# JIMMA UNIVERSITY

## COLLEGE OF NATURAL SCIENCE

## DEPARTMENT OF INFORMATION SCIENCE

# AUTOMATIC TEXT CATEGORIZATION FOR AFAN OROMO NEWS:

# MACHINE LEARNING APPROACH

BY:

Mubarak Taha

January, 2015

Jimma Ethiopia

# AUTOMATIC TEXT CATEGORIZATION FOR AFAN OROMO NEWS: MACHINE LEARNING APPROACH

BY:

Mubarak Taha Adem

Advisor: DebelaTesfaye (Assi.Professor)

Co-Advisor: DulaBoru (MSc)

A Thesis Submitted to the College of Natural Science of Jimma University in Partial Fulfillment of the Requirement for the Degree of Master of science in Electronic and Digital Resource Management

**January, 2015**

**AUTOMATIC TEXT CATEGORIZATION FOR AFAN OROMO NEWS:**

**MACHINE LEARNING APPROACH**

BY:

Mubarak Taha Adem

A Thesis Submitted to the College of Natural Science of Jimma University in Partial Fulfillment of the Requirement for the Degree of Master of science in Electronic and Digital Resource Management

**Approved by Board of Examiners** **Signature** **Date**

**External Examiner**

______________________________ ________________ __________________

**Internal examiner**

______________________________ ________________ __________________

**Advisors**

______________________________ ________________ __________________

______________________________ ________________ __________________

## Dedicated

This work is dedicated to my dearest brother **KEMAL TAHA** passed away before five years and my dearest sister I lost last year **ZAHA TAHA**!

# Declaration

This thesis is my original work and has not been submitted as a partial requirement for a degree in any university

Mubarak Taha Adem

January, 2015

This thesis has been submitted for examination with my approval as University advisor

DebelaTesfaye (Assistance Professor)
January, 2015

# ACKNOWLEDGEMENT

Bismillahirahmanirahim without your help there is nothing would get into its final. You are always with me with your support Alhamdulilahirabbil'aalamin.

I would like to thank my advisors AtoDebelaTesfaye, and AtoDulaBoru for their constructive comments and encouragement on my work. I am also grateful to all staff of the department of Information science and my colleagues of information science students for their support during my stay at the faculty.

I am grateful to my brother Mohammed Taha and his friend MulusewGerbaba this is your input. Only this work is done through you motivation, support, mission and vision you created inside me so thank you thank you for everything you did for me. You are unforgettable great person in my life for everything thank you there is no words which express my feeling for you.

Finally I would want to thank my family for being on my right hand with their support every time. AlhamduLillah.

# ABSTRACT

Automatic text categorization is a supervised learning task, defined as assigning category labels to new documents based on likelihood suggested by a training set of labeled documents. The world is widely changing hence, the impact of the technology and communications revolution has grown greater today. People have realized the importance of archiving and finding information, only nowadays with the advent of computers and the progress of information technology became possible to store and share large amounts of information, and finding useful information from such collections became a necessity.

Currently Oromia Radio and Television Organization are implementing a manual categorization system to categorize their news items in their day-to-day activities although they are using computer system to store and dispatch information using database systems of un organized information system.

The objective of this research is to apply the novel techniques of machine learning approaches to Afan Oromo news text categorization using Naïve Bayes, Sequential Minimal Optimization and J48 classifier algorithm to recommend the best for the problem at hand. The classifiers use Afan Oromo News items of five classes, collected from Oromia Television and Radio Organization and Voice of America AfaanOromoo program for training and testing of the classifiers. Before the implementation of classifiers, document preprocessing is applied on the prepared document. Under preprocessing steps, removing of digits, punctuation marks, extra characters following this compound words are merged and stop words are removed and finally documents are transformed into term matrix with its weighted values to perform the summarization.

Applying the Naïve Bayes, Sequential Minimal Optimization and J48 classifier on Afan Oromo News Text (the training and testing data sets); finally the model is evaluated using the standard measurement of accuracy, precision, recall and F1 measure. Sequential Minimal Optimization classifier achieved the best 92% accuracy, precision of 92% and recall of 92% and outperforms both j48 and Naïve bayes classifier. J48 classifier registered the second best accuracy of 88.5%, precision 88.5% and 88.5% while Naïve Bayes classifier achieve 87% accuracy, 87% precision and 86.9% recall which is the least from all classifier applied in this study. The result shows that Sequential Minimal Optimization support vector, J48 decision tree and Naïve Bayes classifier is encouraging approach for Afan Oromo News Text.

Keywords: Text categorization, machine Learning, naïve Bayes, Sequential Minimal Optimization support vector, J48 decision tree

# TABLE OF CONTENTS

Table of contents page

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF ABBREVIATION

AO = Afan Oromo

AONT = Afan Oromo News Text

AONTD = Afan Oromo News Text Document

NLP =Natural Language Processing

TC = Text Categorization

NB = Naïve Bayesian

NBA =Naïve Bayesian Algorithm

kNN = k Nearest Neighbor

kNNA = k Nearest Neighbor Algorithm

ML = Machine Learning

MLA = Machine Learning Approach

ADCAONT = Automatic Document Categorization for Afan Oromo News Text

AONTC = Afan Oromo News Text Categorization

WSD= Word sense disambiguation

NNet = Neural Network

TF = Term Frequency

IDF = Inverse Document Frequency

SMO = Sequential Minimal Optimization

# CHAPTER ONE

# 1. INTRODUCTION

## 1.1 BACKGROUND OF THE STUDY

Automatic text categorization is concerned with the assignment of documents to predefined categories.It has been successfully applied in many areas that involve the organization, filing, filtering or routing of documents.  These tasks are part of our everyday lives and can be applied to many contexts such as, assigning patents, advertisements or library books into categories, assigning web-pages to YAHOO!-style directories or filtering spam.

In many real- world scenarios, the ability to automatically classify documents into a fixed set of categories is highly desirable. Common scenarios include classifying a large amount of documenteither supervised or unsupervised archival documents such as newspaper articles, legal records and academic papers. For example, newspaper articles can be classified as ’features ’, ’sports’ or ’news ’. Other scenarios involve classifying of documents as they are created. Examples include classifying movie review articles into ’positive ’ or ’negative ’ reviews or classifying only blog entries using a fixed set of labels .

Today Afan Oromo news text has been producing in increasing amount in every day. Effective machine-generated solutions would obviously increase efficiency and productivity. A computer can process information much faster than humans. With the explosion of electronically stored text, efficiency is of increasing importance. Beyond the immediate efficiency gains, however, is the great promise of machines that appear to “read,” machines that examine free text and make correct decisions. These same techniques that make correct general decisions for text categorization can then be adapted to individual tastes, examining great volumes of text and filtering these documents to suit personal interests (Sheth and Maes, 1993).In this research I claim that such techniques are currently feasible, that they are capable of processing huge amount of Afan Oromo News text documents in reasonable times, with high performance is achievable when high-quality sample datas are available.

Open source data mining tool offers powerful techniques for automatically classifying documents. These techniques are predicated on the hypothesis that documents in different categories distinguish themselves by features of the natural language contained in each do c um e nt. Salient features for document classificationmay include word structure , word frequency, and natural language structure in e ach document.

Text categorization (TC) is one of the important tasks in information retrieval. The problemof TC has been active for four decades, and recently attracted many researchers due to the large amount of documents available on the World Wide Web, in emails and in digital libraries. According to (Alsaleem, 2010) described Automated TC involves assigning text documents in a test data collection toone or more of the pre-defined classes/categories basedon their content. Unlike manual classification, which consumes time and requires high accuracy, automated TC makes the classification process fast and moreefficient since it automatically categorizes documents.

The goal of TC task is to assign class labels to unlabeled text documents from a fixed number of known categories. Each document can be in multiple, exactly one, or no category at all. The large availability of online text documents have provided us a very large amount of information. This available information must be organized systematically for its proper utilization. Systematic organization of information facilitates ease of storage, searching, and retrieval of relevant text content for needy application (Tang, 2009). The Text Classification is an important technique for organizing text documents into classes (Maribor, 2007). Automatic Text classification is attractive research area because it relives the organizations from the need of manually organizing document bases, which is not only expensive, time consuming but also error prone (Sebastiani, 2002).

## 1.2 Statement of the problem.

Automatic text categorization is concerned with the assignment of documents to predefined categories and has been successfully applied in many areas that involve the organization, filing, filtering or routing of documents. These tasks are part of our everyday lives and can be applied to many contexts such as, assigning patents, advertisements or library books into categories, assigning webpages to directories or filtering spam. With the growing of Afan Oromo News text user also wants better browsing and retrieving mechanismsand good quality orclassifying Afan Oromo text news.

A manual method of classification leads to vast consumptionof time, manpower, low qualities (high error made) and productivity. With the creation of large volumes of Afan Oromo news text in electronic form,handling huge amount of data in manual classification system is difficult and has some limitations such as; increase costs by hiring additional human resources,has a very low performance, and will also decreases quality. Data mining applications such as text classification, information retrieval and information extraction are believed to perform this action automatically and manual classification system is no longer required as the machine takes over the classification task and addresses the limitations of manual.

In order to make easy access and timely information, news items should be organized in systematic manner. The greater our ability to storeinformation, the more attention must be paid to the problem of organizing and retrieving it. Traditionally, human experts are engaged in classifying news items manually into their predefined classes. Surafel (1995) reported, automatic text classification systems have proven to be just as accurate, correctly categorizing over 90% of the news stories. They are also far faster and moreconsistent, so there has been a switch from manual to automated systems (Surafel, 1995).

More than 80 languages are spoken in Ethiopia. Afan Oromo is the working language for Oromia region. It is the most used languages in electronic form and for electronic media communication purposes in the region, the country and other places in the world. According to (Chen, 2003), information retrieval has attracted significant attention on the part of researchers in information and computer science over the past few decades. In the 1980s knowledge-based techniques also made an impressive contribution to "intelligent" information retrieval and indexing. More recently, researchers have turned to other newer artificial-intelligence based inductive learning techniques, i.e., machine learning. Text categorization, which is also found to be good in IR, is amenable to machine learning techniques where IR is not (Russel and Norvig, 1995).

Automatic text categorization attempts to replace and save human effort required in performing manual categorization. It consists of assigning and labeling documents using a set of pre-defined categories based on document contents. As Yang stated, automatic text categorization has been used in search engines, digital library systems, and document management systems (Yang, 1999). Such applications have included electronic email filtering, newsgroups classification, and survey data grouping. Rachidi for instance uses automatic categorization to provide similar documents feature (Rachidi et al., 2003).

Currently ORTO is using manual classification for categorizing news articles. As mentioned above the manual classification system is time consuming, has low quality, costly and inconsistent. In contrast the automatic document clustering or categorization via application of data miningtools and techniques is believed to be solution for the mentioned problems and also add many values to information retrieval systems which are vital for the speakers and the user of the language. Therefore, one of the primary objectives of automatic Afan Oromo text categorization is for the enhancement and the support of information retrieval tasks to tackle problems, such as information filtering and routing, clustering of related documents, and the classification of documents into pre-specified subject themes as followed in this research. Accordingly, in this research,NB, SMO and J48 classifier which are machine learning text classification approach,areappliedfor learning and testingthe Afan Oromo News text documents to suggest the best one.

## 1.3 Objective of the study

The general and specific objectives of the study are presented as follows:

### 1.3.1 General objective

The general objective of the study is to adapt and apply NB, SMO support vector and J48 decision tree classifier for automatic text categorization of Afan Oromo News text (AONT) to provide comparative result and recommend the best.

### 1.3.2 Specific objective

The specific objectives of this study are the following:

a) To collect, prepare, and preprocess news items suitable for automatic classification.

b) To adapt NB, SMO support vector and J48 decision tree classifier into Afan Oromo News Text (AONT) for categorization of Afan Oromo.

c) To train and testAONT by NB, SMO and J48 classifier algorithm.

d) To measureand report the performance of theclassifiers.

e) To examine the feasibility of NB, SMO and J48 classifier algorithm on AONT.

f) To compare the classifiersin terms performance and time to build a model and finally draw conclusions.

## 1.4 significance of the study

The primary goal of this study is to demonstrate the feasibility of categorizing Afan Oromo news text using NB, SMO and J48 classifier technique.

Automatic text categorization can also help Afan Oromo language speakers and users in tackling of information overloads by over viewing of the document set that were concisely classified by their likeliness of the text contents and subject dealt within the documents; fast access to relevant document with clearly labeled documents which enhance information retrieval. Using such applications users are allowed to look in only to the subject they want and get the documents they need easily and ignore/skip the irrelevant once. This decreases the time required to acquire the relevant documentsonly.

Text categorization (TC) of newspapers can be organized into their associated predefined categories and help organization of the documents.

In News text, classification appears under certain categories such as Sport, Education, Art, Culture, Politics etc. as a result of this research automatic techniques are employed to relieve the pressure and time-consuming activity of manual classification. Text categorization, either automatic or semi-automatic, can lead to vast improvements in productivity including savings in terms of time and human effort.

## 1.5 Scope and limitation of the study

This study, focus on the automatic document classification for Afan Oromo News text. Therefore thisinvestigation deals with Afan Oromo News text documents only. Put differently, this research doesn't involve in developing classification algorithm in other domains i.e. it is only restricted to news documents only. This comparative study is conducted only on three selected algorithms: Naïve Bayesian (NB), SMO and J48 classifier algorithms.

## 1.6 Organization of the Thesis

This thesis is organized in five chapters. The first chapter contains introduction of text classification and background of the text classification, statement of the problem, objective of this study, the significance of research and scope and limitation of the study is included and discussed.

Chapter two is literature review. This chapter discusses formal definition of text categorization and different literatures reviewed on basic concepts of automatic text classification, the application area of text categorization and text categorization steps are reviewed and discussed.

Chapter three is methodology part of this research. This chapter discusses how the data sets collected and the preprocessing steps developed and applied to make the data sets appropriate and ready for classification task is described in this chapter. Discusses steps and how classifiers algorithm used in this study works and finally the tools used to implement classification is discussed.

Chapter Four discusses about implementation and experiment. This section discusses about the implementation of methods proposed in chapter three. The implementation of the classifiers and the result is discussed. Comparison between the classifiers used in this study is discussed and show in tabular and figure form.

Chapter five is the last chapter of this study, discusses conclusion drawn from the study and finally recommendation is described.

# CHAPTER TWO

## 2. LITERATURE REVIEW

The focus in this study is evaluating and optimizing machine learning techniques for Afan Oromo News text categorization. Please also note that the use of machine learning for text categorization is well understood and is exploited in many languages. This section briefly describes the pertinent of this research and the technology that could be applied to automate Afan Oromo News text classification.

### 2.1 Formal Definition of Text Categorization

Text classification (TC – also known as text categorization, or topic spotting) is the task of automatically sorting a set of documents into categories (or classes, or topics) from a predefined set of labels (Feldman and Sanger, 2007; Hill and Lewicki, 2007 ). Texts are assigned to categories based on a likelihood or confidence score that is suggested by a training set of labeled documents corresponding to each category in the assignment. This confidence ranges between either {0,1} or { -1,1} and in order to arrive at a yes/no decision or a plus/minus figure for the inclusion/exclusion of a document in a category, the confidence score must be mapped onto one of the Boolean values {0,1} or one of {—1,1} using thresholds (Edel, 2004).

