Application of Data Mining Classification Algorithms for Afaan Oromo Media Text News Categorization

Etana Fikadu Dinsa[1], Ramesh Babu P[2]

[1 2]College of Engineering & Technology, Wollega University, Post Box No: 395, Ethiopia

Abstract: *This research proposes a model, Afaan Oromo Text categorization, which helps to automatically categorize texts to predefined classes. Text categorization is the task of assigning an electronic document to one or more categories, based on its contents. Document classification can be done manually or automatically. Manual text categorization is carried out by human experts. It requires a certain level of vocabulary recognition and knowledge processing. Automatic classification is a process of classifying documents into a number of classes using machine learning methods. Automatic document categorization reduces searching time, thereby facilitating the searching process.*

In this research, we deal with Itemset method based Afaan Oromo news document categorization using Apriori Algorithm. In text document categorization, each word contained in a document is referred as item. As a part of this work, apriori algorithm is used for generating frequent item in a given text document. Among the automatic classifiers which are applicable on high dimensional data, two of them; Naïve Bayes (NB) and bayes networking have been experimented on the Total data. The data the pre-processed Afaan Oromo text items is organized into categories of nine classes for the experimentation purpose and the experimentation uses 10-fold stratified cross validation for training and test data. The performance of the classification is analyzed to measure the accuracy of the classifiers in categorizing the Afaan Oromo news documents in to specified categories.

The best result obtained by bayes networking Classifier is 97.15% and Naïve Bayes (NB) is 95.666% on nine categories data. This research indicated that bayes networking Classifier is more relevant for categorizing Afaan Oromo news document.

Keywords — *Categorization, Data mining, Apriori, Classifier algorithms, Machine learning and Itemset.*

I. INTRODUCTION

Along with the continuously increasing volume of information availability on the Web, there is a growing interest in getting better ways of accessing these resources. As the amount of information has dramatically grown, finding relevant information among the millions of information resources on the Web is becoming more difficult. Users are currently restricted to browse or follow hyperlinks from one web page to the next and syntactic keyword searches for finding significant information (Abraham, 2013).

Manual Organization of very large volume of electronic information will not be feasible. So, to overcome this problem a mechanism is required for finding, filtering and managing the rapid growth of online information. This mechanism is called text categorization (Meron, 2009 A definition for text categorization can be found at (Sebastini, 2007): "text categorization (TC also known as text classification or topic spotting) is the task of automatically sorting a set of documents into categories (or classes, or topics) from a predefine set".

There are various reasons for using document categorization. Automatic document categorization reduces searching time, thereby facilitating the searching process. Moreover, it facilitates access, when documents are classified based on their concept similarity; we can get hint about what the document actually contains without going through it. Document classification can be done manually or automatically. Manual text categorization is carried out by human experts. It requires a certain level of vocabulary recognition and knowledge processing. There are some problems observed with manual classification. It requires intensive human labor and affects classification results because of inconsistency due to variation in perception, comprehension, and judgment, and for the current Web based knowledge management it is almost impossible. In contrast, automatic classification is a process of classifying documents into a number of classes using machine learning methods (Abraham, 2013).

Machine Learning is defined as "the ability of a machine to improve its performance based on previous results". In other words it is a system capable of learning from experience and analytical observation, which results in continuous self-improvement there by offering increased efficiency and effectiveness (Nigam K, 2008). In general there are four different types of machine learning techniques. They are:

1. Supervised learning.
2. Unsupervised learning.
3. Semi-supervised learning and
4. Reinforcement learning

This research deals with text categorization which is a supervised learning technique. Supervised learning: supervised learning is a machine learning technique that learns from training data set. A training data set consists of input objects, and categories to which they belong. Assigning categories to input objects is carried out manually by an expert. Given an unknown object, supervised learning technique must be able to predict an appropriate category based on prior training.

The aim of news categorization is to assign predefined category labels to incoming news articles. New documents are assigned to pre-defined categories by using a training model which is learned by a separate training document collection (Cagri, 2011). A text categorization approach that uses the unlabeled text collections is called document (text) clustering. Text clustering is used to assign some similar properties of text documents into automatically created groups. It is used to improve the efficiency and effectiveness of text categorization system such as time, space, and quality (Meron, 2009). The standard text clustering algorithms can be categorized into partitioning and hierarchical clustering algorithms. Partitioning clustering algorithm splits the data points into k partition where each partition represents a cluster. Whereas hierarchical clustering algorithm groups data objects to form a tree shaped structure. It can be bottom up approach which each data points are considered to be a separate cluster and clusters are merged based on a criteria or top down approach where all data points are considered as a single cluster and they are splited into number of clusters based on certain criteria (Meron, 2009).

