrectly classified as Culture out of 366 total instances, and 46 instances are wrongly classified as Economy categories out of a total of 570 Education instances.

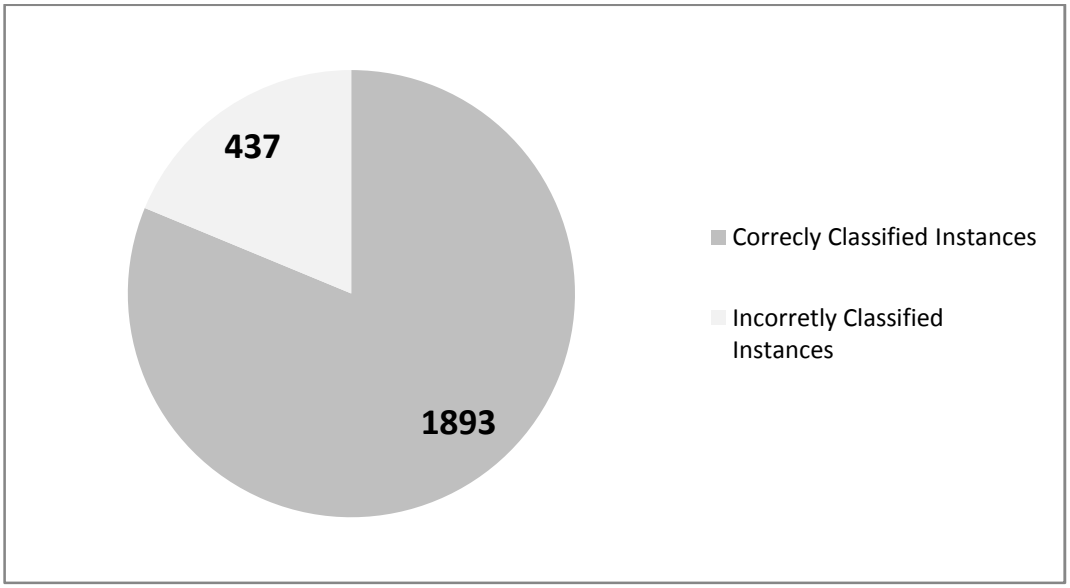
ii. Experiment for Eight News Categories

The eight news categories/classes taken for the experiment are shown in Table 4.12 below with their number of instances.

No.	Categories	No. of instances
1	Accident (አደጋ)	301
2	Economy (ኢኮኖሚ)	570
3	Education (ትምህርት)	366
4	Culture and Tourism (ባህልና ቱሪዝም)	127
5	Science and Technology (ሳይንስና ቴክኖሎጂ)	41
6	Politics (ፖለቲካ)	658
7	Weather (የአየር ሁኔታ)	87
8	Sport (ስፖርት)	180

Table 4.12 Eight News Categories with their Number of Instances

For the eight categories, the LibSVM classifier shows an average accuracy of 1893 (81.2446%) for the dataset of 2,330 instances with 437(18.7554%) classification errors which is shown in Graph 4.2 below.



Graph 4.2 Correctly and Incorrectly Classified Instances for Eight News Categories using Bigram Phrase Structures

=== Detailed Accuracy by Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.874	0.055	0.703	0.874	0.779	0.91	**ACCIDENT**
0.622	0	0.988	0.622	0.763	0.811	**CULTURE**
0.939	0.064	0.827	0.939	0.879	0.937	**ECONOMY**
0.964	0.032	0.851	0.964	0.904	0.966	**EDUCATION**
0.842	0.072	0.821	0.842	0.831	0.885	**POLITICS**
0.146	0	0.857	0.146	0.25	0.573	**SCIENCE**
0.561	0.013	0.777	0.561	0.652	0.774	**SPORT**
0.023	0	1	0.023	0.045	0.511	**WEATHER**

Table 4.13 Eight Category LibSVM Detailed Accuracy by Class

For the eight categories dataset, observation of the values of the TP-Rate, the Precision, the F-measure, and the ROC-Area in Table 4.13 shows that LibSVM classifier has better accuracy for categories with relatively larger instances in the dataset Economy and Education. Categories which have less number of instances are dominated by categories which have large instances. Hence, more instances are incorrectly classified from those categories which have less number of instances. This is observed from the confusion matrix (Table 4.14) in Weather, Sport, and Science categories/classes.

Confirmation of the values used in the class detail accuracy (Table 4.13) is given by the confusion matrix for the eight categories in Table 4.14. For example, the True Positive Rate of the Education class is calculated from the data in the confusion matrix (Table 4.14) as follows:

$$TP = \frac{353}{0+0+13+353+0+0+0+0} = 0.964$$

This means that the TP rate of the Education category (0.964) shows better performance achievement than the TP rate of the Politics category (0.842).

=====Confusion Matrix=====

ACCIDENT	CULTURE	ECONOMY	EDUCATION	POLITICS	SCIENCE	SPORT	WEATHER	Classes
263	0	0	0	26	0	12	0	ACCIDENT
0	**79**	48	0	0	0	0	0	CULTURE
0	0	**535**	35	0	0	0	0	ECONOMY
0	0	13	**353**	0	0	0	0	EDUCATION
27	0	49	23	**554**	0	5	0	POLITICS
18	0	0	0	10	**6**	7	0	SCIENCE
30	0	0	0	49	0	**101**	0	SPORT
36	1	2	4	36	1	5	**2**	WEATHER

Table 4.14 Eight Categories LibSVM Confusion Matrix

iii. Experiment for Twelve News Categories

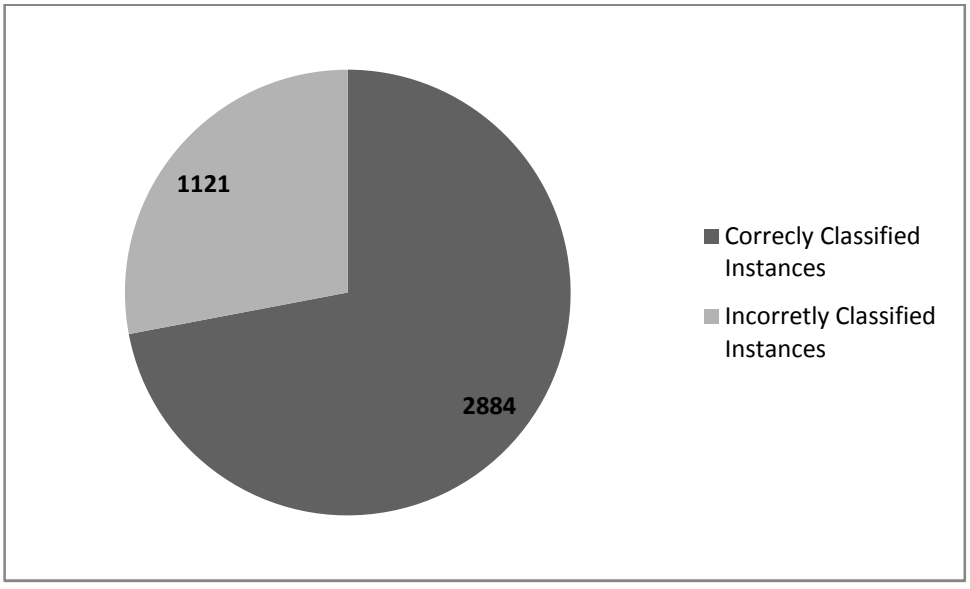
The final experiment using bigram phrase structures is for the whole news categories (12 categories). The twelve news categories with their number of instances used for classification are shown in Table 4.15 below.

No.	News Categories	No. of Instances
1	Culture and Tourism (**ባህልና ቱሪዝም**)	127
2	Economy (**ኢኮኖሚ**)	570
3	Education (**ትምህርት**)	366
4	Health (**ጤና**)	458
5	Law and Justice (**ህግና ፍትህ**)	246

6	Politics (**ፖለቲካ**)	658
7	Social (**ማህበራዊ**)	934
8	Sport (**ስፖርት**)	180
9	Accident (**አደጋ**)	301
10	Weather (**የአየር ሁኔታ**)	87
11	Relations, Defense, and Security (**ግንኙነት፣ መከላከያና ደህንነት**)	37
12	Science and Technology (**ሳይንስና ቴክኖሎጂ**)	41

Table 4.15 Twelve news categories with their number of instances

From the total instances (i.e. 4,005) of the twelve news categories, the LibSVM classifier showed an average accuracy of 2884 (72.01%) while 1121 (27.99%) of the instances were incorrectly classified which is shown in Graph 4.3 below.