According to Sebastini defined Text Categorization is the task of assigning a Boolean value to each pair (dj, ci) € D X C, where D is a domain of documents and C = $(c_1,\ldots,c_{|C|})$ is a set of predefined categories. A value of T assigned to $(d_j, c_i)$ indicates a decision to file $d_j$ under $c_i$,while a value of F indicates a decision not to file $d_j$ under $c_i$ . More formally, the task is to approximate the unknown target function ß: D x C → {T, F} (that describes how documents ought to be classified) by means of a function α: D x C → {T, F} called the classifier (rule,or hypothesis,or model) such that β and α “coincide as much as possible.” How to precisely define and measure this coincidence called effectiveness (Sebstiani, 2002).

As Durgaand Venu said the main aim of text categorization is the classification of documents into a fixed number of pre-determined categories. Every document will be either in multiple, or single, or no category at all. Utilizing machine learning, the main purpose is to learn classifiers through instances which perform the category assignments automatically. This is a monitored learning problem. Avoiding the overlapping of categories is considered as a isolated binary classification problem (Durga and Venu, 2012).

According to Sarasevic categorizing of text is relying only on endogenous knowledge means classifying a document based solely on its semantics, and given that the semantics of a document is a subjective notion, this follows that the membership of a document in a category cannot be decided deterministically (Saracevic, 1975).

As Sebastianiaffirmed, in real world when two human experts decide whether to classify document $d_j$ under category $c_i$ , they may disagree, and this in fact happens with relatively high frequency. For instance news article on Clinton attending Dizzy Gillespie’s funeral could be filed under Politics, or under Jazz,or under both, or even under neither, depending on the subjective judgment of the expert (Sebastiani, 2002).

### 2.1.1 Single-Label versus Multi-label Text Categorization

Depending on the application, there might be different constraints that enforce a given document to be categorized under a given category. A single-label or non-overlapping category happenswhen a given document assigned under only one category. Multi-label or overlapping categories case happen when a given document is assigned under two or more than two category at the same time (Sebastiani, 2002).

### 2.1.2 Category-Pivoted Versus Document-Pivoted Text Categorization

There are two different ways of using a text classifier. Given $d_j \in D$, we might want to find all the $c_i \in C$ under which it should be filed (document-pivoted categorization—DPC); alternatively, given $c_i \in C$, we might want to find all the $d_j \in D$ that should be filed under it (category-pivoted categorization—CPC). This distinction is more pragmatic than conceptual, but is important since the sets Cand Dmight not be available in their entirety right from the start.

DPC is thus suitable when documents become available at different moments in time, e.g., in filtering e-mail. CPC is instead suitable when (i) a new category $c_{|c|+1}$ may be added to an existing set $C=\{c_1,\ldots,c_{|c|}\}$ after a number of documents have already been classified under C, and (ii) these documents need to be re-considered for classification under $c_{|c|+1}$ (Sebstiani, 2002).

### 2.1.3 "Hard" Categorization versus Ranking Categorization

While a complete automation of the TC task requires a True or False decision for each pair ($d_j$, $c_i$), a partial automation of this process might have different requirements. For instance, given $d_j \in D$ a system might simply rank the categories in $C = \{c_1,\ldots,c_{|c|}\}$ according to their estimated appropriateness to $d_j$, without taking any hard decision on any of them. Such a ranked list would be of great help to a human expert in charge of taking the final decision, since she/he could thus restrict the choice to the category (or categories) at the top of the list, rather than having to examine the entire set. Alternatively, given $c_i \in C$ a system might simply rank the documents in D according to their estimated appropriateness to $c_i$; symmetrically, for classification under ci a human expert would just examine the top-ranked documents instead of the entire document set. These two modalities are sometimes called category-ranking TC and document-ranking TC (Yang 1999), respectively, and are the counterparts of DPC (document pivoted categorization) a given document is to be assigned category label(s) and CPC (category pivoted categorization) in which all documents that belong to a given category must be identified.

Regarding to semi automatedLarkey and Croft reported that “interactive" classification systems are useful especially in critical applications in which the effectiveness of a fully automated system may be expected to be significantly lower than that of a human expert(Larkey and Croft, 1996). This may be the case in which the quality of the training data is low, or when the training documents cannot be trusted to be a representative sample of the unseen documents that are to come, so that the results of a completely automatic classier could not be trusted completely.

## 2.2 Basic Concepts of Automatic Text Classification

TC is the classification of documents into a fixed number of pre-defined categories in which each document can be in multiple, exactly one, or no category at all. Generally, TC task goes through three main steps: Text pre-processing, text classification and evaluation. Text pre-processing phase is to make the text documents suitable to train the classifier. Then, the classifier is constructed and adjusted using a learning technique against the training data set. Finally, the classifier gets evaluation by some evaluation measurements i.e. recall, precision, F1-maesure etc. Every language involves its own specific structures which is also the case for Afan Oromo. Afan Oromo (AO) has its own grammatical structures such as subject verbs or object orders and agreements, morphological information, etc. which makes it different from other languages like English. This all contribute to the requirement of specific classification algorithms considering the language unique features.

TC is one of fundamental tasks of text mining in analyzing complex and unstructured data which is concerned about ‘assigning of natural language texts to one or more predefined category based on their content (Dumais, 1998). The concept of text classification has been firstly anticipated in early sixties and it focused on indexing scientific journals using the vocabulary(Feldman, 2007). Latterly, this research field has got more interest due to the fast growth of online documents that holds important and useful knowledge. Therefore, automatic text classification has turned into one of key domains for organizing and handling textual data Currently, there are many applications that are based on the text categorization including: document filtering, spam filtering, automatic metadata generation, classifying web resources under hierarchical catalogues and others (Sebastiani, 2002).

As Lan described text categorization is a long-term research topic which was been actively studied in the communities of Web data mining, information retrieval and statistical learning (Lan, 2005). In the past decade, a number of statistical learning techniques have been applied to text categorization including the k Nearest Neighbor (kNN) approaches, decision trees, Bayesian classifiers, inductive rule learning neural networks and support vector machines (SVM)(Cohen, 1995).

Text categorization (TC) is the task in which texts are categorized into predefined categories based on their contents (Sebastiani, 2002). For example, if texts are represented as a research paper, categories may represent “Computer Science”, “Mathematics”, “Medicine”, etc. The task of TC has various applications such as automatic email classification, web-page categorization and indexing (Feldman, 2007). These applications are becoming increasingly important in today’s information-oriented society especially with the rapid growth of online information, and therefore TC has become one of the key areas for handling and organizing textual data. As mentioned earlier, the goal of TC is the classification of documents into a fixed number of pre-defined categories in which each document can be in multiple, exactly one, or no category at all.

## 2.3 Applications of the Text Categorization

TC goes back to Maron’s (1961) seminal work on probabilistic text classification. Since then, it has been used for a number of different applications (sebastiani, 2002). The assigning of documents to predefined categories is a task that is required in many domains on an everyday basis, such as the labeling of library books or the assignment of patents into associated categories. Until the introduction of automatic solutions, such work has been carried out manually. PubMed1, a service of the National Library of Medicine providing access to over 12 million MEDLINE citations and additional life science journals, spends huge amounts of money each year on human indexers (Edel, 2004).

### 2.3.1 Automatic Indexing for Boolean Information Retrieval Systems

The application that has spawned most of the premature researches in the field (Borko and Bernick 1963; Field 1975; Gray and Harley 1971; Heaps 1973; Maron 1961) is that of automatic document indexing for IR systems relying on a controlled dictionary, the most prominent example of which is Boolean systems. In these latter each document is assigned one or more key words or key phrases describing its content, where these key words and key phrases belong to a finite set called controlled dictionary, often consisting of a thematic hierarchical thesaurus (e.g., the NASA thesaurus for the aerospace discipline, or the MESH thesaurus for medicine).

Usually, this assignment is performed by trained human indexers, and is thus an extremely costly activity. If the entries in the thesaurus are viewed as categories, document indexing becomes an instance of the document categorization task, and may thus be addressed by the automatic techniques described in this thesis. Note that in this case a typical constraint may be that k1 x k2 keywords are assigned to each document, for given k1, k2. Document-pivoted categorization might typically be the best option, so that new documents may be classified as they become available (Addis, 2010; Sebastiani, 2002).

Automatic indexing with controlled dictionaries is closely related to automated metadata generation. In digital libraries, one is usually interested in tagging documents by metadata that describes them under a variety of aspects (e.g., creation date, document type or format, availability, etc.). Some of this metadata is thematic, that is, its role is to describe the semantics of the document by means of bibliographic codes, key words or key phrases. The generation of this metadata may thus be viewed as a problem of document indexing with controlled dictionary, and thus tackled by means of TC techniques (Sebastiani, 2002).

### 2.3.2 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 TC techniques. For instance, at the offices of a newspaper incoming "classified" ads must be, prior to publication, categorized under categories such as Sport, Agriculture, politics, economy, etc.

Most newspapers would handle this application manually those dealing with a high daily number of classified ads might prefer an automatic categorization system to choose the most suitable category for a given ad,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 (Zhang &Oles, 2000; Sebastiani, 2002). According to Yang and Liu reported topic spotting for newswire stories is one of the most commonly investigated applications domains of TC (Yang and Liu, 1999).

### 2.3.3 Document Filtering

Document filtering is the activity of classifying a stream of incoming documents dispatched in an asynchronous way by an information producer to an informationconsumer (Belkin and Croft, 1992). A typical case is a newsfeed, where the producer is a news agency and the consumer is a newspaper (Hayes et al., 1990). 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 (Androutsopoulos et al., 2000; Drucker et al., 1999) and further classify non-junk mail into topical categories of interest to the user.

A document filtering system may be installed at the producer end, in which case its role is to route the information to the interested consumers only, or at the consumer end, in which case its role is to block the delivery of information deemed uninteresting to the user. In the former case the system has to build and update a “profile" for each consumer it serves (Liddy et al., 1994), where as in the latter case a single profile is needed. A profile may be initially specified by the user, thereby resembling a standing IR query, and is usually updated by the system by using feedback information provided by the user on the relevance or non-relevance of the delivered messages.

### 2.3.4 Word sense disambiguation

Word sense disambiguation (WSD) refers to the activity of finding, given the occurrence in a text of an ambiguous (i.e., polynyeous or homonymous) word, the sense this particular word occurrence has. WSD is very important for many applications, including natural language processing, andindexing documents by word senses ratherthan by words for IR purposes. WSD may be seen as a TC task (Gale et al., 1993; Escudero et al., 2000) once we view word occurrence contexts as documents and word senses as categories. Quite obviously, this is a single-label TC case, and one in which document-pivoted TC is usually the right choice.

WSD is just an example of the more general issue of resolving natural language ambiguities, one of the most important problems in computational linguistics. Other examples, which may all be tackled by means of TC techniques along the lines discussed for WSD, are context-sensitive spelling correction, prepositional phrase attachment, part of speech tagging, and word choice selection in machine translation (Addis, 2010; sebastiani, 2002).

## 2.4 Some Text Classifier Techniques

There are many machine learning text classifiers available. This study is conducted by using three popular text classifiers (NB, SMO and J48). There are also other well known classifiers techniques are presented as follow:

### 2.4.1 k-Nearest Neighbors (kNN)

kNNassigns a new test document X to the class that the majority of the k close neighbors to X belongs. According to Mitchell said theclassifier is robust to noiseand quite effective for a large set of training documents (Mitchell, 1997). A major problem involved in the kNN classifier is the "curse of dimensionality". With high dimensional data, the Euclidean distance becomes meaningless and the kNN performs poorly (Surafel, 2003 and Hand, et al 2001). Furthermore, kNN is considered a lazy classifier, since it does not build a model for the training dataand requires more time for classifying objects when a large number of training examples are given. Therefore, nearly all computations take place at the testing time rather than the training time. Therefore, kNN is very in-efficient in terms of both the computational power, and the storage (Yang, et al 1999 and Mitchell, 1997). Another problem of kNN is in choosing k value (Surafel, 2003).

### 2.4.2 Support Vector Machines (SVM)

A Support Vector Machine is a supervised classification algorithm that has been extensively and successfully used for text classification task first applied by Joachims (Joachims, 2002). When learning text classifiers, one has to deal with large number of features. Since SVM use over fitting protection, which does not necessarily depend on the number of features, they have the potential to handle these large feature spaces.However SVM is very time consuming because of more parameters and requires more computation time (Yang, et al 2003).Additionally it suffers from the problem of model parameters where a large number of parameters have to be set in order to provide the optimal solution to a specific problem (Abe, 2005).

### 2.4.3 Neural Network (NNet)

NNet is a network of units, where the input units represent features and the output units represents the category of interest. According to Zurada described the edges connecting the units represent the relations among these units (Zurada, 1992). A basic strength of NNet is its ability to generalize any continuous function. On the other hand, it is very hard to interpret the NNet, or determine why it takes a specific decision. The weakness of NNet is being very slow. Additionally, it’s converge time depends on the network initial conditions (Warner and Misra, 1996). NNets are also affected by the presence of outliers in the training set, since they use the sum-of-square errors, and the problem of the local minima (Abe, 2005).

### 2.4.4 Rocchio Algorithm

As William and Yoram reported the advantages of algorithm are easy to implement, efficient in computation, fast learner and have relevance feedback mechanism but its weakness is low classification accuracy(William and Yoram, 1999).

### 2.4.5 Naïve Bayes Algorithm

Naïve Bayes classifier is a simple probabilistic classifier based on applying Baye's Theorem with strong independence assumptions. This algorithm computes the posterior probability of the document belongs to different classes and it assigns document to the class with the highest posterior probability. This probability model would be independent feature model so that the present of one feature does not affect other features in classification tasks (Irina, 2001). Due to NB classifier's efficiency, simplicity and also has an advantage, that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification, it was implemented in various TC areas including (Surafel, 2003; Duda, eta al 2001; Zhang, 2004; Caruana and Niculescu-Mizil, 2006; Han and Kamber, 2006).

### 2.4.6 J48 Decision tree

J48 builds decision trees from a set of training data, using the concept of information entropy. J48 uses the fact that each attribute of the data can be used to make a decision that splits the data into smaller subsets. Decision tree learning is a way of learning that is used by placing the knowledge in the form of a decision tree. It is used to categorize the types of examples which may come in negative or positive forms. In addition, we can insert more than two types of examples, that is, instead of just positive and negative examples, we can have many other types of examples as well (Wongpun and Srivihok, 2008). Decision tree models are widely used in machine learning and data mining, since they can be easily converted into a set of humanly readable if-then rules (Last et al, 2008).