While there are a number of different texts categorization approaches. Even though the work will focuses on classification algorithm to categorize Afaan Oromo documents in to predefined categories.

A. Afaan Oromo Qubee

The alphabet of Afaan Oromo is often called “Qubee Afaan Oromoo”, alphabets of the Oromo language. The Afaan Oromo alphabet characterized by capital and small letters as in the case of the English alphabet. The major representatives of sources of the sound in a language are the vowels and consonants. Afaan Oromo has 36 basic sounds (10 vowels and 26 consonants) (Wakshum, 2000). Afaan Oromo is a phonetic language, which means that is spoken in the way it is written.

The Afaan Oromoo vowels represented by letters (a, e, o, u and i) are called “Dubbifttuu” in Afaan Oromo and the consonants known as “dubbifamaa” in Afaan Oromo are shown in the following tables 1 and 2 respectively, together with their main articulators. According to (Ladefoged,1955), quoted in (Morka, 2001) Some of the finer anatomical feature involved in speech production include the vocal cords, velum, tongue, teeth, palates, the alveolar ridge, the mouth, and lips. These anatomical components move to different positions to produce various sounds and are referred in articulators.

B. The Afaan Oromo Vowels

Afaan Oromo basically has 10 phonemic vowels, five short and five long vowels, indicated in the orthography by doubling the five vowel letters. Vowel can appear in initial, medial and final positions in a word in Afaan Oromo language. A long vowel is interpreted as a single unit and occurs everywhere a short vowel can occur.

The following examples show some of long vowels at word initial, medial and final positions.
Initial positions: uumaa to mean ‘nature’, eelee to mean ‘pan’, Medial position: keennaa to mean ‘gift’, leexaa to mean ‘single’

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

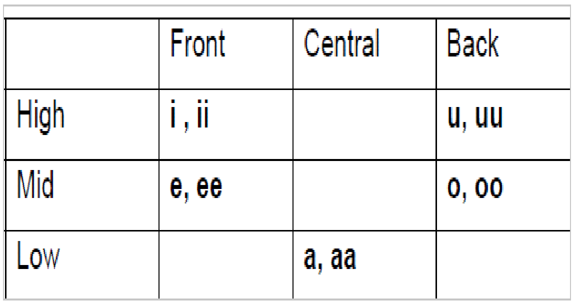
Short and long vowels in word medial position

/i/	[bira]	nearby	[bi:raa]	beer
/e/	[kel:a:]	fence	[ke:l:o:]	wild grass
/a/	[rafu:]	sleep	[ra:fu:]	cabbage
/o/	[boru]	tomorrow	[bo:ru:]	unclean water
/u/	[muka]	wood	[tu:ta]	crowd

All Oromo vowels, both short and long, can occur word initially, medially or finally, though the rounded vowels (/ o / and / u /) only rarely occur in short form in word final position. (Teferi Degeneh, 2015).

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

Table 1 Afaan Oromoo Vowels

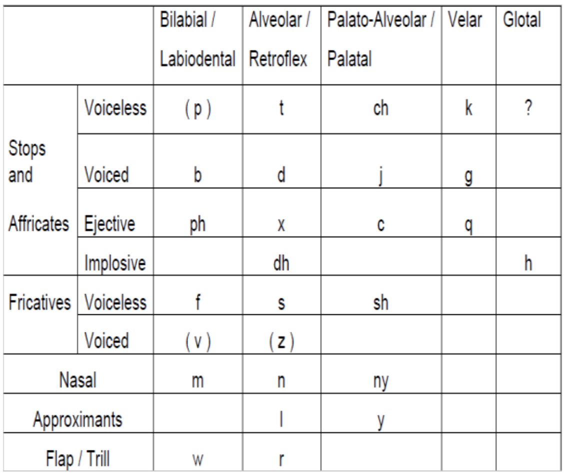


	Front	Central	Back
High	i , ii		u, uu
Mid	e, ee		o, oo
Low		a, aa	

Afaan Oromo vowels are pronounced in sharp and clear fashion which means each and every word is pronounced strongly. For example:
A: Farda, Haadha
E: Gannale, Waabee, Noole, Roobale, colle
I: Arsii, laali, Rafi, Lakki, Sirbbi
O: Oromo, Cilaalo, Haro, caancco, Danbidoollo
U: Ulfaadhu, Guddadhu, dubbadhuu, arba guugu, Ituu