Graph 4.3 Correctly and Incorrectly Classified Instances for Twelve Categories Using Bigram Phrase Structures

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.595	0.027	0.642	0.595	0.617	0.784	**ACCIDENT**
0.748	0.003	0.88	0.748	0.809	0.872	**CULTURE**
0.703	0.004	0.619	0.703	0.658	0.849	**DEFENCE**
1	0.038	0.812	1	0.896	0.981	**ECONOMY**
0.874	0.005	0.941	0.874	0.907	0.934	**EDUCATION**
0.731	0.061	0.606	0.731	0.663	0.835	**HEALTH**
0.297	0.019	0.51	0.297	0.375	0.639	**LAW**
0.67	0.052	0.718	0.67	0.693	0.809	**POLITICS**
0.22	0	0.9	0.22	0.353	0.61	**SCIENCE**
0.884	0.119	0.693	0.884	0.777	0.883	**SOCIAL**
0.044	0.002	0.471	0.044	0.081	0.521	**SPORT**
0.023	0.001	0.4	0.023	0.043	0.511	**WEATHER**

Table 4.16 Twelve categories LibSVM Detailed Accuracy by Class

Observation of the detail statistics from the values of the TP-Rate, the Precision, the F-measure, and the ROC-Area reveals that for 12 categories (Table 4.16) dataset, the LibSVM classifier shows best performance for categories with relatively larger instances in the (Economy, Education and Social). The classifier's accuracy is least for categories with relatively small instances in the dataset i.e. Weather, Sport and Science news categories.

=== Confusion Matrix ===

The accuracy measures are calculated on the basis of the data in Table 4.17, for the confusion matrix of the 12 categories.

ACCIDENT	CULTURE	DEFENCE	ECONOMY	EDUCATION	HEALTH	LAW	POLITICS	SCIENCE	SOCIAL	SPORT	WEATHER	Classified as
179	0	0	0	0	23	31	4	0	64	0	0	ACCIDENT
0	**95**	16	16	0	0	0	0	0	0	0	0	CULTURE
4	6	**26**	0	0	0	1	0	0	0	0	0	DEFENCE
0	0	0	**570**	0	0	0	0	0	0	0	0	ECONOMY
0	0	0	46	**320**	0	0	0	0	0	0	0	EDUCATION
0	0	0	20	0	**335**	7	31	0	65	0	0	HEALTH
42	0	0	0	0	37	**73**	0	0	94	0	0	LAW
0	6	0	35	14	73	6	**441**	0	72	8	3	POLITICS
10	0	0	0	0	1	2	0	**9**	19	0	0	SCIENCE
22	0	0	0	0	43	10	33	0	**826**	0	0	SOCIAL
0	0	0	14	0	34	8	84	0	32	**8**	0	SPORT
22	1	0	1	6	7	6	20	1	20	1	**2**	WEATHER

Table 4.17 Twelve Category LibSVM Confusion Matrix

In this first part of the experiment using bigram phrase structures, it was observed that those news categories with relatively large number of instances have better classification accuracy than those news categories with small instances. The other point which is observed in this experiment is that for small number of news categories bigram phrase structures performed well with performance accuracy of (95.4%) correctly classified instances for the four news categories. However, the performance of the LibSVM classifier using bigram phrase

structures decreased to 81.3% for all 2330 instances of the eight categories, and to 72.01 % for all twelve news categories with total (4,005) number of instances.

Table 4.18 summarizes the results obtained from the experiment for different numbers of news categories for bigram phrase structures.

No. of news Categories	**Correctly Classified Instances**	**Incorrectly Classified Instances**
4	1300 (95.3%)	64 (4.7%)
8	1893 (81.3%)	437 (18.7 %)
12	2884(**72.01**%)	1121 (27.99%)

Table 4.18 Summary of Correctly and Incorrectly Classified Instances at Different Number of News Categories using Bigram Phrase Structures

In the following section the second part of the experiment using trigram phrase structures will be discussed.

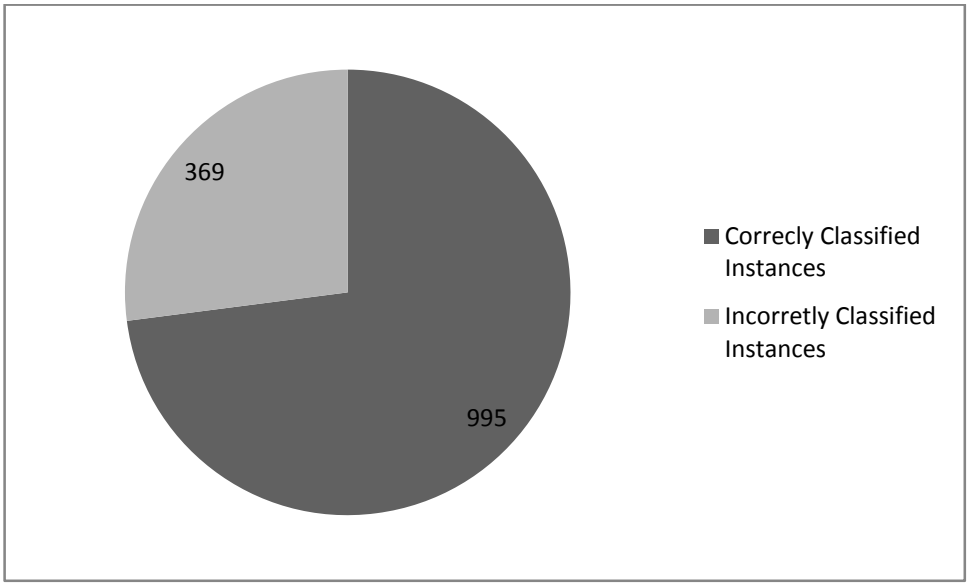
4.6.2 Model Building using Trigrams and Testing Classification Accuracy

The second part of the experiment was performed by using trigram phrase structures which constitute of three sequential words. This experiment like the previous one (i.e. bigram phrase structure) uses the same number of news categories and instances, i.e. the number of news categories, news class labels, and total number of instances per news categories that we implemented for bigram phrase structures experiment have been used here for trigram phrase structures. The only difference between the two experiments is in representing features.

Since the procedures are identical to the first experiment, this experiment also started with four news categories. These are the Accident (**አደጋ**), Culture and Tourism (**ባህልና ቱሪዝም**), Economy (**ኢኮኖሚ**) and Education (**ትምህርት**) news categories. These categories have a total of 1,364 instances (see Table 4. 9).

i. Experiment with Four News Categories

The following graph (Graph 4.4) shows the summary statistics of those news instances which are correctly classified and which are not.



Graph 4.4 Correctly and Incorrectly Classified Instances with Four Categories Using Trigram Phrase Structures

The experiment was performed using 10-fold stratified cross-validation like the first experiment (bigram phrase structure). For this experiment, the overall accuracy of the LibSVM classifier was found to be 995 (72.9%) correctly classified instances and 369 (27.1%) incorrectly classified instances out of the total of 1,364 instances.

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.475	0.078	0.633	0.475	0.543	0.699	**ACCIDENT**
0.989	0.281	0.717	0.989	0.831	0.854	**ECONOMY**
0.716	0.063	0.806	0.716	0.758	0.826	**EDUCATION**
0.205	0	1	0.205	0.34	0.602	**CULTURE**

Table 4.19 Four Category LibSVM Detailed Accuracy by Class Using Trigram Phrases

Observation of the detailed accuracy for each category in table 4.19 reveals that the SVM classifier performs better for those instances which have large number of instances. This is observed from the Economy and Education news categories. The classifier performance is poor for those categories with less number of instances of the Culture and the Accident news categories.

=== Confusion Matrix ===

ACCIDENT	ECONOMY	EDUCATION	CULTURE	Classified as
143	123	35	0	**ACCIDENT**
6	**564**	0	0	**ECONOMY**
37	67	**262**	0	**EDUCATION**
40	33	28	**26**	**CULTURE**

Table 4.20 Four Categories LibSVM Confusion Matrix using Trigram phrases

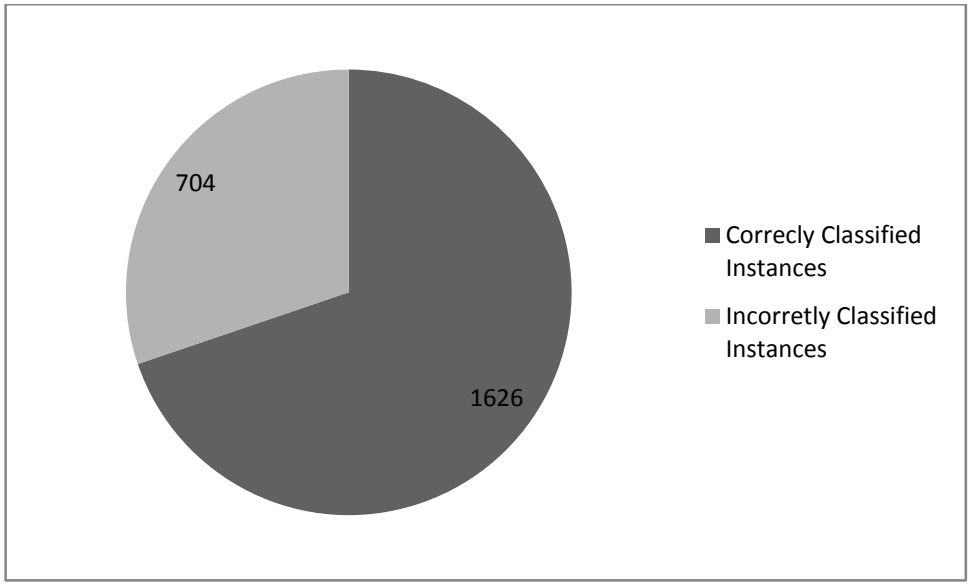
The confusion matrix shown in Table 4.20 confirms the detailed class statistical values used to determine the class level accuracy of the classifier. Hence, in Table 4.20, out of 301 instances of the Accident category, 143 instances are correctly classified as Accident while

the rest 123 and 35 instances are incorrectly classified to Economy and Education classes respectively. No instance is assigned to Culture from the Accident news class. Those number of instances correctly classified to their respective classes can be read diagonally with bold font style.

ii. Experiment for Eight News Categories

The eight news categories/classes taken for the experiment are shown in Table 4.12 with their number of instances.