## 2.6 Text Classification Steps

### 2.6.1 Preprocessing and Document Indexing

Preprocessing is the step of mapping the textual content of a document into a logical view which can be processed by classification algorithms. A general approach in obtaining the logical view is to extract meaningful units (lexical semantics) of a text and rules for the combination of these units (lexical composition) with respect to language. The lexical composition is actually based on linguistic and morphological analysis and is a rather complex approach for preprocessing. Therefore, the problem of lexical composition is usually disregarded in text classification.

#### 2.6.1.1 Removal of Stop Words

Before a document is indexed, the normal procedure in information retrieval and in text classification is to remove stop words. Stop words comprise those words which are neutral to the topic of the document (or query in information retrieval) and would therefore generally contribute very little to the classification of a document. They are often defined by a stop words and include articles, prepositions, conjunctions, pronouns and some high-frequency occurring words. This technique is always performed in IR so as to reduce the number of index terms in a document, to enhance computational efficiency and to minimize the amount of superfluous information in the term space prepositions, conjunctions etc. do not provide information about a document or help in discerning to which category a document belongs.

Many systems use the same generic stoplists consisting of between 300 and 400 words for English. However, research has been conducted into the generation of domain-specific stoplists by (Yang and Wilbur, 1996). Such stoplists are typically much larger than the average domain-independent stoplists so as to make the scaling of categorization systems more tractable when applied to large amounts of data.

Stop words are not removed in experiments using syntactic information to represent the text, for example (Lewis 1992), Chandrasekar and Srinivas (1997).

### 2.6.1.2 Document Representation

Text categorization refers to the automatic labeling of documents based on the nature of the contained text. In order to label documents, systems must first be given access to each document, and the document must be represented by the system in some way (Sable, 2003). Tobuild a classifier model for text categorization using machine learning technique the first step is to generate a representation of each document.All algorithms to text categorization passes through the process of tokenization, feature selection, and creating vector representation of documents. Tokenization is the process of dividing the input into distinct tokens – words and punctuation marks (Zhang &Oles, 2000) is first step and common to most methods of text categorization.

By ignoring lexical composition the logical view of a document $D_j$ can be obtained by extracting all meaningful units (terms) from all documents D and assigning weights to each term in a document reflecting the importance of a term within the document. More formally, each document is assigned an n-dimensional vector $\overrightarrow{Dj}$ = < $w_1$, $w_2$,.,.,$w_n$> whereby each dimension represents a term from a term set **T** . The resulting n-dimensional space is often referred to as Term Space of a document corpus. Each document is a point within this Term Space. So by ignoring lexical composition, preprocessing can be viewed as transforming character sequences into an n-dimensional vector space.

#### 2.6.1.2.1 Term Weighting

a) Term Frequency (TF)

TF calculates the number of times a given term appears in a given document. It measures the importance of each term in a given document and a term with high frequency describes more about the document.

b) Inverse Document Frequency (IDF)

IDF measures, the given terms commonality across all documents. It is calculated by dividing the total number of documents by the number of documents containing the term, then taking the logarithm and quotient.

## 2.7 Machine Learning

There is no conventional algorithm for the task of assigning any as yet unseen documents to a predefined category (Edel, 2004), as no mathematical model of the solution can exist and therefore all we are left to work with in building a classifier are examples. Given a set of examples, we might be able to define input and output values for each given example in the dataset, but we cannot do so for every possible example that exists. It is difficult to generalize from examples to a set of rules or a fixed algorithm for this process. The relationship between the input documents and the desired output category is often too complex to be captured as an algorithm, and so we turn to the technique of machine learning. A machine is said to learn whenever it changes its structure so as to improve expected future performance.

A classic example of an application of machine learning is the speech recognizer. There exists no algorithm to automatically recognize speech from unknown speakers, i.e. no mathematical model can be implemented in order to recognize a person saying, for example, the word learn. For English, for example, we have (or have the potential to obtain) many examples of speech spoken by many different people of different nationalities (English, Irish, American, Australian, Canadian etc). In order to solve the problem of speech recognition we can take a number of examples of different people with different accents saying a particular word and present these examples to the learning machine. The machine can then learn to recognize the word learn by examining a number of examples, some of which may be spoken by British men, some by Irish children, some by American women etc. When the performance of a speech recognition machine improves after hearing many examples of people's speech, we can say the machine has learned (Edel, 2004).

The advantage of a machine learning approach to automatic text categorization is that it is general. Once an implementation of any such method exists, all that is needed to move to a new set of categories is training examples. In fact, creating a corpus of such training examples is often the most time-consuming part of moving to a new set of categories. There are certain text categorization tasks for which the labels are obvious from the start - for example, determining what news group an article comes from - in which case this phase can be skipped; but for most text categorization tasks, automatic creation of a training set does not work well (Sable, 2003).

Since the early '90s, the ML approach to TC has gained popularity and has eventually become the dominant one, at least in the research community. In this approach, a general inductive process (also called the learner) automatically builds a classifier for a category ci by observing the characteristics of a set of documents manually classified under ci by a domain expert; from these characteristics, the inductive process gleans the characteristics that a new unseen document should have in order to be classified under ci. In ML terminology, the classification problem is an activity of supervised learning, since the learning process is "supervised" by the knowledge of the categories and of the training instances that belong to them (Sebastiani, 2002).

Broadly speaking the main two sub fields of machine learning are supervised learning and unsupervised learning (Barber, 2010; Edel, 2004). Supervised learning the training data used to train the learning algorithm consists of many pairs of input/output training patterns - in other words the machine is given the class or output of an input pattern and tries to learn patterns that would arrive at the expected output. The machine learns to adapt based on the experiences of the previous training pattern (Edel, 2004)

### 2.6.1 Training Set and testing Set

The ML approach relies on the availability of an initial corpus $\Omega = \{d_1,\ldots, d_{|\Omega|}\} \subset D$ of

documents pre classified under $C = \{c1,\ldots, c_{|c|}\}$. That is, the values of the total function $\alpha : D \times C \rightarrow \{T, F\}$ are known for every pair (dj , ci) $\in \Omega \times C$. A document $d_j$ is a positive example of $c_i$ if

α (dj , ci ) =T , a negative example of $c_i$ if α (dj , ci ) = F .

In research settings (and in most operational settings too), once a classifier β has been built it is desirable to evaluate its effectiveness. In this case, prior to classifier construction the initial corpus is split in two sets;

A training (-and-validation) set $T\ V = \{d1,\ldots,d_{|TV\ |}\}$. The classifier β for categories $C = \{c1,\ldots.,c_{|C|}\}$ is inductively built by observing the characteristics of these documents;

A test set $Te = \{d_{|TV\ |+1},\ldots, d_{|\Omega\ |}\}$, used for testing the effectiveness of the classifiers. Each $d_j \epsilon$ Te is fed to the classifier, and the classifier decisions $\beta\ (d_j, c_i)$ are compared with the expert decisions $\alpha\ (d_j , c_i)$. A measure of classification effectiveness is based on how often the β (d j ,ci ) values match the α (d j , ci ) values.

### 2.6 Literature Review on Related Works

There are many researchers conducted research on text categorization using one classifier and comparing different classifiers. Some text categorization done in different local languages of Ethiopia and different world languages are reviewed and presented as follows;

Surafel conducted text categorization on Amharic language usingkNN and NB classifier.He conducted four experiments by using three, four, seven and sixteen categories. From his experiment result, he reported that NB classifier performs better than kNN in all experiments. He reported that, both NB and kNN performance was diminishing when the numbers of categories are increased (Surafel, 2003).

Yohannes also conducted Amharic text categorization using two classifiers of Logic Model Tree (LMT) and Support Vector Machine (LibSVM). As he reported both classifiers performance is good. But LibSVM perform better compared to LMT (Yohannes, 2007).

Alemu worked on hierarchical classification of Amharic news items using support vector machine. He also evaluated the performance of the hierarchical classifier over the flat classifier with same data set. The hierarchical classifier performs better than the flat classifiers with same data set. As his experiment result showed that the performance of the classifier increases as it moves down through the hierarchy (Alemu, 2010).

Gebrehiwot conducted research on Tigrigna text categorization from unlabeled documents using repeated bisection and direct k-means for clustering and SVM techniques for classification. From his experiment he reported that, SMO support vector classifiers perform better than j48 and decision tree classifiers (Gebrehiwot, 2011).

Cagri performed experiment by using C4.5, SVM, kNN and NB in order to examine two issues related to Turkish news portals (training set size and robustness) of classifier in terms of time line. Increasing training set size results in accuracy improvement with C4 .5 and SVM classifiers. This increase is not consistent for kNN. But according to his report for NB, small train sets can perform well and NB was also robust in terms of time difference between train and test sets (Cagri, 2011).

Duwairi compared three popular text classification algorithms (kNN, NB, and Distance-Based classifier). Based on her experimental results she reported that, NB outperforms the other two algorithms (Duwairi, 2007).

Abdullah and Mohammed conducted comparative study between three classifier algorithms. The classifiers they used for comparison are SVM, NB and C4.5. As they reported NB outperforms both SVM and C4.5 classifiers using percentage split. On the other hand the Naïve Bayes and SMO classifiers achieved the same accuracy (83.7%) using 10 folds cross validation, while C4.5 achieved the lowest accuracy (79.70%).

Kanaan et al also compared three (kNN, NB, and Rocchio) classification algorithms, they discovered that NB was the best performing algorithm (Kanaan et al, 2009).

# CHAPTER THREE

# METHODOLOGY

## 3. Introduction

In order to build the model for Afaan Oromo News Text Categorization model (AONTC), different text corpuses are prepared for five categories. After corpus is prepared different document preprocessing (tokenization, removal of digits and punctuations, stopwords removal) techniques are applied. Following document preprocessing stage document indexing and representation performed. Proceeding to this step, the three text classifier algorithms selected for this study (NB, SVM and J48) was applied using Weka open source classification tool. Finally Performance of each algorithm was measured.

## 3.1 Document Preprocessing

Document preprocessing is the primary and important part in text classification task. In this step there are different methods applied in order discriminate the informative part of the document and reduce unnecessary words (stop words), punctuation marks and digits.

### 3.1.1 Tokenization, Eliminating Digits, removing of Punctuation marks and Extra Characters

Tokenization is the process of splitting the text into a set of tokens depending on specified rule. It can be tokenized in word based, sentence based or other. In this research word based tokenization was applied by specifying the whitespace delimiter. Punctuation marks, extra characters and digits do not indicate the contents of the document. They have no importance in the document and should have to be removed to reduce the size and get the content bearing words. To perform this action in this work the algorithm 3.1 were applied.

Algorithm 3.1: Tokenize, remove digits, punctuation marks and extra characters

```
Open the file
Read list of punctuation marks
Do
    Read the file and normalize into lowercase
    Assign string to hold the file content
    For token in string split by whitespace
        If token is alpha
            If token is not in punctuation marks list
                Continue
            Else
                Remove the token from the file
            End If
        Else
            Remove the token from the file
        End If
    End For
While End File
```

### 3.1.2 Compound words

Afaan Oromo has many compound words which can be written in different formats. It may written by separating each word with hyphen (-) or as a single word. For example Afan Oromo compound words may written as ("sablammii" or "sab-lammii" (nation), " sardiidaa" or "sardiidaa" (fox))this form. Words written as a single word does not have any problem. Most of the time AO compound words were written by putting hyphen (-) between them. Afan Oromo words written separately by (-), should have to merge together in order to make it a single word as well as to normalize (if there is compound word which has already written as a single word) in the documents. In such case word with the same meaning in a sentence but have different writing format cause it to be treated as distinct word. Treating words having the same meaning but different writing format independently decreases the weight of the word and increases the size of attribute. So in order to address this problem, we prepared Afan Oromo word lists that may appear first and adopted Gebrehiwot's combining Tigrigna compound words algorithm (Gebrehiwot, 2011) as shown in algorithm 3.2.

Algorithm 3.2: Combining Afaan Oromo words

```
Open the file
Do
    Read the file
    Assign the content to string
    For word in string split by space
        If word in compound word list
            Combine first word with next word
        End if
    End for
While end file
```

### 3.1.3 Stop words

In Afaan Oromo there are words which appear commonly in every document. Those words do not describe a given document since they appear many times in many documents. For this work a list of Afaan Oromo stop words were taken from the previous work of (Gezehagn, 2012) and the algorithm 3.3 was applied in order to remove the words from the documents.  Some of AO stop words are (“aanee”, “agarsiisoo”, “akka”, “garas”, “koo”, “kun”).

Algorithm 3.3: Afan Oromo stopword removal

```
Read stopword lists
Open the file
Do
    Read the file
    Convert the file into lowercase
    Tokenize the file and assign to a string
    For token in string
        If token not in stopword lists
            Continue
        Else
            Remove the token from the file
        End If
    End For
While End file
```

## 3.2 Term weighting

To apply machine learning techniques for Afan Oromo news text, documents were indexed and transformed into a representation as suitable for the technique. According to Salton and his friends reported, Vectorspace model is the most widely used method for document representation (Salton et al, 1975). In this model, each document is represented as a vector d. Each dimension in the vector d stands for a distinct term in the term space of the document collection. A term in the document stands for a distinct single-word with its weight.

$d = (w_1, w_2, \ldots, w_{|T|})$_______________________________________3.1

There are various term weighting approaches most of which are based on the following characteristics;

- The relevance of a word to the topic of a document is proportional to the number of times it appears in the document.
- The discriminating power of a word between documents is less, if it appears in most of the documents in the document collection.

Boolean weighting is the simplest method for term weighting. In this approach, the weight of a term is assigned to be 1 if the term appears in the document and it is assigned to be 0 if the term does not appear in the document. This technique does not consider the frequency of the term, rather the weighting is the presence or the absence the term. For this reason it is not widely used.

Term Frequency (TF) weighting is also a simple method for term weighting. In this method, the weight of a term in a document is equal to the number of times the term appears in the document, i.e. to the raw frequency of the term in the document.

Term Frequency×Inverse Document Frequency (TF × IDF) Weighting technique was selected and implemented for this study. This approach is the most common method used for term weighting method. Since Boolean weighting and TF weighting do not consider the frequency of the term throughout all the documents in the document corpus, but TF × IDF takes into account this property. In this approach, the weight of term i in document d is assigned proportionally to the number of times the term appears in the document, and in inverse proportion to the number of documents in the corpus in which the term appears.

$wt_{ij} = tf_{ij} \times \log(N/N_i)$____________________________________3.2

Equation 3.2 indicates that $tf_{ij}$ is the term frequency term i in a document j, loglog(N/$N_i$) is the inverse document frequency of the term, N is the total document number in the corpus, and $N_i$ is the number of documents the term appears.

TF×IDF weighting approach weights the frequency of a term in a document with a factor that discounts its importance if it appears in most of the documents, as in this case the term is assumed to have little discriminating power. This approach was applied in this study to weight terms and to represent the documents.

## 3.3 Categorization Algorithms

To classify Afan Oromo text documents into different categories, three popular text classifiers were selected. The selected classifier algorithms are Naïve Bayesian (NB), Support Vector Machine (SVM) and J48. Using Weka open source from prepared text corpus classification tasks was implemented and classification model was constructed. Finally, performance of each algorithm was analyzed and reported.