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

Table 2 The Afaan Oromo consonants



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

The consonants p, v, and z only occur in loan words. Because, there are no native words in Afaan Oromo that formed from these characters. However, in writing Afaan Oromo language, they are used to refer to foreign words such as “polisii” (“police”).

In Afan Oromo, like in other languages there is word and sentence boundaries, the blank character (space) shows the end of one word (Workineh Tesema, 2017). Moreover, parenthesis, brackets, quotes are being used to show a word boundary. Furthermore, sentence boundaries punctuations are almost similar to English language i.e. a sentence may end with a period (.), a question mark (?), or an exclamation point (!)(Getachew, 2014).

C. Afaan Oromoo Gemination

Gemination happens when a spoken consonant is pronounced for an audibly longer period of time than a short consonant. Gemination is also significant in Afaan Oromo. That is, consonant length can distinguish words from one another; for example, the word “baruu” in Afaan Oromo is to mean ‘to learn’ is ungeminated (called ‘laafaa’ in Afaan Oromo) , whereas the word “barruu” in Afaan Oromoo is to mean ‘palm of hand’ is geminated (called ‘jabaa’ in Afaan Oromo). The difference between the two words “baruu” and “barruu” is the number of consonants appearing together, which makes a difference in meaning.

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

On the other hand, few words are stressed on the first syllable. These words always have a combination consonant.
Examples: Nyaadhu, Nyaara , Dhugaa

II. DATA SET PREPARATION

Documents which contain Afaan Oromo will be collected from different sources to evaluate the Afaan Oromo text categorization was selected and Prepared as there is no previous research and corpora in Afaan Oromo for evaluating Afaan Oromo text categorization. The corpus size is 1101text documents written in Afaan Oromo language and with Latin alphabets which is called Qubee‘.

The corpus is built from the official website of Oromia broadcast network (OBN), Oromia media network(OMN), Voice of America Radio Afaan Oromo language and other Internet based sources. The prepared corpus consists of 9 news text categories items written on different topics. Finally, performances of the system were tested by conducting different experiments using the collected documents.

III. DESIGN AND DEVELOPMENT OF THE AUTOMATIC CATEGORIZER FOR AFAAN OROMO TEXTS

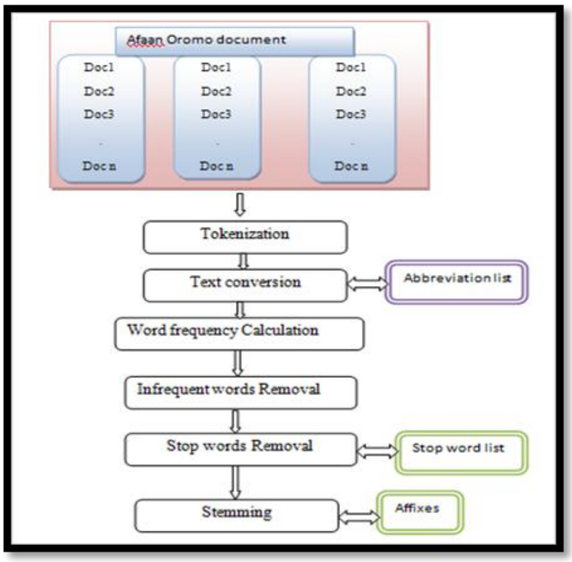
In the design and implementation process of itemset method automatic Afaan Oromo document categorization by using Apriori algorithms to generate frequent items. The major activities in this thesis are pre-processing, generating frequent items, processing and categorizer phase. Pre-processing includes tokenization, stop-word removal, stemming, and keyword term selection (shown in fig.1). Then it uses different text classifier algorithms to predict the documents to their predefined categories.