For the eight news categories, the LibSVM classifier shows an average accuracy of 1626 (69.7854 %) from the total dataset of 2,330 instances with 704 (30.2146 %) classification errors which is observed in Graph 4.5 below.



Graph 4.5 Correctly and Incorrectly Classified Instances with Eight Categories using Trigram Phrase Structures

=== Detailed Accuracy by Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Classes
0.449	0.054	0.551	0.449	0.495	0.697	ACCIDENT
0.982	0.165	0.659	0.982	0.789	0.909	ECONOMY
0.443	0.049	0.625	0.443	0.518	0.697	EDUCATION
0.142	0	0.947	0.142	0.247	0.571	CULTURE
0.967	0.112	0.772	0.967	0.858	0.927	POLITICS
0.244	0.001	0.833	0.244	0.377	0.622	SCIENCE
0.511	0.007	0.86	0.511	0.641	0.752	SPORT
0.149	0	0.929	0.149	0.257	0.574	WEATHER

Table 4.21 Eight Category LibSVM detailed accuracy by Class using Trigram Phrases

Confirmation of the values used in the class detail accuracy is given by the confusion matrix for the eight categories in Table 4.24

=== Confusion Matrix ===

ACCIDENT	CULTURE	ECONOMY	EDUCATION	POLITICS	SCIENCE	SPORT	WEATHER	Classes
135	117	31	0	16	0	2	0	ACCIDENT
10	**560**	0	0	0	0	0	0	ECONOMY
36	67	**162**	0	94	0	6	1	EDUCATION
33	33	20	**18**	18	2	3	0	CULTURE
0	22	0	0	**636**	0	0	0	POLITICS
1	3	14	0	12	**10**	1	0	SCIENCE
16	23	9	0	40	0	**92**	0	SPORT
14	25	23	1	8	0	3	**13**	WEATHER

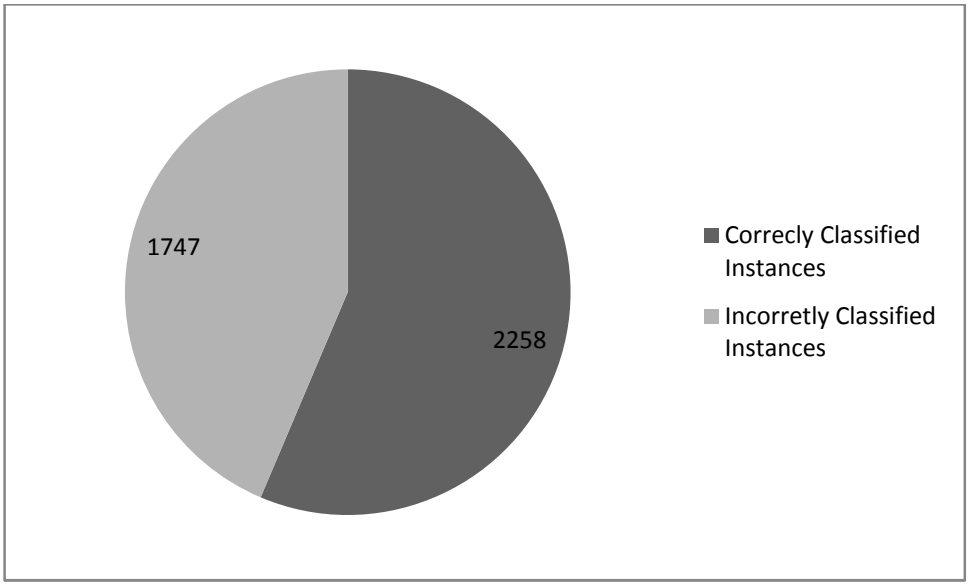
Table 4.22 Eight Categories LibSVM Confusion Matrix using Trigram Phrases

For the eight categories dataset, observation of the values of the TP-Rate, the Precision, the F-measure, and the ROC-Area in Table 4.22 shows that LibSVM classifier has better accuracy for categories with relatively larger instances in the dataset Economy and Politics. However, for those categories which have less number of instances the classifier shows less performance accuracy. This is observed from the confusion matrix (Table 4.22) in the Weather, Culture, and Science news categories/classes.

iii. Experiment for Twelve Categories

The final experiment using trigram phrase structures is for the whole news categories (12 categories). The twelve news categories with their number of instances used for classification are shown in Table 4.15.

From the total instances (i.e. 4,005) of the twelve news categories using trigram phrase structures, the LibSVM classifier showed an average accuracy of 56.4 % (2258) while 43.6 % (1747) instances incorrectly classified. This is shown in Graph 4.6 below.



Graph 4.6 Correctly and Incorrectly Classified Instances for Twelve Categories using Trigram Phrase Structures

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Classified as
0.05	0.016	0.203	0.05	0.08	0.517	**ACCIDENT**
0.024	0	1	0.024	0.046	0.512	**CULTURE**
0.865	0.001	0.914	0.865	0.889	0.932	**DEFENCE**
0.94	0.154	0.503	0.94	0.656	0.893	**ECONOMY**
0.224	0.014	0.612	0.224	0.328	0.605	**EDUCATION**
0.456	0.016	0.783	0.456	0.577	0.72	**HEALTH**
0.423	0.024	0.533	0.423	0.472	0.699	**LAW**
0.927	0.107	0.631	0.927	0.751	0.91	**POLITICS**
0.024	0	1	0.024	0.048	0.512	**SCIENCE**
0.623	0.191	0.497	0.623	0.553	0.716	**SOCIAL**
0.433	0.003	0.886	0.433	0.582	0.715	**SPORT**
0.069	0	1	0.069	0.129	0.534	**WEATHER**

Table 4.23 Twelve Categories LibSVM detailed accuracy by class using Trigram Phrases

Observation of the detailed statistics from the values of the TP-Rate, the Precision, the F-measure, and the ROC-Area reveals that for 12 categories (Table 4.23) dataset, the LibSVM classifier shows best performance for categories with relatively larger instances in the Economy and Politics news categories. The classifier's accuracy is least for categories with relatively small instances in the dataset i.e. Weather, Culture and Science news categories.

=== Confusion Matrix ===

The accuracy measures are calculated on the basis of the data in Table 4.24, for the confusion matrix of the 12 categories.

ACCIDENT	CULTURE	DEFENCE	ECONOMY	EDUCATION	HEALTH	LAW	POLITICS	SCIENCE	SOCIAL	SPORT	WEATHER	Classified as
15	0	0	92	10	10	27	17	0	127	0	0	**ACCIDENT**
2	**3**	0	28	7	2	4	10	0	71	0	0	**CULTURE**
0	0	**32**	0	1	0	0	0	0	4	0	0	**DEFENCE**
0	0	0	**536**	0	0	0	0	0	34	0	0	**ECONOMY**
7	0	0	65	**82**	23	10	80	0	99	0	0	**EDUCATION**
19	0	0	83	12	**209**	32	26	0	76	1	0	**HEALTH**
14	0	0	52	0	3	**104**	14	0	59	0	0	**LAW**
0	0	0	22	0	0	0	**610**	0	26	0	0	**POLITICS**
0	0	0	1	11	2	0	**8**	**1**	18	0	0	**SCIENCE**
13	0	0	137	4	16	16	162	0	**582**	4	0	**SOCIAL**
1	0	1	28	0	0	0	33	0	39	**78**	0	**SPORT**
3	0	2	21	7	2	2	7	0	35	2	**6**	**WEATHER**

Table 4.24 Twelve Category LibSVM Confusion Matrix using Trigram phrases

From the second experiment it can be observed that like the first experiment the LibSVM performs well for those news categories which have large number of instances. Secondly, with the increased number of news categories the performances of the classifier decreased using trigram phrase structures like that of the first experiments. This means that for both phrase structures (bigrams and trigrams) increasing the number of news categories will tend to decrease the classifier accuracy. This occurred because while the total number of features increased the probability of similarity of statistical values among features across different news categories will increase. As the similarity of statistical values increases for different features which belong to different news categories, the classifier categorizes instances incorrectly. This phenomenon will contribute for the decrement of the classifier while increasing the news categories.