### 3.3.1 Naive Bayes (NB)

Naive Bayes is a statistical algorithm that is based on Bayesian theorem. A Bayesian classifier tries to estimate the conditional probability that an input document belongs to a category. It compares "text in a document d" to "text that would be generated by the model associated with a category c." Then it computes an estimate of the likelihood that d belongs to c.

In text categorization, NB calculates probability values in order to assign category labels.Firstly, prior category probabilities are calculated. P ($c_i$) is prior probability that document $d_i$ is in $c_i$ if we knew nothing about "the text in $d_i$." Then we multiply it with the probability that $d_i$ is generated by $c_i$. The result is called the posterior probability P ($c_i$| $d_i$), which can be computed from the product of the prior probability P ($c_i$) and the likelihood P ($d_i$ | $c_i$) according to Bayes theorem:

$$P(c_i \mid d_i) = \frac{P(di|ci)\,P(ci)}{P(di)} \qquad \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 3.3$$

Since the probability that a document $d_i$ occurs in the corpus, P ($d_i$), is a fixed value for a given document $d_i$, we do not need to estimate it. The estimation of the posterior probability P ($c_i$ | $d_i$) is thus converted to the estimation of the prior probability P ($c_i$) and the likelihood P ($d_i$ | $c_i$). If the terms of the input document are assumed to be conditionally independent given the category, the likelihood P ($d_i$ | $c_i$) can be simply calculated by multiplying the likelihood of category $c_i$ with respect to each term:

$$P(d_i \mid c_i) = \prod_{k=1}^{|T|} P(\mathrm{tk} \mid \mathrm{ci}) \qquad \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 3.4$$

Where $t_k$ is the weight of the $k^{th}$ term in document $d_i$, and |T| is the total number of terms. The probability distributions p ($c_i$) and P ($t_k$ | $c_i$) are usually assumed to have known parametric forms, and the learning task is essentially the estimation of the parameters.

### 3.3.2 Support Vector Machines

Support Vector Machines (SVM) is a technique introduced by Vapnik, which is based on the Structural Risk Minimization principle (Burges, 1998; Cortes and Vapnik, 1995). The idea of structural risk minimization is to and a hypothesis h for which we can guarantee the lowest true error. The true error of h is the probability that h will make an error on an unseen and randomly selected test example (Joachims, 1998).

SVM is designed for solving two-class pattern recognition problems. The problem is to find the decision surface that separates the positive and negative training examples of a category with maximum margin. Figure 3.1 illustrates the idea for linearly separable data points. A decision surface in a linearly separable space is a hyper plane. The dashed lines parallel to the solid line show how much the decision surface can be moved without leading to a misclassification of data.

Margin is the distance between these parallel lines. Examples closest to the decision surface are called support vectors.

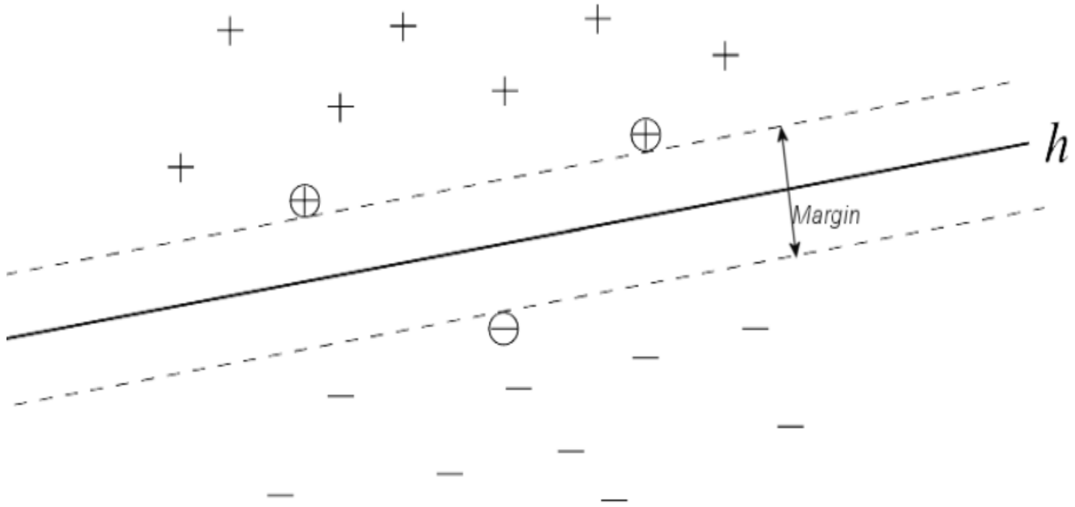


Figure 3.1. Support vector machines find the hyper plane h that separates positive and negative training examples with maximum margin. Support vectors are marked with circles.

The hyperplane for a linearly separable space can be defined by a linear function:

$wx + b = 0$ ________________________________________________3.5

Where x is a document to be classified, w weighting vector and b constant are learned from the training set. The SVM problem is to find w and b that satisfy the following constraints:

Minimize $\|w\|^2$ ________________________________________________3.6

So that $\forall i$: $y_i [wx + b] \geq 1$ ________________________________________3.7

Here, $i \in \{1,2, ..., N\}$,where N is the number of documents in the training set; and yi equals +1 if document xi is a positive example for the category being considered and equals -1 otherwise.

For this study to classify Afan Oromo documents into different categories, we used Support Minimal Optimization (SMO) algorithm. SMO is an SVM algorithm that is particularly suited for linear SVMs and sparse datasets. It exploits the sparseness of the data to improve performance.The optimization problem is broken down in simple, analytically solvable problems which are problems involving only two Lagrangian multipliers. Thus, the SMO algorithm consists of two steps:

1. Using a heuristic to choose the two Lagrangian multipliers.

2. Analytically solving the optimization problem for the chosen multipliers and updating the SVM.

It replaces the quadratic programming inner loop of the SVM algorithm with a heuristic analytic quadratic programming step. It breaks down the quadratic programming problem into a series of smaller quadratic programming problems and at every step chooses to solve the smallest possible optimization problem. Additionally, SMO requires no matrix storage since only two Lagrangian classifiers are solved at a time.

### 3.3.3 J48 Decision Tree

J48 is the final algorithm implemented to classify Afan Oromo documents. A decision tree text classifier is a binary tree T where each inner node is labeled by a term $t_k$ and each branch defines a test on this term deciding if a branch should be taken or not. Leaf nodes are representing classes $ci \in \boldsymbol{C}$ of documents **D.**

Classifying a document $D_j$ means recursively traversing the decision tree by deciding in each inner node which branch should be taken. The decision is based on the representation of a document $D_j$ (i.e. the term vector$\overrightarrow{\boldsymbol{dj}}$ ) and the decision rule for this branch. Classification stops if a leaf node is reached. The class corresponding to the leaf node is assigned.

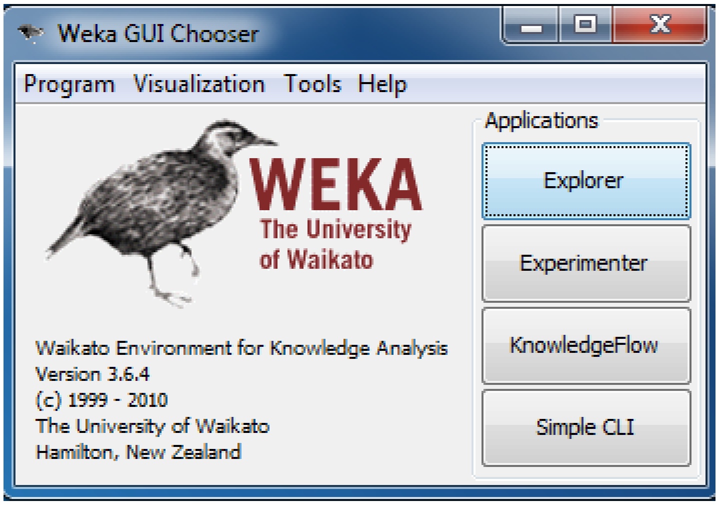
Algorithm for decision tree induction constructs the tree in a top-down recursive divide-and-conquer  manner. As discussed in some works, the decision tree algorithm steps were summarized as indicated below (Quinlan, 1993; Witten and Frank, 2005).

Algorithm 3.4: J48 decision tree classifier algorithm

- ➢ At start, all the training examples are at the root
- ➢ Examples are partitioned recursively based on selected attributes
- ➢ Test attributes are selected on the basis of a heuristic or statistical measure
- ➢ The algorithm stop partitioning in one of the following conditions:
    - ✓ All samples for a given node belong to the same class
    - ✓ There are no remaining attributes for further partitioning – majority voting is employed for classifying the leaf
    - ✓ There are no samples left

## 3.4. Text Classifier Tool

For this study WEKA (Waikato Environment for Knowledge Analysis) was used for text classification. It is a popular suite of machine learning software written in Java, developed at the University of Waikato (Hallet al, 2009; Witten and Frank, 2005). It is free software available under the GNU General Public License. WEKA provides a large collection of machine learning algorithms for data preprocessing, classification, clustering, association rules, and visualization.

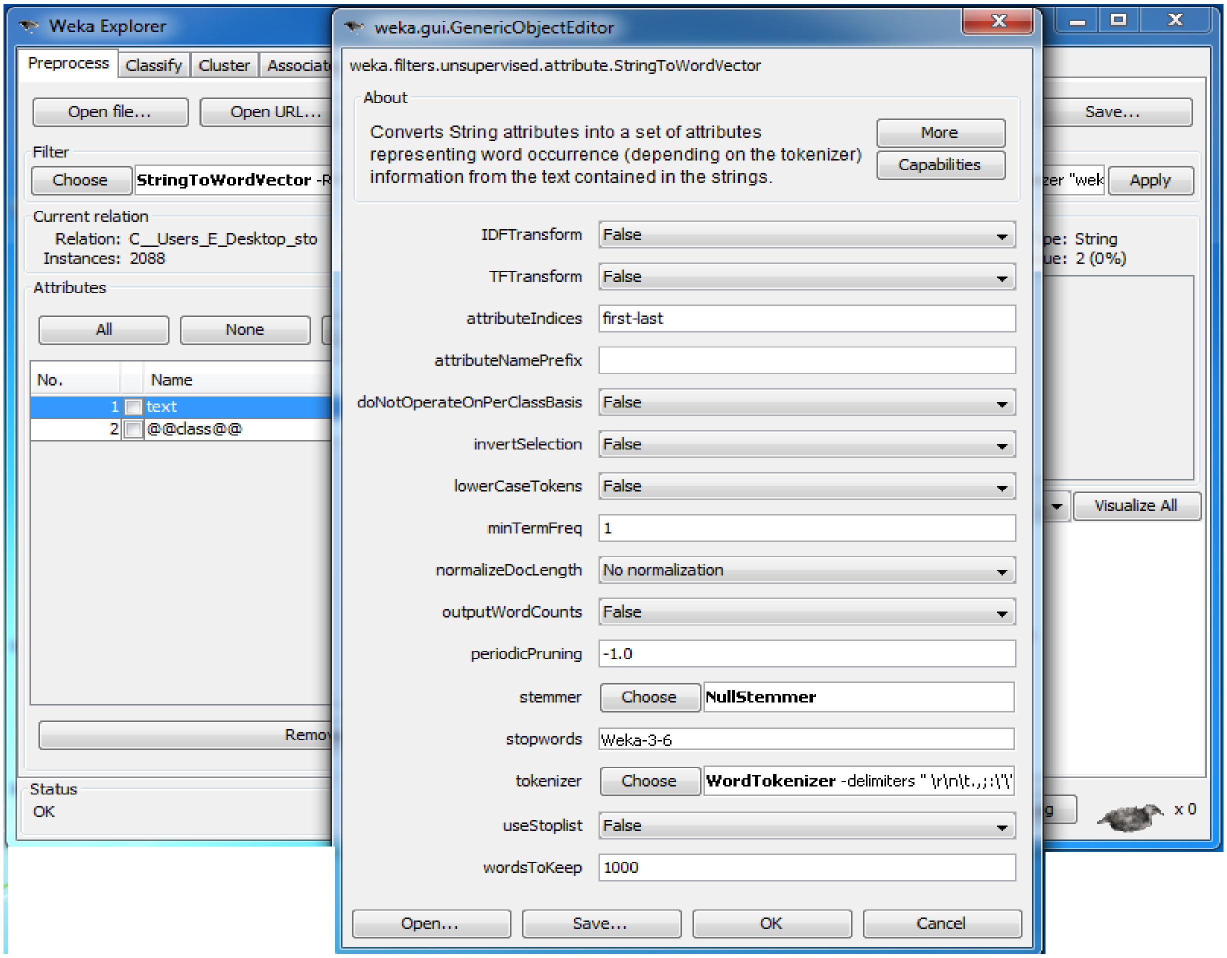


Waikato Environment for Knowledge Analysis
Version 3.6.4
(c) 1999 - 2010
The University of Waikato
Hamilton, New Zealand

Figure 3.1: WEKA GUI chooser.

It can be seen from figure 3.1 that it is a big workbench for data analysis and machine learning where a majorityof the most common algorithms used in data mining have been implemented and are readyto be used. The WEKA suite is divided into 3 different graphical interfaces as seen in figure 3.1, namely the Explorer, Experimenter and the Knowledge Flow. There is also a command line terminal mode where you can call the different methods with function calls directly. In the Explorer, the user gain quickly access to all the features in WEKA and can freely analyze the data.The Experimenter on the other hand is more focused on setting up machine learning experiments where you can more easily compare different algorithms against each other. Finally, in the Knowledge flow the user can use the same methods as in the Explorer, but instead of applying a certain function one at a time the user can set up complex flows that do the whole chain from reading the data to plotting the result in a graph.

For this study, the researcher used Weka explorer interface. As the explorer opened, it redirects to preprocess menu and under this menu there are option tools such as open file, open url, open DB etc exist to open datas and choose option tool to choose filter algorithm. We used the open file option to open the prepared data for the study. Next to this, using WEKA StringToWordVector tool options under filters option (filters → unsupervised →attribute →StringToWordVector) with different combinations, we setup the term weighting combinations to the data in order to convert words into numeric values with its weight. Finally convert the data into ARFF (Attribute Relationship File Format) which was a single flat form of the entire file.



StringToWordVector -R

Current relation
Relation: C__Users_E_Desktop_sto
Instances: 2088

| No. | Name |
|---|---|
| 1 | □ text |
| 2 | □ @@class@@ |

Status
OK

er "wek
pe: String
ue: 2 (0%)

weka.filters.unsupervised.attribute.StringToWordVector

Converts String attributes into a set of attributes representing word occurrence (depending on the tokenizer) information from the text contained in the strings.

| | |
|---|---|
| IDFTransform | False |
| TFTransform | False |
| attributeIndices | first-last |
| attributeNamePrefix | |
| doNotOperateOnPerClassBasis | False |
| invertSelection | False |
| lowerCaseTokens | False |
| minTermFreq | 1 |
| normalizeDocLength | No normalization |
| outputWordCounts | False |
| periodicPruning | -1.0 |
| stemmer | Choose NullStemmer |
| stopwords | Weka-3-6 |
| tokenizer | Choose WordTokenizer -delimiters " \r\n\t.,;:\'\ |
| useStoplist | False |
| wordsToKeep | 1000 |

Figure 3.2: WekaStringToWordVector options

After preprocessing, converting the data into ARFF file format and the researcher implement the selected classifier (NB, SVM and J48) algorithm in order to classify preprocessed Afan Oromo document into different categories. Finally, the time to build the model by each classifier and using the standard accuracy measurements the detail accuracy of each algorithm is measured.