This research involve different subjects like Accident, Gada system, religion, politics, education, health, Economy, sport and agriculture. For this study 1101 different textual documents were collected from different news media such as Oromia Radio and TV website, VOA (Voice of America) Afaan Oromo service, www.Oromiya.com, online educational resources and other resources sources in Afaan Oromo available on the web.

Hence, text classification model is developed from the training data sets using bayes net and Naive Bayes network classifiers. Here after, the developed model is evaluated using the test data sets. Finally, the system comes with the category of Afaan Oromo news text documents model.

Data Pre-processing

In order to convert both training as well as test data from the original data to suitable data as an input to Apriori algorithm, there are several operations required. Fig.1 describes the block diagram of data processing phase which was applied in this research.



Tokenization

The corpus which is a set of sentences first tokenized into words. Since, Afaan Oromo uses Latin alphabet the sentences can split using similar word boundary detection techniques like the use white space in English. Tokenization is the process of breaking parsed text into pieces, called tokens (Birmingham, 2010). During this phase punctuations and any non Afaan Oromo characters are removed and only Afaan Oromo words are selected from each document. For example consider the sentence "Immammataa fi deemsa hojii barnootaa Oromiyaa keessaa irratti yeroo dhaa yerootti qorannaan nigeggeeffama. " from a document which belongs to category education is tokenized in to a set of words on the white space as shown below.

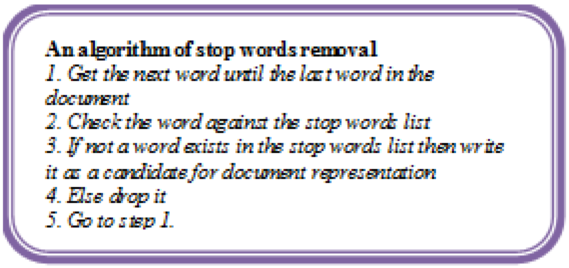
'Imammataa' 'fi' 'deemsa' 'hojii' 'barnootaa' 'Oromiyaa' 'keessaa' 'irratti' 'yeroodhaa' 'yerootti' 'qorannaan' 'nigeggeeffama'.

In Afaan Oromo language, it is common to write some words in shorter form using "/" (forward slash) or "." (dot). In this preprocessing phase the short forms of words are automatically identified and manual replacement to the appropriate form has been done.

After text conversion, removing rare words has been done. To do this first of all frequency of each words determined and those words which have frequency less than 10 considered as rare words. Those words identified as rare words excluded from further processing and the items those have frequency above 10 is used to predict the text categories.

Stop word Removal

After tokenization take place, we have removed Afaan Oromo stop words, hence it has no effect on meaning of the words. In this work, Stop word removal is used to remove stop words from the selected contexts because the absence or presence of these words has no contribution to identify appropriate sense. Not all tokenized words are necessary for this work hence, one word carry the meaning than other words and other words that have no own meaning. For instance, words such as ('as', 'achi', 'irra', 'keessaa', 'jala'), conjunctions ("'fi', 'akkasumas', 'kana malees'). Since stop words do not have significant discriminating powers in the meaning of ambiguous words; we filtered stop-words list to ensure only content bearing words are included. To store stop words there is no system which is developed has a sub-system (module) which can load all the available Afaan Oromo documents. So, that manually the stop words are selected and stored in the individual corpus. Once those words are identified as stop words, they were not further processed. Since the influence of those irrelevant words for the categorization purpose is illuminated, it helps to save both time to process the texts and storage space.



An algorithm of stop words removal

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

After stop words were cleaned from corpus stemming operation was implemented on corpus using python programming language.

Processes to Generate Frequent Itemsets
There are two steps to categorize documents in this research. These steps are; find all frequent itemsets and generate strong association rules from frequent itemsets. To do those two important activities the following diagrammatically shown high level design is applied.