Finally, the performance of the classifier using trigram phrases is poor for all news categories tested. For example, for the four news categories it was observed that only 995 (72.9 %) correctly classified while 369 (27.1) incorrectly from the total instances 1364. In the eight and twelve news categories the performance of the classifier was extremely poor with the following results. For the eight news categories 1626 (69.7%) correctly and 704 (30.3%) incorrectly classified instances were observed from the total instances of 2330. In the twelve news categories the result of the classifier was found to be 2258 (56.4 %) correctly classified instances and 1747 (43.6 %) incorrectly classified instances from all 4,005 instances. This is summarized in Table 4.25.

No. of News Categories	Correctly Classified Instances	Incorrectly Classified Instances
4	995 (72.9 %)	369 (27.1)
8	1626 (69.7%)	704 (30.3%)
12	2258 (**56.4** %)	1747 (**43.6** %)

Table 4.25 Summary of Correctly and Incorrectly Classified Instances at Different Number of News Categories using Trigram Phrase Structures

In the next section, the performance of the LibSVM classifier using two phrase structures (bigram phrases and trigram phrases) is discussed. Following that the appropriate phrase structure will be suggested for the automatic Amharic news text classification using phrase based feature representations.

4.6.3 Comparison of Bigrams and Trigrams in Amharic News Text Classification

Based on the results of the two experiments done above we will compare and contrast bigram and trigram phrase structures here under.

As we observed in the first experiment (bigram) of four news categories Accident (**አደጋ**), Culture and Tourism (**ባህልና ቱሪዝም**), Economy (**ኢኮኖሚ**) and Education (**ትምህርት**) with total instances of 1,364, the LibSvm classifier correctly classified 1300 (95.3079%) instances and 64 (4.6921%) incorrectly instances using bigram phrases. Whereas this same four news categories with the same number of instances were classified as 995 (72.9472 %) correctly classified instances and 369 (27.0528 %) incorrectly classified instances using trigram phrases.

When we increase the number of news categories to eight [Accident (**አደጋ**), Culture and Tourism (**ባህልና ቱሪዝም**), Economy (**ኢኮኖሚ**), Education (**ትምህርት**), Politics (**ፖለቲካ**), Science (**ሳይንስ**), Sport (**ስፖርት**) and Weather (**የአየር ፀባይ**)] with 2,330 total number of instances the performance of LibSVM using bigram phrases was 1893 (81.2446%)correctly classified and 437 (18.7554 %) incorrectly classified instances. Whereas with trigram phrases the performance of LibSVM decreased to 1626 (69.7854 %) correctly classified and 704 (30.2146 %) incorrectly classified instances.

Finally, with all twelve news categories of the Accident (**አደጋ**), Culture and Tourism (**ባህልና ቱሪዝም**), Economy (**ኢኮኖሚ**), Education (**ትምህርት**), Politics (**ፖለቲካ**), Science (**ሳይንስ**), Sport (**ስፖርት**), Weather (**የአየር ፀባይ**), Social (**ማህበራዊ**), Law (**ህግ**), Defense (**መከላከያ**), and Health (**ጤና**) with 4,005 total instances the performance of the LibSVM was analyzed for both bigrams and trigrams phrases. While 2884 (72.01%) of the 4005 instances classified correctly and 1121 (27.99%) incorrectly using bigram phrases, only 2258 (56.4 56%) instances were classified properly and 1747 (43.6 %) instances wrongly using trigram phrases.

The following table (Table 4.26) summarizes the experimental results for both phrase structures using different number of news categories discussed above.

No. News Categories	Category Name	No. Instances per Category	Total Instances considered	Phrase Structures used	Correctly Classified Instances	Incorrectly Classified Instances
Four	Accident	301	**1364**	bigram	1300 (95.3%)	64 (4.7%)
	Culture	127				
				trigram	995 (72.9 %)	369 (27.1)
	Economy	570				
	Education	366				
Eight	Above four	1364	**2330**	bigram	1893 (81.3%)	437 (18.7 %)
	Politics	658				
	Science	41				
				trigram	1626 (69.7%)	704 (30.3%)
	Sport	180				
	Weather	87				
Twelve	Above eight	2330	**4,005**	bigram	2884(**72.01**%)	1121 (27.99%)
				trigram	2258 (**56.4** %)	1747 (43.6 %)
	Social	934				
	Defense	37				
	Law	246				
	Health	458				

Table 4.26 Summary of Experimental Results using Bigram and Trigram Phrase Structures at Different Category Levels

Table 4.26 summarizes the results of the classifier at different number of news categories for both bigram and trigram phrase structures. From the result of the experiment, bigram phrase structures shows better classification accuracy than trigram phrase structures for the automatic Amharic news text classification system using phrase based approach. Moreover, for small number of news categories the classifier performance was better than those with large number of news categories. Conversely, with large number of news categories for both

bigram and trigram phrase structures the result of the classifier was relatively poor compared with small number of news categories.

CHAPTER FIVE

CONCLUSIONS AND RECOMANDATIONS

5.1 Conclusions

The volume of electronic documents within corporate archives and repositories are increasing exponentially from time to time in this information age. This huge availability of electronic documents is causing information overload, a major problem which currently faces many information processing organizations like Ethiopian News Agency (ENA). The issue in this information age is, therefore, how to use the best from this massive available information.

Various researches have been conducted in different contexts to devise methods which can enable to change threats into opportunities for wise use of information to counter act information overload. 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 information; as a result, automatic classification using *bag-of-words* feature representation is one selected area of research.

Following that a number of researches have been conducted on automatic processing of news texts for Amharic language using *bag-of-words* feature representation. However, using words as features could result in losing the intended meaning when the concept is created from two or more sequential words. Thus, in order to maintain this concept, a different approach from previous works have been proposed and implemented in this research. This approach is using bag-of-phrases (bigram and trigram phrase structures).

Preprocessing, feature representation, and testing were the major steps for the accomplishment of this study. Preprocessing the data is worked out before the datasets are fed into the classifier.

For feature preparation, because of the problems of the Amharic writing system and unavailability of Amharic document processing tool, the focus of the research was on developing tools which facilitates efficient automatic classification of Amharic documents.
To this end, much attention was given on the processing of the source data by developing the following tools:

- Amharic character (Fidel) controller
- stemmer that remove prefixes and suffixes
- stop word and number remover

Using such preprocessing methods, significant reduction of features (which enabled 10 to 20% feature reduction) is achieved. The final features are selected by applying dimension reduction technique called Document Frequency (DF) thresholding. The importance of features is determined using TFIDF weighting schemes.

In feature representations, two forms of phrase structures have been developed and tested. These are bigram and trigram features. After features are represented by bigram and trigram features and their weights are identified using the TFIDF schemes, phrase matrix have been generated and saved as CSV file format. The CSV files have been imported to the LibSVM classifier using the GUI of WEKA application package.

Finally, the experimental testing was performed using bigram and trigram phrase structures for four, eight and twelve news categories.

This study tried to see the potential application of SVM for the automatic classification of Amharic news texts using phrase based approach. Under this umbrella, effectiveness of news text classifier at increasing level of news categories using bigram and trigram phrase structures has been investigated. From this work on automatic classification of Amharic news texts using phrases, we can conclude the following points:

- The best accuracy using bigram phrase structures has been obtained at four categories, which is 95.3%. The overall accuracy using bigram phrases structures for the eight categories has been recorded as 81.3% and for all twelve categories it has been scored to 72.01%. On the other hand, for trigram phrase structures, the best accuracy recorded at four categories is 72.9 % and for eight categories it has been recorded as 69.7% and the least accuracy has been obtained at twelve categories that accounts to 56.4%.
- Using bigram phrase structures the LibSVM classifier has best classification performance at larger number of instances
- Misclassified news items are mostly from those news categories which have less number of instances
- In the course of training using LibSVM classifier, it is found that computational time increases as the number features and news categories increase.
- LibSVM has significantly better accuracy for features represented by bigram phrases than features represented by trigram phrases
- LibSVM classifier showed better accuracy for categories with relatively large number of instances for both bigram and trigram phrase structures

5.2 Recommendations

The results of this research showed that phrase based feature representations particularly using bigram phrase structures can be implemented to avoid semantic relations that exist between words. However, more research needs to be conducted to exploit the full potential of automatic news classification system.

The following are some of the areas identified in this research for future work.