## 3.5 Performance Measures

There are various methods to measure the performance of classifiers. To measure the performance of the classifier for Afan Oromo documents, we used the standard and most commonly used effectiveness measurements such as Accuracy, Precision, Recall and F-measure. Accuracy refers the percentage of correct predictions made by the model when compared to the actual classifications.

Table 3.1: Contingency table for class $C_i$

| Class ci | | Assigned by expert? | |
|---|---|---|---|
| | | Yes | No |
| Assigned by classifier? | Yes | TPi | FPi |
| | No | FNi | TNi |

Where

TPi True positive: Those assessments where system and human expert agree for a label assignment.

FPi False positive: Those labels assigned by the system that does not agree with expert assignment.

FNi False negative: Those labels the system failed to assign as they were by human expert.

TNi True negative: Those non assigned labels that also were discarded by the expert.

Based on the contingency table3.1 the values Precision, Recall and F-measure can be computed as shown on equation below.

Precision (P) = $\frac{T\ P}{T\ P + FP}$ ________________________________________________ 3.3

Precision indicates for the percentage that if a document is assigned to the class, this assignment will be correct.

Recall (R) = $\frac{T\ P}{TP + FN}$ __________________________________________________3.4

Recall is an indicator for the percentage of how many documents of a class are classified correctly.

F-measure is the harmonic average of precision and recall. Defined as:

F = $\frac{2PR}{P + R}$ __________________________________________________________3.5

# CHAPTER FOUR

# EXPERIMENT AND EVALUATION

This research was conducted, relayed on the data sets obtained from two sources. The data sets then preprocessed using python programming language and make ready for classification task. Classification of Afan Oromo was applied using Weka open source application package by thee selected classification algorithms NB, SMO and J48 classifier. Finally, the classifiers performance was analyzed by using standard performance measurements of Accuracy, precision, recall and F-measure.

## 4.1 Proposed System Architecture

The proposed Afan Oromo document classification has five steps as shown in figure 4.1. It includes; data collection, data preprocessing, training, testing and finally results and evaluation. In the first step the data sets was collected for five categories. In the second stage the collected data sets were preprocessed; digits, punctuation mark and extra characters were removed from the documents, then Afan Oromo compound words were merged together and also stopwords were excluded from the documents. Finally, the documents were transformed in TF X IDF matrix weighted values. In the third step the classifiers were trained using the training data sets (which were 66% of the total document) and in fourth step the testing data sets (the remaining 34%) were used to test the classifier performance. Finally in the fifth step the result obtained from testing documents were evaluated using standard performance measurements.

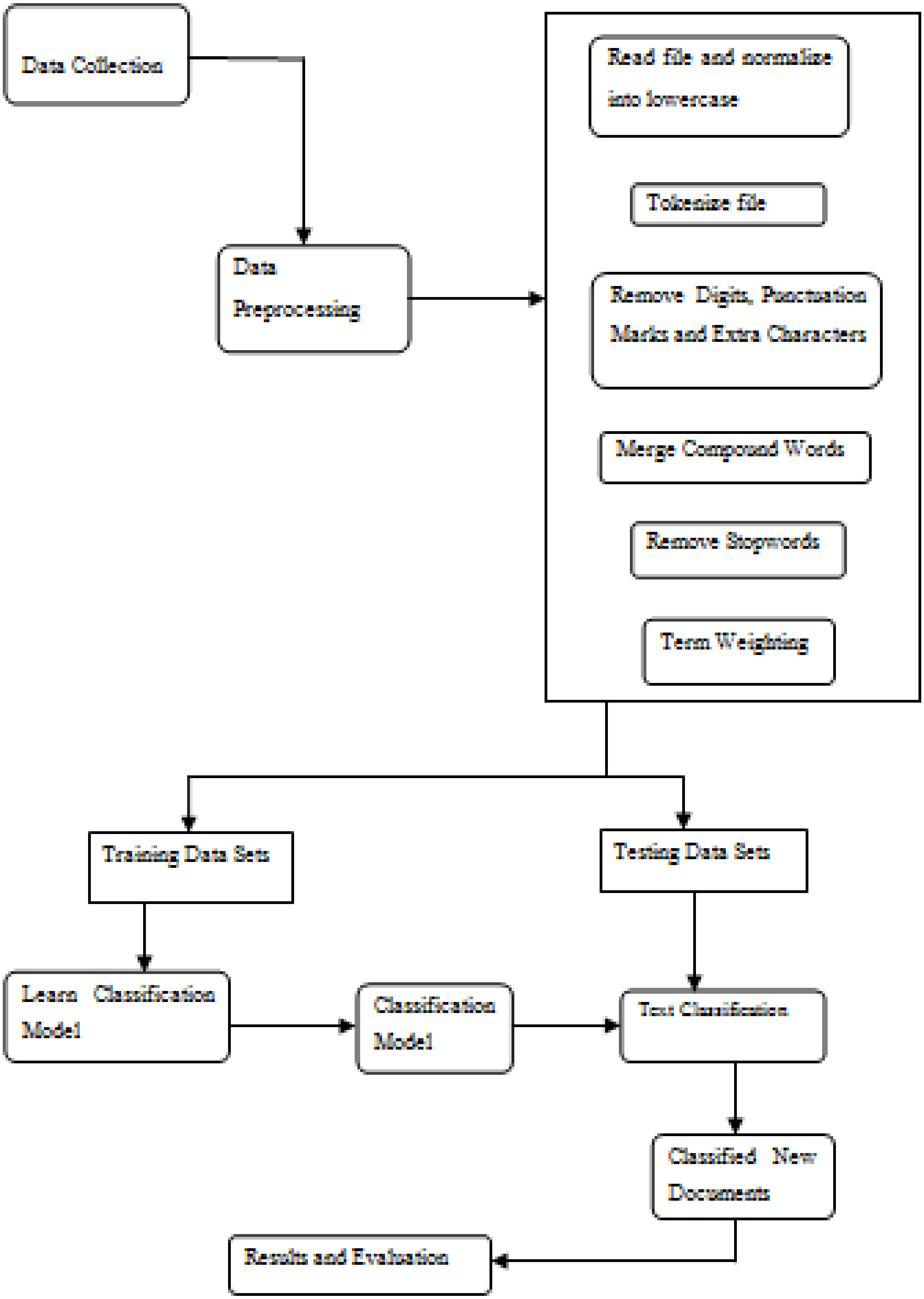


Figure 4.1: Afan Oromo Text Categorization System Architecture

## 4.2dataset Collection and preparation

The datasets for this study were collected from two sources. Oromia Television and Radio Organization (ORTO)[1] and VOA Afaan Oromoo[2] program, were used as a source or this study. To collect the document for this study we used Inspire webdisk2 software to download all contents of the websites in order to access it offline. From ORTO website, the researcher prepare 1723 (one thousand seven Hundred twenty three) text document from five different programs called “Fayyaa” (Health), “Siyaasa” (Politics), “Ispoortii” (Sport), “Bizinasii fi Ikoonoomiksii” (Business and Economics) and “Qonnaa” (Agriculture) and from VOA AfaanOromoo program 418 (four hundred eighteen) text documents were prepared from three different programs namely “Ispoortii” (Sport), “Fayyaa” (Health) and “DirreeDimokraasii” (Democracy Field). Finally, merged the documents prepared from ORTO and VOA AfaanOromoo which was under the same program (“Ispoortii” (Sport), “Fayyaa” (Health) and “Siyaasa” (Politics)). Totally 2088 (two thousand eighty eight) text documents were prepared for this study.

[1] http://www.orto.gov.et/

[2] http://www.voaafaanoromoo.com/

Table 4.1: Detail of ORTO and VOA AfaanOromoo Datasets

| No | Category Name | ORTO | VOA AfaanOromoo | Total |
|---|---|---|---|---|
| 1. | “Bizinasii fi Ikoonoomiksii” (Business and Economics) | 432 | | 432 |
| 2. | “Fayyaa” (Health) | 258 | 113 | 371 |
| 3. | “Ispoortii” (Sport) | 254 | 131 | 385 |
| 4. | “Qonnaa” (Agriculture) | 482 | | 482 |
| 5. | “Siyaasa” (Politics) | 297 | 121 | 418 |
| Total | | 1723 | 418 | 2088 |

## 4.3Afan Oromo Document Preprocessing

Before implementing the text classification task in this study the researcher preprocessed the collected documents in order to make it ready for classification. The preprocessing performed in this study includes different tasks such as tokenizing, normalizing, eliminating digits, removing punctuation marks and extraneous characters, stop word removal and compound word normalization.

Tokenization is one of the text preprocessing tasks applied in Afan Oromo text documents. It is the process of splitting the document into words, sentences, paragraphs or lines. For this research we split Afan Oromo text documents into words.

After tokenization, since the document contains the combination of digits, punctuation marks and extra characters which were unnecessary in the documents should have to be reduced from the documents. In order to reduce the documents contents to its explanatory part, those digits punctuation marks and extra characters were eliminated from Afan Oromo documents.

In Afan Oromo writing system compound words can be written in two forms. It can be written either by combining two words for example (“**Sar**diidaa” (fox), “r**og**sadee” (triangle)) or separating them by hyphen (-) for instance (“Sar-diidaa” (fox), “rog-sadee” (triangle)). In such writing form, the words considered as they are different words but not. So we normalized and make Afan Oromo compound as single word by merging the words together.For Afan Oromo compound wordswe prepared lists ofwordsthat may appear first, and then read the tokens in the document if it exist in prepared lists ofwords then it merge with the next token.

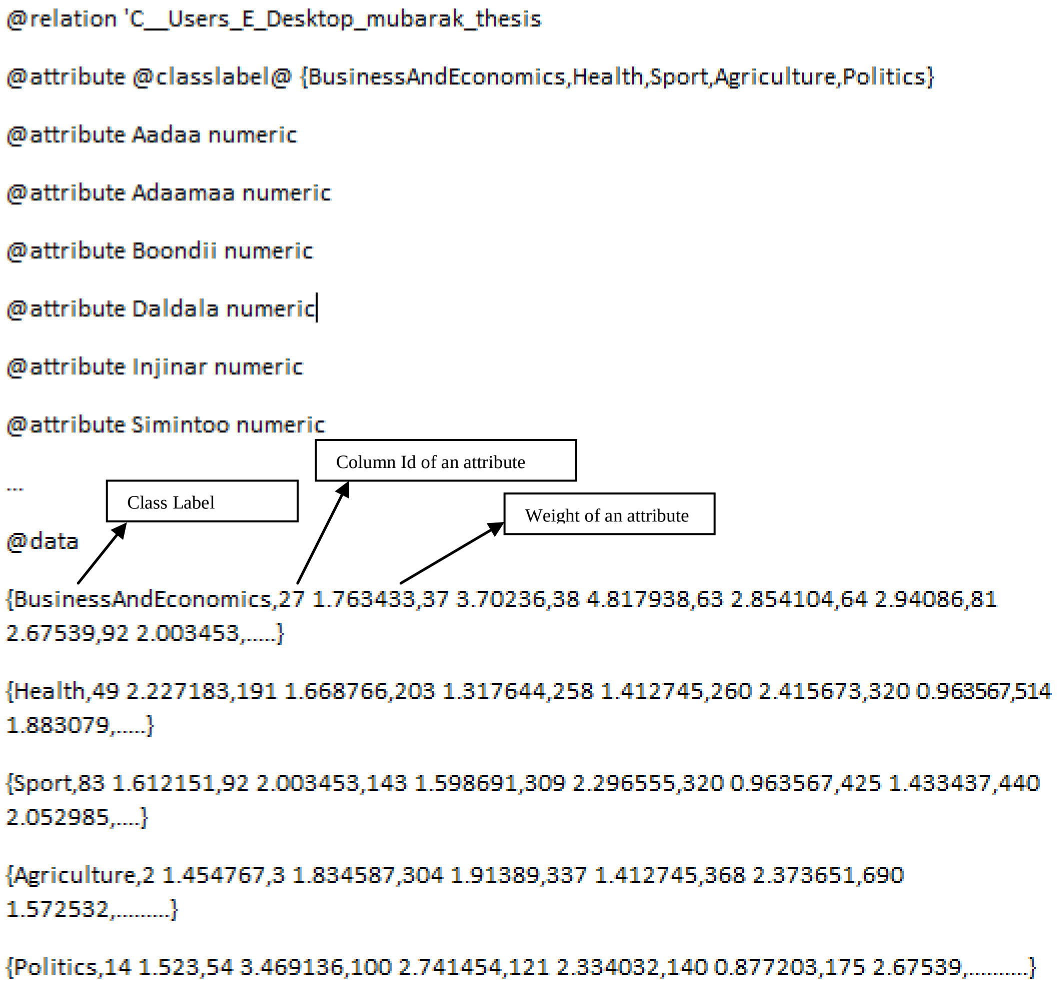
Stop-words are most frequent terms which are common to every document. They have no discriminating power to distinguish one document from the other. Those words are usually article, preposition and etc which are insignificant for that document and do not bring any effect on the document by their removal. For this research we used 222 total number of Afan Oromo stop word lists and exclude from the documents.

## 4.4Term Weighting and Document Representation

After preprocessing the documents, term weighting task was the technique we used to representdocument. For this work we kept all terms left after preprocess to represent the document. In order to represent the documents and to keep the weight of each terms across all documents, the researcher used TF x IDF technique. We used TF x IDF techniques in order to eliminate bias and to normalize weight of the terms. For this study 3861 (three thousand eight hundred sixty one) feature sets were used to represent the documents.

## 4.4 Data Conversion

For this research Weka data mining tools was selected to apply automatic Afan Oromo text documents classification and experimentation purpose. In order to apply classification algorithm we converted our data sets that was appropriate to Weka tools which was ARFF (Attribute Relationship File Format). ARFF files format has two distinct sections as shown in figure 4.2. The first section contains the header information and the second section contains the data information. The header of the ARFF file contains the name of the relation (data set name), list of attributes (the columns in the data) and type's of the attribute. The last section is the data section which shows the class of the document, value of the attribute and its weight in the given documents.



```
@relation 'C__Users_E_Desktop_mubarak_thesis

@attribute @classlabel@ {BusinessAndEconomics,Health,Sport,Agriculture,Politics}

@attribute Aadaa numeric

@attribute Adaamaa numeric

@attribute Boondii numeric

@attribute Daldala numeric

@attribute Injinar numeric

@attribute Simintoo numeric

...

@data

{BusinessAndEconomics,27 1.763433,37 3.70236,38 4.817938,63 2.854104,64 2.94086,81
2.67539,92 2.003453,.....}

{Health,49 2.227183,191 1.668766,203 1.317644,258 1.412745,260 2.415673,320 0.963567,514
1.883079,.....}

{Sport,83 1.612151,92 2.003453,143 1.598691,309 2.296555,320 0.963567,425 1.433437,440
2.052985,....}

{Agriculture,2 1.454767,3 1.834587,304 1.91389,337 1.412745,368 2.373651,690
1.572532,.........}

{Politics,14 1.523,54 3.469136,100 2.741454,121 2.334032,140 0.877203,175 2.67539,..........}
```

Figure 4.2: ARFF input file format for Weka tool

## 4.5Experiment

In this study NB, SVM and J48 decision tree classifiers were used to build classification model and to classify Afan Oromo text documents into different categories. In order to carry out the experiment we split the total data sets as training data sets 66% and the remaining 34% as testing data set. The training and testing data sets for this experiment was randomly selected with specified percentage split of the total data sets. Generally 2088 (two thousand eighty eight) data sets were prepared for this study.

Table 4.2: Training and testing data sets

| No | Category Name | Training sets | Testing sets | Total |
|---|---|---|---|---|
| 1. | Agriculture | 318 | 164 | 482 |
| 2. | Business and Economics | 285 | 147 | 432 |
| 3. | Health | 245 | 126 | 371 |
| 4. | Politics | 276 | 142 | 418 |
| 5. | Sport | 254 | 131 | 385 |
| | Total | 1378 | 710 | 2088 |

### 4.5.1. Classification Using NB classifier

Weka support Bayes Net, Complement Naive Bayes, DMNBtext, Naïve Bayes, Naive Bayes Multinomial, Naïve Bayes Simple and Naïve Bayes Updateable. But for this study we carried out the experiment using Naïve Bayes classifier. We used 1378 (one thousand three hundred seventy eight) instances which is 66% of the total data sets, as training data sets and 710 (Seven hundred ten) instances which is 34% of the total data sets as testing data sets. Out of the testing data sets NB correctly classified 617 instances and while 93 instances were classified incorrectly. Time taken to build model was 13.15 seconds.

```
                 === Summary ===

Correctly Classified Instances       617          86.9014 %
Incorrectly Classified Instances      93          13.0986 %
```

Weka provides different types of options for measuring the performance a classifier. We presented here the confusion matrix and detailed accuracy as follows;

Table 4.3: Confusion matrix of NB classifier

| Agriculture | BusenessandEconomics | Health | Politics | Sport | |
|---|---|---|---|---|---|
| 143 | 12 | 0 | 6 | 3 | Agriculture |
| 7 | 117 | 1 | 12 | 10 | BusenessandEconomics |
| 5 | 0 | 109 | 9 | 3 | Health |
| 6 | 2 | 11 | 123 | 0 | Politics |
| 0 | 0 | 6 | 0 | 125 | Sport |

From the above table we can see that 143 instances of agriculture we classified as correctly (as True Positive TP) of class Agriculture while 21 instances were incorrectly classified and which were assumed by classifier as False Positive (FP). Totally 164 actual instances of Agriculture class were presented for classifier.

Based on the above confusion matrix table the detailed performance of NB classifier is shown below in table 4.4.

Table 4.4: Detailed accuracy of NB classifier by class

| TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---|---|---|---|---|---|---|
| 0.872 | 0.033 | 0.888 | 0.872 | 0.88 | 0.961 | Agriculture |
| 0.796 | 0.025 | 0.893 | 0.796 | 0.842 | 0.926 | BusenessandEconomics |
| 0.865 | 0.031 | 0.858 | 0.865 | 0.862 | 0.928 | Health |
| 0.866 | 0.048 | 0.82 | 0.866 | 0.842 | 0.931 | Politics |
| 0.954 | 0.028 | 0.887 | 0.954 | 0.919 | 0.967 | Sport |
| 0.869 | 0.033 | 0.87 | 0.869 | 0.869 | 0.943 | |

Based on confusion matrix table 4.3 the detail of accuracy was calculated and presented on the above table 4.4. As we see from the table 4.3 the best accuracy was registered by class “Sport” with 95.4% followed by class “Agriculture” of 87.2%. The least accuracy was recorded by class “BusenessandEconomics” with 79.6% of accuracy. The best precision was recorded by class “BusenessandEconomics” with 89.3% followed by class “Agriculture” and “Sport” with 88.8% and 88.7% respectively, while class “Politics” registered the least precision 82%.  Looking to recall level of each class, the best recall was witnessed on class “Sport” with 95.4% followed by “Agriculture” and “Politics” with 87.2% and 86.6% respectively whereas 79.6% was the least recall recorded by “BusenessandEconomics” class. Class “Sport” registered best F-measure of 92% whereas 84.2% is the least registered by class “Politics” and “BusenessandEconomics”. Generally, NB classifier recorder the weighted average of accuracy 87%, precision 87 %, recall 87% and F-measure of 87%.

### 4.5.2. Classification Using SVM classifier

Weka 3.6.4 version we used for the experiment in this study has different types SVM classifier. We executed our experiment using SMO (polykernel) classifier of Weka. From Total 710 testing instances SMO classified correctly 653 instances which is 91.9718 % and classified incorrectly 57 instances which is 8.0282 %. The total time taken to build model is 5.87 seconds.

```
                    === Summary ===

Correctly Classified Instances        653          91.9718 %
Incorrectly Classified Instances       57           8.0282 %
```

The confusion matrix of SMO classifier is show on table 4.5 as follows:

Table 4.5: Confusion matrix of SMO classifier

| Agriculture | BusenessandEconomics | Health | Politics | Sport | |
|---|---|---|---|---|---|
| 153 | 6 | 0 | 5 | 0 | Agriculture |
| 0 | 137 | 0 | 0 | 10 | BusenessandEconomics |
| 0 | 0 | 108 | 9 | 9 | Health |
| 3 | 0 | 6 | 133 | 0 | Politics |
| 0 | 3 | 6 | 0 | 122 | Sport |

Based on the above confusion matrix, the performance of SMO classifier is shown in table 4.6.

Table 4.6: Detailed accuracy of SMO classifier by class

| TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---|---|---|---|---|---|---|
| 0.933 | 0.005 | 0.981 | 0.933 | 0.956 | 0.97 | Agriculture |
| 0.932 | 0.016 | 0.938 | 0.932 | 0.935 | 0.975 | BusenessandEconomics |
| 0.857 | 0.021 | 0.9 | 0.857 | 0.878 | 0.936 | Health |
| 0.937 | 0.025 | 0.905 | 0.937 | 0.92 | 0.964 | Politics |
| 0.931 | 0.033 | 0.865 | 0.931 | 0.897 | 0.97 | Sport |
| 0.92 | 0.019 | 0.921 | 0.92 | 0.92 | 0.964 | |

The detail accuracy for each class is computed from confusion matrix table 4.5 and shown in table 4.6 above. As we see from the detail accuracy table 4.6 the best accuracy was recorded by class “Politics” with 93.7% while the least accuracy was 85.7% which was registered by class “Health”. When we look at precision level of classes the best was 98.1% of precision registered by class “Agriculture” whereas 86.5% is the least precision recorded by “Sport” class. The best recall was registered by class “Politics” 93.7% while the least was 85.7% registered by “Health” class. Coming across to F-measure, 95.6% is the top F-measure from the lists was recorded by class “Agriculture” but 87.8% is the least recorder by “Health” class. Generally, SMO classifier registered weighted average of accuracy 92%, precision 92.1%, recall 92% and F-measure of 92%.

### 4.5.3. Classification Using Decision Tree classifier

Weka 3.6.4 version we used for the experiment in this study support different types decision tree classifier. This research experiment was carried out by using J48 decision tree classifier. J48 correctly classified 628 out of 710 instances, while the remaining 82 instances were incorrectly classified. The total time taken to build model is 327.85 seconds.

```
                    === Summary ===
Correctly Classified Instances        628        88.4507 %
Incorrectly Classified Instances       82        11.5493 %
```

Table 4.7: Confusion matrix of J48 classifier

| Agriculture | BusenessandEconomics | Health | Politics | Sport | |
|---|---|---|---|---|---|
| 156 | 6 | 0 | 2 | 0 | Agriculture |
| 13 | 118 | 2 | 1 | 13 | BusenessandEconomics |
| 3 | 3 | 105 | 9 | 6 | Health |
| 8 | 0 | 4 | 130 | 0 | Politics |
| 0 | 6 | 6 | 0 | 119 | Sport |

Based on the above confusion matrix, the performance of J48 classifier is shown in table 4.8below.

Table 4.8: Detailed accuracy of J48 classifier by class

| TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---|---|---|---|---|---|---|
| 0.951 | 0.044 | 0.867 | 0.951 | 0.907 | 0.962 | Agriculture |
| 0.803 | 0.027 | 0.887 | 0.803 | 0.843 | 0.95 | BusenessandEconomics |
| 0.833 | 0.021 | 0.897 | 0.833 | 0.864 | 0.951 | Health |
| 0.915 | 0.021 | 0.915 | 0.915 | 0.915 | 0.968 | Politics |
| 0.908 | 0.033 | 0.862 | 0.908 | 0.885 | 0.985 | Sport |
| 0.885 | 0.03 | 0.885 | 0.885 | 0.884 | 0.963 | |

The detail accuracy for each class is computed from confusion matrix table 4.7 and shown in table 4.8 above. As we understand from the detailed accuracy table 4.8 the best accuracy was achieved by class “Agriculture” with 95.1% while the least accuracy was 80.3% which was registered by class “BusenessandEconomics”. When we look at precision levels of each class the best was 91.5% of precision registered by class “Politics” whereas 86.2% is the least precision recorded by “Sport” class. The best recall was registered by class “Agriculture” 95.1% while the least was 80.3% registered by “BusenessandEconomics” class. Regarding to F-measure, 91.5% is the top from the lists recorded by class “Politics” whereas 84.3% is the least recorder by “BusenessandEconomics” class. Generally, J48 classifier registered weighted average of accuracy 88.5%, precision 88.5%, recall 88.5% and F-measure of 88.4%.

## 4.6. Comparison and Discussion

Among three classifiers applied on Afan Oromo documents, SMO achieved the highest average accuracy of 92%, followed with average accuracy of 88.5 % of J48 classifier. NB registered the least average accuracy of 87%. Table 4.9 shows the comparison of NB, SMO and J48 classifier in terms of correctly classified instances, incorrectly classified instances and total time taken by each classifier to build model.

Table 4.9: Comparison of NB, SMO and J48 classifiers

| | NB | | SMO | | J48 | |
|---|---|---|---|---|---|---|
| Correctly Classified Instances | 617 | 87% | 653 | 92% | 628 | 88.5% |
| Incorrectly Classified Instances | 93 | 13% | 57 | 8% | 82 | 11.5% |
| Time taken to build a model | 13.51 seconds | | 5.87 seconds | | 327.85 seconds | |

As we understand from the above table J48 suffered a long time to build model. SMO required the least time to build the model.

Figure 4.3 and 4.4 below shows the three classifiers average accuracy and total time taken to build model.

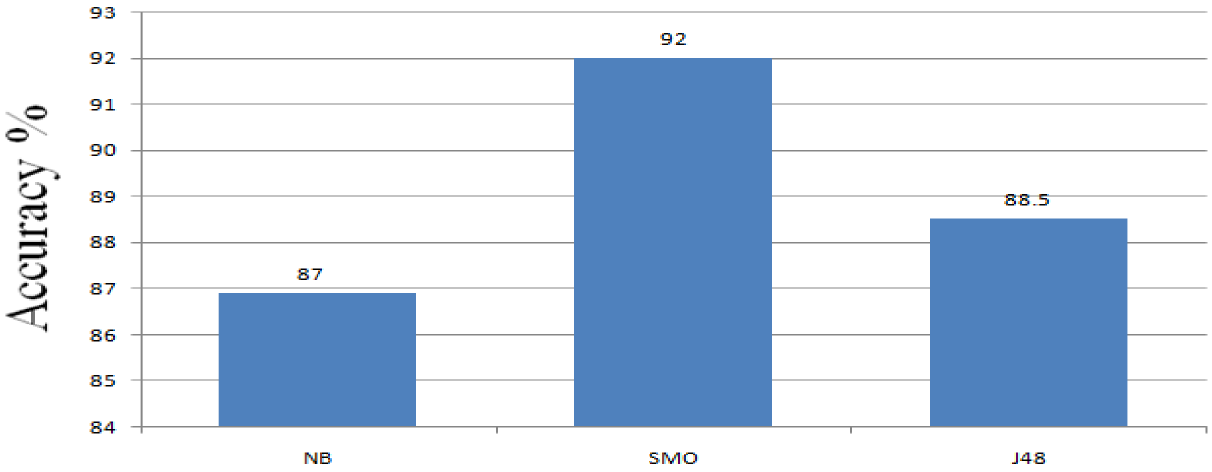


Figure 4.3: classifiers average accuracy

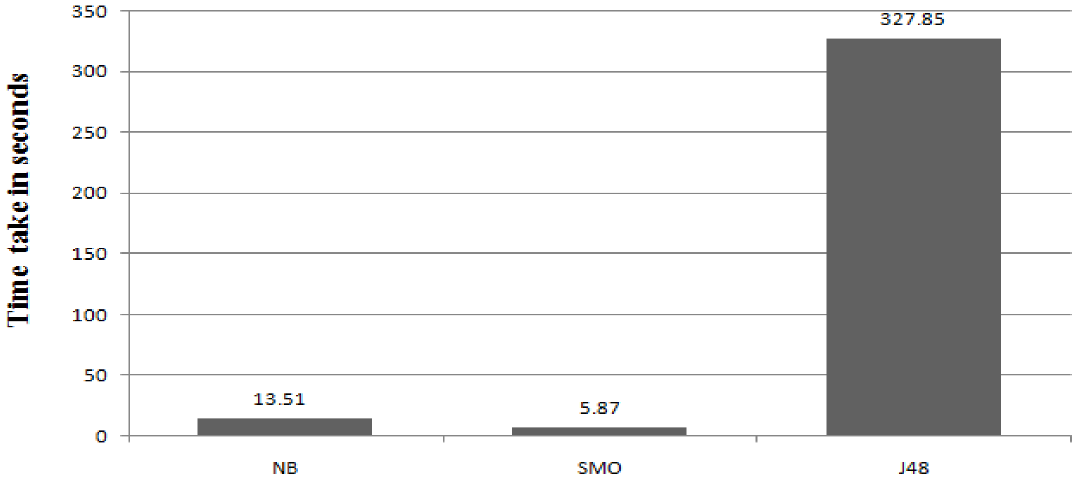


Figure 4.4: Total time taken to build model

Table 4.10 shows the details of precision, recall and F-measure of NB, SMO and J48 classifier

| Class | NB | | | SMO | | | J48 | | |
|---|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | F-Measure | Precision | Recall | F-Measure | Precision | Recall | F-Measure |
| Agriculture | 0.888 | 0.872 | 0.88 | 0.981 | 0.933 | 0.956 | 0.867 | 0.951 | 0.907 |
| BusenessandEconomics | 0.893 | 0.796 | 0.842 | 0.938 | 0.932 | 0.935 | 0.887 | 0.803 | 0.843 |
| Health | 0.858 | 0.865 | 0.862 | 0.9 | 0.857 | 0.878 | 0.897 | 0.833 | 0.864 |
| Politics | 0.82 | 0.866 | 0.842 | 0.905 | 0.937 | 0.92 | 0.915 | 0.915 | 0.915 |
| Sport | 0.887 | 0.954 | 0.919 | 0.865 | 0.931 | 0.897 | 0.862 | 0.908 | 0.885 |
| Weighted Avg. | 0.87 | 0.869 | 0.869 | 0.921 | 0.92 | 0.92 | 0.885 | 0.885 | 0.884 |