- Highly customized tool for correcting character (Fidel) variations was developed in this research. Full-fledged Amharic spelling checker, which addresses the various causes of character variations need to be developed.
- In this research a number of tools have been developed to reduce word variants to their root form. These separate and incomplete efforts should be replaced by the availability of a complete Amharic stemmer.
- Even though phrases have high discriminating power in identifying one news categories from others, they are negatively affected by having less number of frequencies for less number of documents. As a result, their statistical values which are fundamental properties for the classifier will ultimately be affected. Therefore, combining bag-of-words and bigram-phrases should be tested by other researchers instead of words alone to get the best result for Amharic news text classification.
- The performance of the classifier at increasing number of news categories did not improve using phrase based approach. At increasing labels of news categories the classifier shows decreasing results as it was reported by other researchers using bag-of-words feature representation. Therefore, using different approaches to feature representation like *ontology* could be tested by future researchers.

REFERENCE

1. Androutsopoulos, I., et al. Learning to filter spam e-mail: a comparison of a Naive Bayesian and a memory-based approach, 2000
2. Atelach Alemu and Lars Asker. *Dictionary based Amharic-French Information Retrieval*. Department of Computer and System Science, Stockholm University, 2005
3. Atelach Alemu and Lars Asker: *Applying Machine Learning to Amharic Text Classification*, 2006
4. Baeza-Yates, R. and Ribeiro-Neto: *Modern Information Retrieval*. Addison- Wesley: New York, 1999
5. Benbrahim, and M. A. Barmer: *Neighborhood Exploitation in Hypertext Categorization*, In Research and Development in Intelligent Systems XXI, Springer-Verlag, 2005
6. Bender, et.al. The Ethiopian Writing System. In *Languages in Ethiopia*. London: Oxford University Press, 1976.
7. Blumberg and Atre, *Automatic classification: Moving to the Main Stream*, 2003
8. Boley et al.: *Supporting Document-Category Management: An Ontology-Based Document Clustering Approach*, 1999
9. Calvo, R.; Lee, J.; and Li, X.: *Managing Content with Automatic Document Classification. Journal of Digital Information*, 2004.
10. Caropreso, S.Matwin, F.Sebastiani, A learner-independent evaluation of the usefulness of statistical phrases in automated text categorization, Text database and Document Management: Theory and Practice, 2001.
11. Cheng, and Wu. ACS: *An Automatic Classification System. Journal of the American Society for Information Science*. 21(4):289-299, 1995.

12. Crawford **:***Phrases and Feature Selection in E-Mail Classification*, 2006

13. Duda R. and Hart P.*Pattern Classification and Scene Analysis*, Wiley, New York, 1973

14. ENA. *Problems of the Existing News Classification System and Suggested Solutions*. Unpublished, July 2006.

15. Enser, P. G. B. *Automatic Classification of Book Material Represented by Back-of the Book Index*. *Journal of Documentation*. 41(3):135-155, 1985

16. Fürnkranz: *A Study Using n-gram features for Text Categorization.* Technical Report OEFAI-TR-98-30, Austrian Research Institute for Artificial Intelligence, Vienna, Austria1, 1998.

17. Getachew Haile. *The Problems of the Amharic Writing System.* Unpublished, 1966.

18. Giorgino, T. *An Introduction to Text Classification*, 2004.

19. Haddis Alemayehu, **ፍቅር እስከ መቃብር**, 1958

20. J. Han and M. Kamber. Data Mining: Concepts and techniques (2nd ed.). Morgan Kaufmann Publishers, 2006.

21. J. L. Fagan. *Experiments in automatic phrase indexing for document retrieval*: a comparison of syntactic and non-syntactic methods. PhD thesis, Department of Computer Science, Cornell University, Ithaca, US, 1987.

22. J.P.Lewis, *Tutorial on SVM*, CGIT Lab, USC, 2004.

23. Joachim, Thorsten. *A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. In Proceedings of the 14th International Conference on Machine Learning ICML97*. pp. 143-151, 1996.

24. Klein, B. *Text Categorization or Classification*, 2004

25. Kwok, J. L. *The Use of Title and Cited Titles as Document Representation for Automatic Classification*. *Information Processing and Management*. 2:201-206, 1975.

26. Leslau, Wolf: *An Amharic Text Book of Everyday Usage*. University of California, Los Angles, 1965.

27. Lewis. *An evaluation of phrasal and clustered representations on a text categorization task*. In N. J. Belkin, P. Ingwersen, and A. M. Pejtersen, editors, Proceedings of SIGIR-92, 15th ACM International Conference on Research and Development in Information Retrieval, pages 37–50, Kobenhavn, DK,. ACM Press, New York, US., 1992

28. Liao, C.; Alpha, S. and Dixon, P.: Feature Preparation in Text Categorization, 2003

29. Losee, R. M. and Hass, S. W. Sublanguage Terms: Dictionaries, Usage, and Automatic Classification. *Journal of the American Society for Information Science.* 46(7):519- 529, 1995.

30. Luhn, H. *The automatic creation of literature abstracts*. IBM Journal of Research and Development, PP159–165, 1958.

31. M. F. Caropreso, S. Matwin, F. Sebastiani. *Statistical Phrases in Automated Text Categorization*, 2001

32. Mladenic, D. and Grobelnik, M. *Word sequences as features in text learning*. In *Proceedings of the 17th Electrotechnical and Computer Science Conference (ERK-98)* (pp. 145- 148), Ljubljana, Slovenia, 1998

33. Nega Alemayehu and Peter Willett. Stemming of Amharic Words for Information Retrieval. Literary Linguistic Computing Vol. 17, No.1, 2002.

34. Nello C.and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, 2000.

35. P.N. Tan, M. Steinbach, and V. Kumar: *Introduction To Data Mining*, Pearson Education, Inc, 2006

36. S. T. Dumais, J. Platt, D. Heckerman, and M. Sahami. *Inductive learning algorithms and representations for text categorization*. In G. Gardarin, J. C. French, N. Pissinou, K. Makki, and L. Bouganim, editors, Proceedings of CIKM-98, 7th ACM International Conference on Information and Knowledge Management, pages 148–155, Bethesda, US,. ACM Press, New York, US, 1998

37. Salton, Gerard. 1989. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Reading: Addison-Wesley Publishing Company.

38. Schapire, R, Singer, Y., and Singhal, A. *Boosting and Rocchio Applied to Text Filtering*. In Croft *et. al*. (Ed.), *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval* (pp. 215-223). New York: ACM Press, 1998

39. Scott, Matwin, Feature engineering for text classification, Proceedings of ICML-99, 16th International Conference on Machine Learning, 1999.

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

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

42. Sebastiani, Fabrizo. *Machine Learning in Automated text classification*. In ACM Computing Surveys. Vol.34NO.1, pp1-47, 2002.

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

44. Surafel Teklu: *Automatic Categorization Of Amharic News Text: A Machine Learning Approach,* Master Thesis, Addis Ababa University, 2003

45. Vapnik. *The Nature of Statistical Learning Theory*. Springer, Berlin, 1995.

46. Wang, Y, Zang, H., Spencer, B. and Yan, Y. *A Text Categorization Approach for Match-Making in Online Business Tendering*. Journal of business and technology, 2005.

47. Worku Kelemework: *Automatic Amharic Text News Classification*: A Neural Networks Approach**,** Master Thesis, Addis Ababa University, 2009

48. Y. Yang. *An Evaluation of Statistical Approaches to Text Categorization.* The Netherlands: Kluwer Academic Publishers, 1999.

49. Yohannes Afework. *Automatic Amharic Text Categorization*. Thesis. Addis Ababa University: Addis Ababa, 2007

50. Zelalem Sintayehu. *Automatic classification of Amharic news items*: The case of Ethiopian News Agency., Master Thesis, 2001

51. Zhixu Li, Pei Li, Wei, Hongyan Liu, Jun He, Tao Liu and Xiaoyong Du AutoPCS: *A Phrase-Based Text Categorization System for Similar Texts*, Springer LNCS 5446, pp. 369-380, 2009

Appendix 1: Amharic Characters ('Fidel') (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	ቹ ču	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	ኩ ku	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	ኹ hu	ኺ hi	ኻ ha	ኼ he	ኽ h	ኾ ho					
ዘ zä	ዙ zu	ዚ zi	ዛ za	ዜ ze	ዝ z	ዞ zo			ዟ $z^{w}a$		
ዠ žä	ዡ žu	ዢ ži	ዣ ža	ዤ že	ዥ ž	ዦ žo					
የ yä	ዩ yu	ዪ yi	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 $g^{w}ə$
ጘ dä	ጙ du	ጚ di	ጓ da	ጔ de	ጕ d	጖ do			጗ $d^{w}a$		
ጠ ṭä	ጡ ṭu	ጢ ṭi	ጣ ṭa	ጤ ṭe	ጥ ṭ	ጦ ṭo			ጧ $ṭ^{w}a$		
ጨ ćä	ጩ ću	ጪ ći	ጫ ća	ጬ će	ጭ ć	ጮ ćo			ጯ $ć^{w}a$		
ጸ şä	ጹ şu	ጺ şi	ጻ şa	ጼ şe	ጽ ş	ጾ şo			ጿ $ş^{w}a$		
ፀ şä	ፁ şu	ፂ şi	ፃ şa	ፄ şe	ፅ ş	ፆ şo					
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 2: Amharic Punctuation Marks Unicode

No.	Punctuation mark	Symbol	Purpose
1	The four dots or Full stop	።	Mark end of a sentence
2	Colon	፥	Separate words in a sentence: which not common
3	word space	፡	Separate words in a sentence: however, the current practice is being white space
4	Question mark	?	Placed at the end of questions
5	Exclamation mark	!	Used at the end of sentences that show exclamation
6	Comma	፣	Used to separate word lists
7	Semi-colon	፤	Used like semi-column
8	Preface Colon	፦	Used
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.