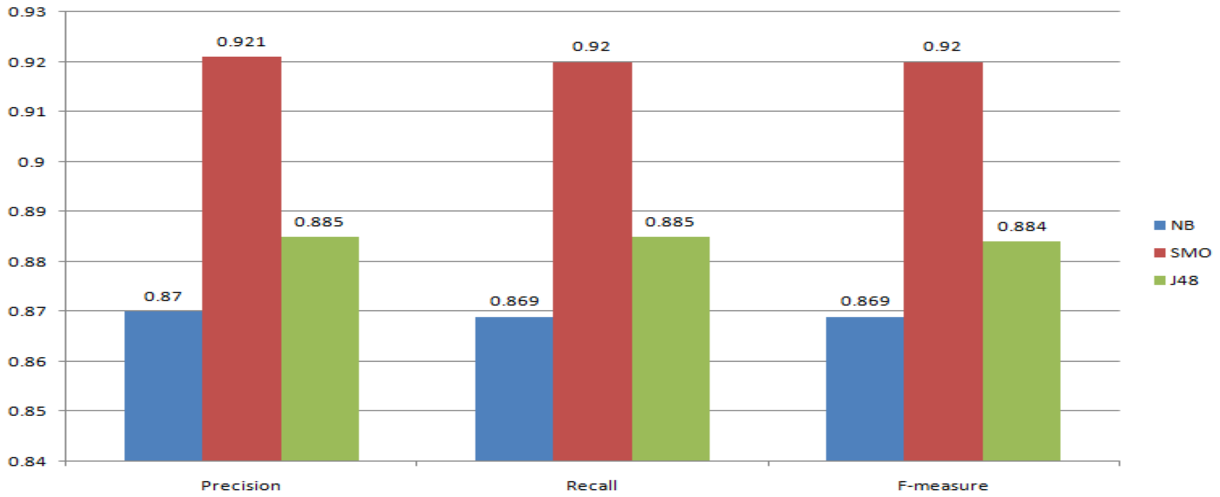


Figure 4.5: Weighted average of Precision, Recall and F-measure for NB, SMO and J48

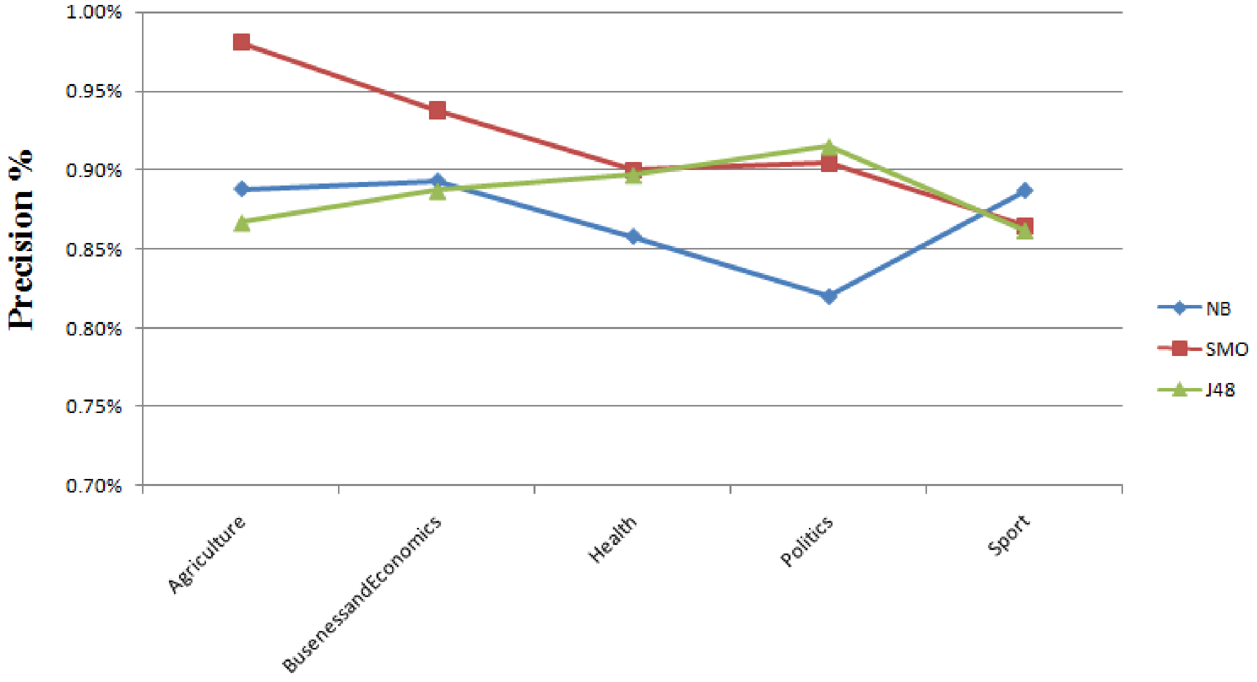


Figure 4.6: precision of all classes for NB, SMO and J48 classifiers

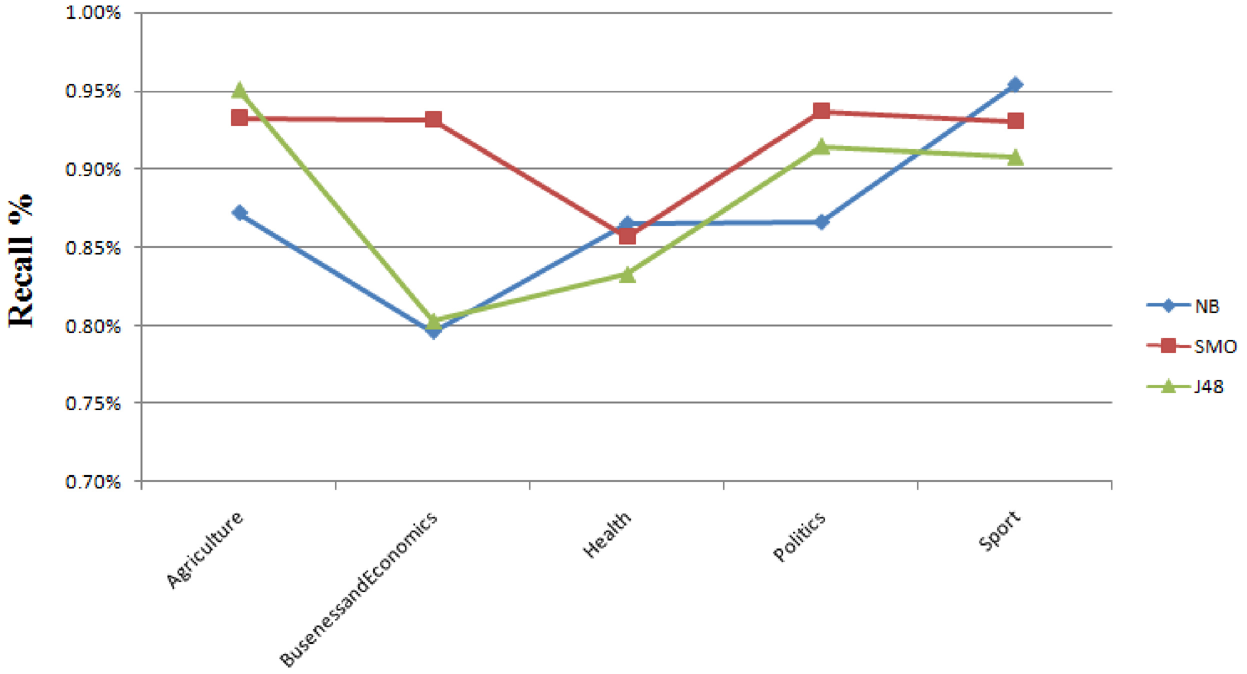


Figure 4.7: Recall of all classes for NB, SMO and J48 classifiers

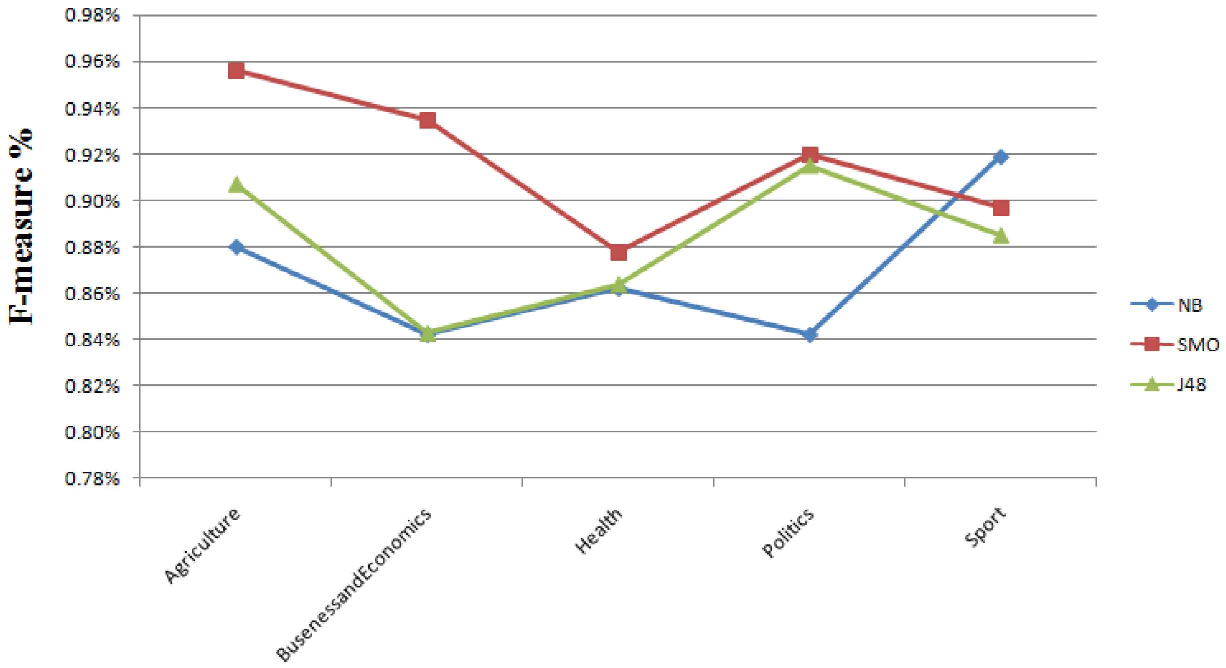


Figure 4.8: F-measure of all classes for NB, SMO and J48 classifiers

# CHAPTER FIVE

# CONCLUSIONS AND RECOMMENDATIONS

## 5.1 Conclusions

This thesis work has presented the application of NB, SVM and J48 decision tree classifier technique for Afan Oromo News Text Categorization (AONTC) using machine learning technique. The validity of the approachwas experimentally demonstrated. Experiments have been performed with the help of training and testing documents sets which are obtainedfrom ORTO and VOA AfaanOromoo program websites. Certain preprocessing techniques were applied on the documents and the implementation of the classifiers is followed after as clearly explained in Chapter 3. All the experimental results obtained aretabulated, and evaluated in as discussed in Chapter 4.

The objective of the research is to test the applicability of machine learning approach to Afan Oromo news text categorization and the result achieved proved the possibility of the approach.

Finally, classifiers was trained and tested on the dataset grouped in to 66% and 36% for training and testing respectively.The result obtained from the experiments shows NB classified 617 instances correctly which is 87%of testing instances and 93 instances or 13% of testing instances were incorrectly classified**.** In addition to this NB registered 87% precision, 87% recall and 87% F-measure. SMO classified correctly653 instances or 92% of testing instances while 57 instances are incorrectly classified. The third classifier J48 correctly classified 88.45% or 628 instances and incorrectly classified the remaining 82 instances. So from the experiment the best result is registered by SMO classifier with 92% accuracy, precision of 92% and recall of 92%.

Next to SMO, J48 recorded best accuracy 88.5%, precision 88.5% and 88.5% while the NB is least with 87% accuracy, precision 87% and recall 86.9%. But from the experiment we can understand that, all classifiers perform good result.

## 5.2 Recommendations

The result of this research indicated that machine learning approach of NB, SMO and J48 text classifier is applicable for automatic Afan Oromo news text categorization. However, continuous researches have to be conducted to get better results. So, I recommend the following points.

- The availability of standardized text corpus facilitates text categorization researches. Nevertheless, there is no established text corpus for text classification purposes. Hence, I recommend the need to develop Afan Oromo text corpus.
- This work is tested by using supervised machine learning approaches of NB, SMO and J48 classifier algorithm to automatically classify Afan Oromo news articles. There are also many other algorithms available for automatic text classification either supervise or unsupervised machine learning algorithms. Hence, using different machine learning algorithms (supervised or unsupervised) have to be tested for Afan Oromo.
- Stemmer did not apply for this research. I recommend for future to apply stemmer and measure the performance of the classifier.
- This research considers the single-label classification which assigns a given document only to one category. There may be a need one document may be assigned to more than one category (multi-label classification). Accordingly, research has to be done on multi-label classification for Afan Oromo texts.

# REFERENCES

Abdullah, H., and Mohammed, Al-Kabi., 2012. Comparative Assessment of the Performance of Three WEKA Text Classifiers Applied to Arabic Text. Vol. 21, No. 1, pp. 15- 28.

Abe.S., 2005. Support Vector Machines for Pattern Classification. Advances in Pattern Recognition. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Addis, A., 2010.Study and Development of Novel Techniques for Hierarchical Text categorization. PhD thesis, University of Cagliari, Italy.

Aditya, C., 2010. A Comparative Study on Text Categorization. Master's thesis in Degree of Computer Science School of Computer Science Howard R. Hughes College of Engineering University of Nevada, Las Vegas.

Alemu, K., 2010. Hierarchical Amharic News Text Classification. MSc Thesis. Addis Ababa University, Addis Ababa, Ethiopia.

Alsaleem, S., 2011.Automated Arabic Text Categorization Using SVM and NB,International Arab Journal of e-Technology, Vol. 2, No. 2.

Androutsopoulos, I.,Koutsias, J., Chandrinos, V., And Spyropoulos, C., 2000. An experimental comparison of naive Bayesian andkeyword-based anti-spam filtering with personal e-mail messages. In Proceedings of SIGIR-00, 23rdACM International Conference on Research and Development in Information Retrieval (Athens, Greece, 2000).