Appendix 3: 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 4: List of Stop words

1 ነዉ
2 በያዝነው

3 ሀዋሳ
4 ነው
5 አንድ
6 በአንድ
7 አስታወቁ
8 ጋራ
9 ዒስጥ
10 ሁለት
11 ሶስት
12 ዛሬ
13 እንደገለጹት
14 እየሰጠ
15 ዲክተር
16 እያካሄደ
17 መሆኑ
18 ሀዋሳ
19 አዋሳ
20 ያለ
21 እንደገለፁት
22 ነበር
23 ገለጡ
24 ገለፁ
25 ገለጹ
26 ተገለጠ
27 ተገለፀ
28 ሀዋሳ
29 ተገለጸ
30 ተባለ
31 መቶ
32 አምስት
33 እስከ
34 ዕስከ
35 ይፋ
36 ታዬ
37 ታየ
38 አስታወቁ
39 አስገነዘቡ
40 ዋለ
41 መካከል
42 ምክንያት
43 ሰባት
44 ሠባት
45 አንዳንዴ
46 መጀመሪያ
47 ዘተኛ
48 ያህል
49 ነገ
50 ትናንት
51 መላ
52 ዙሪያ
53 አሳዬ
54 አሣዬ
55 ልክ
56 ዲላ
57 ባሕርዳር
58 ድርጅቶች
59 አስር
60 አሥር
61 ዐስር
62 ዐሥር
63 አለ
64 ደብረብርሃን
65 ደብረብርሀን
66 ደብረብርሐን
67 ቢንሻንጉል
68 ዕኔ
69 እኔ
70 ስምንት
71 ሥምንት
72 ያልሆኑ
73 የሆኑ
74 የተባለ
75 ጉጂ
76 ሰሜናዊ
77 ሠሜናዊ
78 አለ
79 ስል
80 ሥል
81 ረገድ
82 የሚል
83 የሚለ
84 የሚባለ
85 ጉምዝ

86 አሶሳ
87 ሲካሄድ
88 ሲካሔድ
89 ሃያ
90 ሀያ
91 ባዶ
92 ጉራጌ
93 ጎንደር
94 ባለፈ
95 ባለፈው
96 ሮቤ
97 ሶዶ
98 ወላይታ
99 ያስፈልጋል
100 አብዛኛው
101 ብዙሃን
102 ብዙሃኑ
103 ብዙሐን
104 ዋ
105 ጎባ
106 ባሌ
107 እስካሁን
108 ዕስካሁን
109 ሩብ
110 አሰላ
111 ዐሥላ
112 አሥላ
113 አዋሳ
114 አዋሣ
115 መንዝ
116 ያለ
117 እሁድ
118 ዕሑድ
119 ዕሁድ
120 እሑድ
121 ሰኞ
122 ማክሰኞ
123 አሮብ
124 ዕሮብ
125 ነገሌ
126 ቦረና
127 ቦሌ
128 ብቻ
129 ዕለት
130 እለት
131 ይጠበቅባቸዋል
132 አሳሰቡ
133 አሣሠቡ
134 አሣሰቡ
135 መሆን
136 መካሄድ
137 አብዛኞች
138 አብዛኛዎቹን
139 ትልቁ
140 ትንሹ
141 ትላልቅ
142 ትልቅ
143 ትንሽ
144 ነቀምት
145 ነቀምቴ
146 ቤንቺ
147 ማጂ
148 ጂጂጋ
149 ጅጅጋ
150 በኩል
151 ደረስ
152 ሻሸመኔ
153 ሸሬ
154 ማይጨው
155 አዳ
156 ሚኒስቴር
157 ጠቅላይ
158 ነበር
159 ይሆናል
160 እንዲሁ
161 ዕንዲሁ
162 ጌዲዮ
163 ኤልባቡር
164 ጎጃም

165 እንዳለ
166 ታወቀ
167 ታውቀዋል
168 ይቀጥላል
169 ቀጥሎ
170 ሲካሄድ
171 አካሄዱ
172 ጎፋ
173 ጋሞ
174 ጋሞጎፋ
175 ቡርቃ
176 አካሄደ
177 ተካሄደ
178 ተካሔደ
179 መጡ
180 አርባምንጭ
181 አካባቢ
182 በሙሉ
183 ሙሉ
184 አላማጣ
185 ብዙ
186 አስፈላጊው
187 አስፈለገ
188 ነጭ
189 ሐድያ
190 ሃድያ
191 ሀድያ
192 ዜሮ
193 ዎልቂጤ
194 ውልቂጤ
195 ደባርቅ
196 ሰመራ
197 ሠመራ
198 ሁመራ
199 ሁመራ
200 ሀረር
201 ሐረር
202 ጎዴ
203 ደጋሐቡር
204 ደጋሃቡር
205 ደጋሀቡር
206 በኩላቸው
207 ዕጥፍ
208 እጥፍ
209 መተማ
210 ኮምቦልቻ
211 ደሴ
212 መቀሌ
213 እንደስላሴ
214 ዕንደስላሴ
215 እንደሥላሴ
216 ዕንደሥላሴ
217 ያካሂዳሉ
218 ታውቃል
219 ሲዳማ
220 መስክ
221 ምን
222 አዋለ
223 ጊምቢ
224 የሚሆን
225 የሚሆኑ
226 የሚያህል
227 የሚያህሉ
228 አርማጭሆ
229 ዎዲህ
230 ወዲህ
231 ሁለቱ
232 ያልሆነ
233 እያካሄዱ
234 ነበረው
235 ነበሩ
236 መትሃራ
237 መትሓራ
238 አዴግራት
239 ወልድያ
240 ዳባት
241 ናዝሬት
242 አዳማ
243 ከፋ

244 ጂማ
245 አጋሮ
246 ሚዛንቴፒ
247 ደበረታቦር
248 አለምያ
249 ሐሮምያ
250 ሐሮሚያ
251 ደንቢያ
252 ሜጫ
253 ቡሬ
254 ደብረማርቆስ
255 አዊ
256 ዞን
257 ወረዳ
258 ክልል
259 አሁን
260 ቀጣይ
261 በሚቀጥሉት
262 በሚቀጥለው
263 ወራት
264 ወሮች
265 አመት
266 አመታት
267 ቅዳሜ
268 ሐሙስ
269 ሃሙስ
270 ሀሙስ
271 አርብ
272 ኤሊባቡር
273 ኮንሶ
274 ወሊሶ
275 ሰቆጣ
276 ሞጣ
277 ደጀን
278 ወረዳ
279 ቀበሌ
280 የተውጣጡ
281 ምንጊዜም
282 ቢሆን
283 ቢሆንም
284 ባይሆንም
285 ከተቻለ
286 መጠን
287 ሁሉንም
288 ሆነው
289 ሆኖ
290 በሚል
291 በተባለ
292 በሚባል
293 በተባሉት
294 በሚባሉት
295 ሊያካሂዱ
296 ያካሂዳሉ
297 ዳሞት
298 ደጋዳሞት
299 ፉነተሰላም
300 ፍኖተሰላም
301 ፍኖተሠላም
302 አቸፈር
303 አዋበል
304 ዝዋይ
305 ፈፅሞ
306 ተፈጸመ
307 ተፈፀመ
308 ተጠናቀቀ
309 ይጠናቀቃል
310 ሣምንት
311 ሳምንት
312 ሳምንታት
313 ሣምንታት
314 ወራቶች
315 በሚቀጠሉት
316 አመታቶች
317 ወራቶች
318 መቄት
319 ሆነው
320 ግለሰብ

321 ግለሰቦች
322 ሲለ
323 አስታወቁ
324 አስታውቀዋል
325 ነጥብ
326 ያህለ
327 ያሕለ
328 ሞጆ
329 ሶስቱ
330 ሰበታ
331 ከነማ
332 ፊንታሌ
333 ናት
334 ናቸው
335 ተባለ
336 ተባለ
337 ይባላል
338 ይባላለ
339 ተያይዞ
340 ይቀርባል
341 ቀረበ
342 በሚሆን
343 በሚሆኑት
344 በሚሆን
345 ጊዜ
346 ተፈላጊ
347 አስተነቀቀ
348 አሳወቀ
349 መሆኑን
350 ተሰማ
351 ተቀሰቀሰ
352 ተሠማ
353 ተቀሠቀሠ
354 አሰፈላጊ
355 ለአንድ
356 ለሁለት
357 ለእያንዳንዱ
358 ለእያንዳንዳቸው
359 በሙለ
360 ተደረገ
361 እየተደረገ
362 ይደረጋል
363 ተመለሰ
364 ተመለሠ
365 ይመለሳለ
366 ገቡ
367 ገባ
368 ይገባል
369 ይገባለ
370 ሰጠ
371 ተጀመረ
372 ይጀመራል
373 ሊጀመር
374 ይችላል
375 ወደ
376 ይመጣለ
377 መጣ
378 ወዲያ
379 ከነገ
380 ከዛሬ
381 ጋምቤላ
382 ሆሳሕና
383 ሆሣሀና
384 በነገለ
385 በአዳማ
386 በአዋሳ
387 በአዋሣ
388 ሕዳር
389 መስከረም
390 ጥቅመት
391 በመጪው
392 ታሕሳስ
393 ታህሳስ
394 በአብዛኛው
395 ካለ·ት
396 ከተባለ·ት
397 ገሚሱ
398 መጀመሪያው
399 መጀመሪዎች
400 መጀመሪያዎቹ

401 መጨረሻ
402 መጨረሻዎች

403 መጨረሻዎቹ
404 ተነገረ
405 ተናገሩ
406 ይነጋገራሉ
407 ተወያዩ
408 ይወያያል
409 ይወያያሉ
410 ይጠበቃል
411 እንደሚሆን
412 እንደሚሆኑ
413 ተገቢ
414 ይገባል
415 ተሰጠ
416 ተሰጡ
417 ጀመሩ
418 ጀመረ
419 ተጀመረ
420 ሊሰጥ
421 ተሰጡ
422 ይሰጣል
423 ይሰጣሉ
424 እየሰጡ
425 ተከታተሉ
426 እየተከታተሉ
427 እየተደረገ
428 ሰጠ
429 አገኙ
430 አገኘ
431 በላይ
432 በታች
433 ቀነሱ
434 እየቀነሰ
435 እየቀነሠ
436 መጣ
437 ቃል
438 ተገባ
439 ጠየቁ
440 ተጠየቀ
441 ሊጠየቅ
442 ሊጠናቀቅ
443 ተጠናቀቀ
444 ይጠነቀቃል
445 ተመቻቹ
446 ሊመቻች
447 ይመቻቻል
448 ተጣለ
449 ታዬ
450 ይታያል
451 አበረታች
452 መቀጠል
453 ተጠናክሮ
454 ሊሆን
455 ተግባራዊ
456 ይተገበራል
457 መተግበር
458 ይጠበቅባቸዋል
459 ይቻላል
460 እንደሚቻል
461 አሳይታለች
462 ታሳያለች
463 ውጤቶች
464 መታየት
465 ጀምረዋል
466 ይዛ
467 ሀረሪ
468 ሁመራ
469 አሳይታ
470 ጊንጪ
471 ማጂ
472 ቤንቺማጂ
473 ሊቀጠሉ
474 ተከናወነ
475 ይከናወናል
476 እንደሚከናወን
477 እየተከናወነ
478 እየተከናወኑ
479 መሆኑን
480 ሊጀምሩ
481 ደረሰ
482 ታስተናግዳለች

483 እንደምታስተናግድ
484 ገለጹ
485 ሰፋ
486 እየሰፋ
487 መምጣቱ
488 በቂ
489 በበቃት
490 ተቻለ
491 ተገናኙ
492 ይገናኛሉ
493 እንደሚገናኙ
494 መጠን
495 ሁኔታ
496 ሁኔታዎች
497 ተከፈተ
498 ተከፈቱ
499 ተመርቀው
500 ተመረቀ
501 አስመረቁ
502 አስመረቀ
503 ያስመርቃል
504 ይመረቃሉ
505 ተብሎ
506 ይነገራል
507 ግለጽ
508 ግልፅ
509 ይገለፃል
510 ተሳተፉ
511 ተሳተፈ
512 ይሳተፋሉ
513 ተሳታፊዎች
514 እንደሚሳተፉ
515 ተሳታፊዎች
516 ታወቁ
517 ይታወቀሉ
518 እንደሚታዎቅ
519 እነደሚታዎቁ
520 ይከፈታል
521 ዝግጅት
522 ተዘጋጀ
523 ይዘጋጃል
524 አዘጋጁ
525 ተዘጋጁ
526 ይዘጋጃሉ
527 እነደሚዘጋጅ
528 እንደሚያዘጋጁ
529 እነደሚታወቅ
530 ያሳውቃል
531 ያሳውቃሉ
532 ይታወቃል
533 በቅርብ
534 በቅርቡ
535 አቅራቢያው
536 በአቅራቢያቸው
537 ወደሚገኝ
538 ድርጅቱ
539 ተደራጁ
540 ተደራጅተው
541 መደራጀታቸው
542 ተጠቀሚ
543 ቢቢሲ
544 ኢዜአ
545 ሰሜን
546 ደቡብ
547 ምዕራብ
548 ምስራቅ
549 በሰሜን
550 በደቡብ
551 በምዕራብ
552 በምስራቅ
553 ሀገር
554 ሃገር
555 ሀገሮች
556 በሀገሪቱ
557 በሃገሪቱ
558 ከሁለት
559 ተሰበሰበ
560 ይሰበሰባል

561 ይገኛል
562 ተገኘ
563 መገኘቱ
564 አመላካች
565 በአንድ
566 ከአንድ
567 የማይጠበቅ
568 ይጠበቃል
569 ትውልድ
570 ከቀጠዩ
571 ብዙ
572 እንደሚጠበቅ
573 አረጋገጠ
574 አወጡ
575 ተደመደመ
576 በዚሁ
577 በዚህ
578 አይነት
579 አካሄድ
580 እንደማይቻል
581 አስገነዘቡ
582 ወይዘሮ
583 ተመሰረተ
584 ይመሰረታል
585 ክፈለከተማ
586 አራዳ
587 ይመሰረታል
588 ሊመሰረት
589 ሃሳብ
590 ሃሳባቸውን
591 እንደሚሆን
592 ይታያል
593 መስጠት
594 አሳዩ
595 አነጋገሩ
596 ዋና
597 ጽሐፊ
598 ፕሬዝዳንት
599 ግርማ
600 ወልደጊዮርጊስ

601 ቀዳሚ
602 ሺህ
603 ተቀብለው
604 ተቀበሉ
605 ይቆማል
606 እነደሚቆም
607 ከነናቸው
608 ጥረቱ
609 ጥረት
610 አገኘች
611 እንደሚገኚ
612 እንደሚገኝ
613 አስታወቀ
614 ይደረጋል
615 ይደረጋሉ
616 እንደሚደረግ
617 ጋር
618 ያለው
619 ያላት
620 አዝማሚያ
621 ተካተተ
622 ታካተዋል
623 እቅድ
624 እቀዳቸው
625 ላይ
626 ያስፈልጋል
627 ተከናውነዋል
628 መንገዝ
629 ምንጃር
630 አጠናክራ
631 እንደምትቀጥል
632 ይቀንሳል
633 ሊደረግ
634 ርብርብ
635 መረባረብ
636 ተመለከተ
637 አመለከተ
638 አዜብ
639 መስፍን
640 አዲሱ