Barber, D., 2010.Bayesian Reasoning and Machine Learning.

Belkin, N., And Croft, W. B., 1992. Information filtering and information retrieval: two sidesof the same coin? Commun.ACM 35, 12.

Bishop, C., 2006. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Borko,H., and Bernick, M., 1963.Automatic documentclassification. J. Assoc. Comput.

Mach. 10, 2.

Boyapati, V., 2002.Improving hierarchical text classification using unlabeled data.

Burges, C., 1998. “A Tutorial on Support Vector Machines for Pattern Recognition”,Data Mining and Knowledge Discovery, Vol. 2, No. 2, pp. 121–167.

Cagri, T., 2011.Text Categorization and Ensemble Pruning in Turkish News Portals.Master’s thesis, department of Computer Engineering and Graduate School of Engineering, BilkentUniversity, Turkey.

Caruana, R., Niculescu-Mizil, A., 2006. An empirical comparison of supervised learning algorithms, In the Proc. of the 23rd int. conf. on Machine learning.

Changuel, S., Labroche, N., and Bouchon-Meunier, B., 2009. Automatic web pages author Extraction. LNAI 5822, Springer-Verlag Berlin Heidelberg.

Chen, H., 2009. Machine Learning for Information Retrieval, Neural Networks.

Cohen,W., 1995. Text categorization and relational learning. In 12th International Conference on Machine Learning (I CML)

Cortes, C., and Vapnik, V., 1995. Support-vector networks: Machine Learning, 20(3):273–297.

Crammer, K., and Singer, Y., 2003.A Family of Additive Online Algorithms for Category Ranking.

Debela, T., 2010.Designing a Stemmer for Afan Oromo Text: A hybrid approach. Master’s thesis, School of graduate studies, Addis Ababa University, Ethiopia.

Deerwester, S.T.,Dumais, T.K,. Landauer, G.W., Furnas, and R.A. Harshman., 1990.Indexing by latent semantic analysis. Journal of the American Society of Information Science, 41(6):391-407.

Drucker, H., Vapnik,V., and Wu, D., 1999. Automatic text categorization and its applications to text retrieval.

Duda, R., Hart, P., Stork, D., 2001."Pattern Classification", 2nd Ed, Wiley Interscience.

Dumais, S., Platt, J., Sahami, M. & Heckerman, D. 1998. Inductive learning algorithms and representations for text categorization.

Durga, B., and Dr .Venu, G., 2012. Text Categorization and Machine Learning Methods: Current State of the Art. Global Journal of Computer Science and Technology Volume XII, Issue XI, Version I, 37.

Duwairi, R., 2007. "Arabic text Categorization", In the Int. Arab journal of information technology, 4, 2.

Edel, G., 2004. Automatic Text Categorization of Racist WebPages, a thesis submitted to Dublin City University, for the degree of Master of Science.Dublin City University.

Escudero, G., M`arquez, L., And Rigau, G., 2000.Boosting applied to word sense disambiguation. In Proceedings of ECML-00, 11th European Conference on Machine Learning (Barcelona, Spain, 2000), 129–141.

Fang,Y., Parthasarathy, S., and Schwartz, F., 2001. Using clustering to boost text classification, ICDM Workshop on Text Mining (TextDM'01).

Feldman, R., Sanger, J., 2007.The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press,

Field, B., 1975. Towards automatic indexing: automatic assignment of controlled-language indexing and classification from free indexing. J. Document. 31, 4, 246–265.

Gale, W. A., Church,K.W., AND Yarowsky, D. 1993. A method for disambiguating word senses in a large corpus. Comput.Human. 26, 5, 415–439.

Gebrehiwot, A., 2011. A Two Step Approach for Tigrigna Text Categorization. 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 University, Addis Ababa, Ethiopia.

George, H.,Kohavi, R., and Pfleger,K., 1994. Irrelevant Features and the Subset Selection Problem.In ICML, pages 121-129.

Gray,W.A., and Harley, A. J., 1971.Computer-assisted indexing. Inform. Storage Retrieval.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann,P., Witten, I. 2009. The WEKA Data Mining Software: An Update. SIGKDD Explorations, 11(1).

Hand, D., Mannila, H., and Smyth, P., 2001.Principles of Data Mining (Adaptive Computation and Machine Learning). MIT Press.

Han, J., and Kamber, M., 2006. Data Mining: Concepts and Techniques, (2nd Ed), the Morgan Kaufmann Series in Data Management Systems.

Hayes, P. J., Knecht, L. E., and Cellio, M.J.,1988.A news story categorizationsystem. In Proceedings of the second conference on applied natural language processing, ANLC ’88, pages 9 – 17, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hayes,P.J.,Andersen, P. M., Nirenburg,I.B., AND Schmandt, L. M. 1990. Tcs: a shell for content-based text categorization. In Proceedings of CAIA-90, 6th IEEE Conference on Artificial Intelligence Applications (Santa Barbara, CA, 1990), 320–326.

Heaps, H., 1973.A theory of relevance for automatic document classification. Inform Control 22, 3, 268–278.

Hill, T., and Lewicki P., 2007. Statistics Methods and Applications, (1st Ed), StatSoft, Tulsa, OK.

Ingwersen, P., 2002. Information Retrieval Interaction, 1st ed. London: Taylor Graham Publishing.

Irina, R., 2001. An Empirical Study of the Naïve Bayes Classifier, Proc. of the IJCAI-01 Workshop on Empirical Methods in Artificial Intelligence.

Isa, D., Lee, L.H., Kallimani, V.P., and RajKumar, R., 2008. Text document pre-processing with the Bayes formula for classification using the support vector machine

Joachims,T., 1998. Text categorization with Support Vector Machines: Learning with many relevant features. Proceedings of ECML-98, 10th European Conference on Machine Learning, pp. 137-142.

Joachims, T., 2002.Learning to classify text using SVM", Kluwer Academic Publishers.

Kanaan, G., Al-Shalabi, R., Ghwanmeh, S., Al-Ma'adeed, H., 2009. A comparison of text-classification techniques applied to Arabic text, Journal of the American Society for Information Science and Technology, 60(9).

Kim, S., Han, K., Rim, H., and Myaeng, S., 2006.Some effective techniques for naïve bayes text classification. IEEE Transactions on Knowledge and Data Engineering, vol. 18, no.Proceedings of SIGIR.

Kula, K., Varma, V., and Pingali, P., 2008. Evaluation of Oromo-English Cross-Language Technologies Research Center. Information Retrieval IIIT, Hyderabad, India.

Lafferty, J., McCallum, A., and Pereira, F., 2001.Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in `ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Lan, M., Tan, C., Low, H,, and Sung, S., 2005. A comprehensive comparative study on term weighting schemes for text categorization with support vector machines. In Posters Proc. 14thInternational World Wide Web Conference, pages 1032–1033.

Larkey,L.S., and Croft, W. B., 1996.Combining classifiers in text categorization. In Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval (Zurich, Switzerland, 1996), 289–297.

Last, M., Markov, A., and Kandel, A., 2008. Multi-lingual Detection of Web Terrorist Content, In: Chen, H. (Ed.), WISI, Lecture Notes in Computer Science, Springer - Verlag, 3917 16-30.

Lewis, D., Yang, Y,.Tony, G.,and Li, F.,2004.A New Benchmark Collection for Text Categorization Research.

Lewis, D. D., 1992. An evaluation of phrasal and clustered representations on a text categorization task. In Proceedings of SIGIR-92, 15th ACM International Conference on Research and Development in Information Retrieval (Copenhagen, Denmark, 1992), 37–50.

Li,H., and Jain, A. K., 1998.Classification of text documents.Comput. J. 41, 8, 537–546.

Liddy, E. D., Paik,W., and Yu, E.S., 1994. Text categorization for multiple users based on semantic features from a machine-readable dictionary. ACM Trans. Inform. Syst. 12, 3, 278–295.

Maribor, S., 2007.Text Categorization for Multi-label Documents and Many Categories.

Maron, M., 1961. Automatic indexing: an experimental inquiry. J. Assoc. Comput. Mach. 8,3,404–417.

Meyer, C., 2008. On Improving Natural Language Processing through Phrase-based andone-to-oneSyntactic Algorithm, Msc. Thesis, Kansas State University Manhatan, Kansas.

Mitchell, T., 1997.Machine Learning.The MIT Press.

Nils, J., 1998.Introduction to Machine Learning: An Early Draft of a ProposedTextbook. Stanford University, Stanford, CA 94305.

Norbert, F., and Buckley,C., 1991. A probabilistic learning approach for document indexing.ACM Transactions on Information Systems, 9, 3.

Porter, M. F., 1980. An algorithm for suffix stripping, Program, 14, 3, 130-137.

Quinlan, J., 1998. Data mining tools See5 and C5.0. Technical Report, RuleQuest Research.

Quinlan, R., 1993. “C4.5: Programs for Machine Learning”, Morgan Kaufmann Publishers, San

Mateo, CA.

Rabiner, R., 1989. `"A tutorial on hidden markov models and selected applications in speech Recognition", Proceedings of the IEEE 77(2), 257-286.

Russel, S., and Peter, N., 1995.Artificial Intelligence: A Modern Approach, New Jersey: Prentice Hall,

Sable, C.,2003. Robust Statistical Techniques for the Categorization of Images Using Associated Text.PhD thesis, Columbia University.

Salton, G., C. Yang, and A. Wong, 1975. A Vector-Space Model for Automatic Indexing, Communications of the ACM, Vol. 18, No. 11, pp. 613–620.

Saracevic, T., 1975. Relevance: a review of and a framework for the thinking on the notion in information science. J. Amer. Soc. Inform. Sci. 26, 6, 321–343.

Schütze, D., and Pedersen,J.O., 1995.Comparison of classifiers and document representations for the routing problem. In the Proceedings of the 15thAnnual International ACM SIGIR International Conference on Research and Development in Information Retrieval, pages 229-237, Seattle, Washington, USA.

Sebastiani, F., 2002.Machine Learning in Automated text classification.In ACM Computing Surveys.Vol. 34 No.1.

Sheth, B., and Maks, P., 1993. Evolving agents for personalized information filtering. In Proceedings of the IEEE CAIA-93. IEEE, New York, 345-352.

Sorensen, A., and O'Riordan,C., 1997.Profiling with the INFOrmertext filtering agent. J.UCS: Journal of Universal Computer Science, 3(8): 988-999.

Surafel, T., 2003.Automatic Categorization of Amharic News Text: a machine learning Approach, Master Thesis at SISA.Addis Ababa University.

Tang, L., Rajan,V.K., Narayanan, 2009. Large Scale Multi-Label Classification via MetaLabeler, In Proceedings of the Data Mining and Learning.

Vapnik, V. N., 1995. The nature of statistical learning theory, Springer-Verlag New York,Inc., New York, NY, USA.

Warner. B., and Misra, M., 1996.Understanding neural networks as statistical tools. The American Statistician, 50(4):284–293.

William, C., and Yoram, 1999. Context-sensitive learning method for text categorization, Proc. of SIGIR 96, 19th International Conference on Research and Development in Informational Retrieval, vol. 17, Issue 2, pp-307-315.

Witten, I. and Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques. San Francisco, USA: Morgan Kaufmann.

Wongpun, S., and Srivihok, A., 2008. Comparison of Attribute Selection Techniques and Algorithms in Classifying Bad Behaviors of Vocational Education Students, In proceedings of 2nd IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST), Australia, 526-531.

Yang, Y., and Liu.X., 1999.A re-examination of text categorization methods,Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR'99), pp 42--49,.

Yang, Y., 1999.An evaluation of statistical approaches to text categorization. Inform. Retr. 1,1–2, 69–90.

Yang, Y., Zhang, J., and Kisiel, B., 2003. A scalability analysis of classifiers in text categorization. In Proc. of the 26th ACM International Conference on Research and Development in Information Retrieval (SIGIR'03), pages 96–103, Toronto, Canada, ACM Press, New York, United States.

Yohannes, A., 2007. Automatic Classification of Amharic News Text. MSc Thesis. Addis Ababa University, Addis Ababa, Ethiopia.

Zhang, H., 2004. The Optimality of Naïve Bayes, In FLAIRS2004 conference.

Zhang, W., Yoshida, T., and Tang, X., 2007.Text classification using multi-word features. In proceedings of the IEEE international conference on Systems, Man and Cybernetics, pp. 3519 – 3524.

Zhang, T., and Frank, J. Oles.,2000.Text categorization based on regularized linear classification methods. New York,

Zurada, J., 1992. Introduction to Artificial neural systems. West publishing company, Saint Paul, Minnesota.

# APPENDIX

## Appendix I: Afan Oromo stop word lists

aanee
agarsiiso
akka
akkam
akkasum
akkum
akkuma
ala
alatti
alla
amma
ammo
ammoo
an
ana
ani
ati
bira
booda
booddee
dabalate
dhaan
dudduub
dugda
dura
duuba
eega
eegana
gararraa
garas
garuu
giddu
gidduu
gubbaa
ha
hamma
hanga
henna
hoggaa
hogguu
hoo
hoo
illee
immoo
ini
innaa
inni
irra
irraa
irraan
isa
isaa
isaaf
isaan
isaani
isaanii
ishiirraa
ishiitti
ishiitti
isii
isiin
isin
isini
isinii
isiniif
isiniin
isinirraa
isinitti
ittaanee
itti
itumallee
ituu
ituullee
jala
jara
jechaan
jechoota
jechuu
jechuun
kan
kana
kanaa
kanaaf
kanaafi
koo
kun
lafa
lama
malee
manna
maqaa
moo
na
naa
naaf
naan
naannoo
narraa
natti
nu
nu'i
nurraa
nuti
nutti
nuu
nuuf
nuun
nuy
odoo
ofii
oggaa
oo
siin
silaa
silaa
simmoo
sinitti
siqee
sirraa
sitti
sun
tahullee
tana
tanaaf
tanaafi
tanaafuu
ta'ullee
ta'uyyu
ta'uyyuu
tawullee
teenya
teessan
tiyya
too
tti
utuu
waa'ee
waan
waggaa
wajjin

eegasii
ennaa
erga
ergii
f
faallaa
fagaatee
fi
fullee
fuullee
gajjallaa
gama
isaaniitiin
isaanirraa
isaanitti
isaatiin
isarraa
isatti
isee
iseen
ishee
ishii
ishiif
ishiin
kanaafi
kanaafuu
kanaan
kanaatti
karaa
kee
keenna
keenya
keessa
keessan
keessatti
kiyya
osoo
otoo
otumallee
otuu
otuullee
saaniif
sadii
sana
saniif
si
sii
siif
warra
woo
yammuu
yemmuu
yeroo
yommii
yommuu
yoo
yookaan
yookiin
yoolinimoo
yoom

## Appendix II: List of Afan Oromo special words which co-occur with other words.

Al

Sab

Saal

Wal

Kor

Kal

Man

Rog

Tarm

Qar

Gar

Bar

Bir

Gar