641 ለገሰ
642 ሚኒስትሩ

643 ሚኒሰቴር
644 መስሪያ
645 ጥንቃቄ
646 ሊደረግ
647 እንደሚገባ
648 አፀደቀ
649 አፀደቁ
650 ጸደቀ
651 ፀደቀ
652 አጸደቀ
653 አጸደቁ
654 ይኖርበታል
655 እንደሚኖርበት
656 እንደሚኖርባቸው
657 መግለጹ
658 መግለፁ
659 ለመስራት
660 እንደሚሰራ
661 አለበት
662 ይረዳል
663 እየሰራች
664 ተግባራት
665 ለማከናወን
666 ሲከናወን
667 መቆየቱ
668 እንደሚችል
669 አመርቂ
670 ለውጦች
671 ለውጥ
672 የተሻለ
673 ተፈረመ
674 ተፈራረሙ
675 ይፈራረማሉ
676 እንደሚፈረም
677 ይፈረማል
678 ፈረሙ
679 ይፈርማሉ
680 ማሳሰቡን
681 ተፈራረመ
682 ፈረመች
683 ባለበት
684 እንደሚኖርበት
685 ይኖርባቸዋል
686 አጥጋቢ
687 ለጀመረው
688 ለተጀመረው
689 የተጀመሩት
690 ሊጠናቀቅ
691 ይቀረዋል
692 ቀርቶታል
693 በማድረግ
694 በሚካሄደው
695 በሚያካሂዱት
696 ሕዝበ
697 ህዝበ
698 ማቃለል
699 ይቃለላል
700 ተቃለለ
701 ይገባዋል
702 ምቹ
703 ሁኔታን
704 ይፈጠራል
705 እንደሚፈጠር
706 ማሳሳቡን
707 ይጠበቃሉ
708 ማድረግ
709 እንደሚገባው
710 እንደሚገባቸዉ
711 ይገባቸዋል
712 ለጀመረው
713 መነቃቃት
714 ስምንተኛ
715 ሠባተኛ
716 መደበኛ
717 አንደኛ
718 ሁለተኛ
719 ሶስተኛ
720 አራተኛ
721 አምስተኛ
722 ስድስት
723 ስድስተኛ

724 ዘጠነኛ
725 ባለፈው
726 ሚኒሰቴሩ
727 ነዋሪዎች
728 ሚና
729 ተጨወተ
730 ይጫወታል
731 አስመዘገበች
732 ተኮር
733 አደረጉ
734 ቁጥር
735 ተመቻቹ
736 ተዘረጉ
737 ተዘርግተዋል
738 ሰጠ
739 ተያያዘ
740 ተያያዙ
741 ጉዳዮች
742 መልስ
743 ክርክር
744 የሚስችለ
745 ያስችላል
746 ይደርሳል
747 ስለ
748 በየጊዜው
749 የተከናወኑት
750 ያከናወናቸው
751 ሽዋ
752 ዲስትሪክት
753 ክፍለሀገረ
754 ሀገራት
755 ሃገራት
756 ክፍለሃገር
757 ትኩረት
758 ሰጥቶ
759 ይሰራል
760 እንደሚሰራ
761 ዘመቻ
762 አንጥቦባቸዋል
763 የሚውለ
764 የሚመክር
765 የሚሆን
766 ለተጠቃሚዎች
767 ተጠቃሚ
768 ተጠቃሚዎች
769 እያዘጋጀ
770 እይተዘጋጀ
771 ያተኮረ
772 ማተኮር
773 እንደሚያከፋፍል
774 አከፋፈለ
775 ገዝቶ
776 ግዥ
777 መዘርዝር
778 ቀረበ
779 ሊቀርብ
780 ተረከበ
781 ቁሳቁስ
782 ሊዘረጋ
783 ሊጀምር
784 አዘጋጅቶ
785 መንቀሳቀስ
786 መዘጋጀቱን
787 መአጋጀታቸውን
788 መገኘት
789 ለማደረግ
790 ሊያደርግ
791 ለሚወስዱ
792 ሚወስዱ
793 ቀርፆ
794 የሚያድርጉ
795 እየተንቀሳቀሰ
796 ካለት
797 በሊበን
798 ሊገነባ
799 እንቅስቃሴ
800 ንሃሴ
801 ንሀሴ
802 ሐምሌ

803 ሀምሌ
804 ሰኔ
805 ግንቦት
806 ሚያዚያ
807 ጥር
808 በቁ
809 ይበቃሉ
810 ዝግጁ
811 በሚበልጥ
812 ሚበልጥ
813 የሚያድርጉ
814 ሚያድርጉ
815 ይጠናቀቃል
816 ይገመታል
817 ያቋቋማል
818 ይቋቋማሉ
819 ተቋቋሙ
820 እንደ
821 እየ
822 አኖሩ
823 የተጀመረዉ
824 የተሰራዉ
825 የተገነባዉ
826 የተገነባው
827 ለሚገነባው
828 በለሳ
829 ተቀመጠ
830 አስቀመጡ
831 አያሌዉ
832 ተጠናቋል
833 እየተጠናቀቀ
834 ያለዉ
835 እየተፋጠነ
836 ይጠናቀቃሉ
837 መካሄድ
838 መፋጠን
839 ዙር
840 ለማጠናቀቅ
841 እቅድ
842 አደረገ
843 ጥሪ
844 አቀረቡ
845 ድሬዳዋ
846 ተገነባው
847 ሊሠራ
848 ወጪ
849 ተጣለ
850 ተዘጋጅተዋል
851 ሰለጠኑ
852 ደርሻ
853 ይቀበላል
854 ተቀበለ
855 እንደሚቀበል
856 ሊቀበል
857 ለመጀመሪያ
858 የሚሆኑ
859 ያስገነባቸዉ
860 አዳዲስ
861 ማስመዝገቧን
862 ሚኒስትር
863 ጥሪዎች
864 መስጠታቸውን
865 ያስፈልጋሉ
866 ሊሰጡ
867 ጠቃሚ
868 አስመዘገቡ
869 እንደሚያስመዘግቡ
870 ተግባራዊ
871 ይደረጋል
872 ተሸጋገረ
873 ተይዟል
874 እያሻሻለ
875 እንደሚገኝ
876 ሰዓት
877 ሆኗል
878 መገባቱን
879 ሰበሰበ
880 ተዘጋጅቷል

881 መንደፉን

882 ታቅዷል

883 ታቀደ

884 ሆነዋል

885 መሆናቸዉን

886 መሆናቸውን

887 ተሸላሚ

888 ያከናውናል

889 እየሆነ

890 መጥታል

891 እያገዙ

892 አስረከበ

893 ቀዳማይት

894 ሆነች

895 ሊሆን

896 ሊሆኑ

897 ልትሆን

898 ተመዘገበች

899 ዘመን

900 ሰሞኑን

901 ዜና

902 ገለፀች

903 ያደረገችው

904 ተወደሰ

905 አለው

906 ይዟል

907 ተያዘ

921 በነፍስ ወከፍ

922 አዲስ አበባ

923 ቤንቺ ማጂ

924 ክፈለ ከተማ

925 ክፍለ ሀገር

926 ክፍለ ሃገር

927 ተዘጋጅቶ

928 ለውይይት

929 ቁሳቁሶችን

930 መሰጠት

931 ክትትል

932 እንዲያደርግ

933 ወደተግባር

934 አሰላጡን

935 ትግበራ

936 በተጠናከረ

937 መልኩ

938 ዓ.ም

939 ጠ/ይ

940 ወ/ሮ

941 ዶ/ር

942 1ኛ

943 1ብር

944 1ሚሊዮን

945 1ቢሊዮን

946 ከ1ሺህ

947 10ኛ

908 መያዙን

909 መደበ

910 ተመደበ

911 ቀን

912 ባህር ዳር

913 በሕር ዳር

914 ደብረ ብርሃን

915 ደብረ ብርሀን

916 ደብረ ብርሓን

917 ሚዛን ቴፒ

918 አሰሸ ተፈሪ

919 አሰበ ተፈሪ

920 ፈጽሞ

948 1ለ0

949 ከ10ሺህ

950 100ኛ

951 ከ1000

952 11ኛ

953 ከ11ሺህ

954 16ኛ

955 2ኛ

956 2ሚሊዮን

957 2ቢሊዮን

958 ከ2ሺህ

959 2ለ0

960 2ለ1

961 2ለ1

962 25ኛ

963 3ኛ

964 3ሚሊዮን

965 3ቢሊዮን

966 ከ3ሺህ

967 3ለ0

968 3ለ2

969 4ሚሊዮን

970 ከ4ሺህ

971 5ኛ

972 5ቢሊዮን

973 ከ5ሺህ

974 50ኛ

975 6ኛ

976 ከ6ሺህ

981 ከ8ሺህ

982 9ኛ

983 9ብር

984 ከ9ሺህ

981 ከ8ሺህ

982 9ኛ

983 9ብር

984 ከ9ሺህ

977 7ኛ

978 ከ7ሺህ

979 8ኛ

980 8ብር

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

PHRASE BASED AMHARIC NEWS TEXT CLASSIFICATION

BY

ZELEKE ABEBAW

APPROVED BY
EXAMINING BOARD:

LEMMA LESSA (M.Sc.),
CHAIRPERSON_____
WONDWOSSON MULUGETA (M.Sc.), ADVISOR_____
ERMIAS ABEBE (M.Sc.), EXAMINER_____

DECLARATION

THIS THESIS IS MY ORGINAL WORK AND HAS NOT BEEN SUBMITTED FOR DEGREE IN ANY OTHER UNIVERSITY AND ALL SOURCES OF MATERIALS USED FOR THE THESIS HAVE BEEN DULY ACKNOWLEDGED.

_____

ZELEKE ABEBAW KASSA

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

_____

WONDWOSSON MULUGETA (M.Sc.)