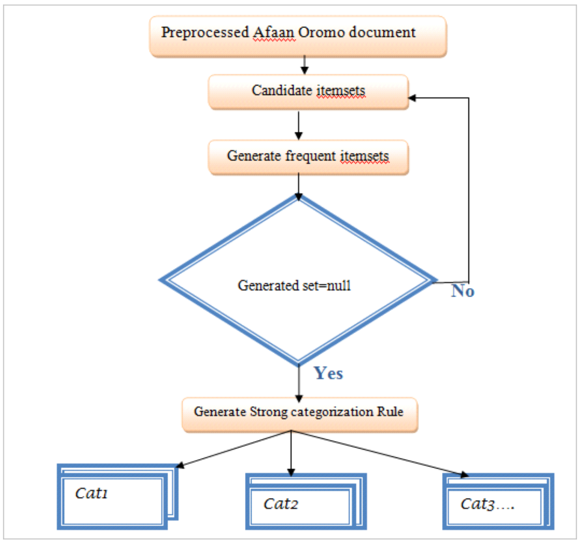

Preprocessed Afaan Oromo document
Candidate itemsets
Generate frequent itemsets
Generated set=null
No
Yes
Generate Strong categorization Rule
Cat1
Cat2
Cat3….

Fig .2 Model for document categorization using itemset method

The following text describes the low level extended Apriori algorithm design to associate terms with document categories;

Input: The dataset and minimum support (min_sup)
Output: The maximum frequent itemset

1. Scan the transaction database to get the frequency of each term.
2. K 1.
3. Find frequent itemset, Lk from Ck, the set of all candidate itemset.
4. Form Ck+1 from Lk.
5. Prune the frequent candidates by removing itemset from Ck whose elements with less than 10 times in Lk..
6. Modify the entry in memory to be zero for the itemset which are not occurring in any of the candidates in Lk .
7. Check the Size of Transaction (ST) attribute and remove transaction from data base where ST<=k.
8. K k+1.
9. Repeat 5-8 until Ck is empty or transaction database is empty.

Step 5 is called the frequent itemset generation step. Step 6 is called as the candidate itemset generation step and step 7-9 are prune steps. Details of first two steps are described below. Frequent itemset generation: Scan D and count each itemset in Ck, if the count is greater than min_sup, then add that itemset to Lk.

Candidate itemset generation: For k = 1, C1 = (all itemset of length = 1).
For k > 1, generate Ck from Lk-1 as follows:
The join step:
Ck = k-2 way join of Lk-1with itself.
If both {a1,..,ak-2, ak-1} & {a1,.., ak-2, ak} are in Lk-1, then add {a1,..,ak-2, ak-1, ak} to Ck.

After the preprocessing phase building an inverted index was done. Inverted Index (Ajit Kumar, etal, 2011), is an indexing approach which can help to map a data with a given content. In this research inverted index is applied to produce data structure of each term in a way which indicates where it is located. There are two types of inverted indexes (Ajit Kumar et al, 2011), these are referred as record level inverted index and word level inverted index. In this work, record level inverted index is applied. A record level inverted index contains a list of references to documents for each term. Whereas a word level inverted index additionally contains the positions of each word within a document. The following simple example illustrates the concept of Inverted Indexing. Assume that there are 4 documents named D1, D2, D3 and D4. The content of each of the document is presented as follows:

D1: Namtichi kitaaba dubbisaa jira.
D2: Namtichi kitaaba dubbisuu jaallata
D3: Baratichi kitaaba qaba.
D4: Baratichi fi namitichi fedhii kitaba dubbisuu qabu.

To build inverted index, it is required to select each term from all the documents and keeping record of in which documents the terms are found. Table 3 shows how a simple inverted index can be generated for the above four documents. For the term "Kitaaba" document identification code 1, 2, 3 and 4 used. This is because documents D1, D2, D3 and D4 contain the term “Kitaaba" which is categorized under education with maximum frequency.

Table 3 a record level term index

Items	Document identification code
Namtichi	{1,2,4}
Kitaaba	{1,2,3,4)
Dubbisaa	{1}
Dubbisuu	{1,2}
Jaallata	{2}
Baratichi	{3,4}
Fedhi	{4}
Qaba	{3}
Qabu	{4}

In the table 3 above examples the issue of stemming the Afaan Oromo terms is not considered. However in the actual experiment while examining the Apriori algorithm to categorize Afaan Oromo documents stemming was done for each term to find frequent items.

The following Table 4.2 shows few terms from the above document dataset along with their document frequencies

Table 4 shows few terms from document dataset along with their document frequencies

Items	Frequency of each word
Nam	3
Kitaaba	4
Dubbis	2
Jaallata	1
Baratichi	2
Fedhi	1
Qaba	2

As shown in above table 4 the document frequency of a term "kitaaba" is shown as 4 because it occurs in 4 documents out of 4. Here 4 is the sum of all terms (such as "kitaaba","kitaabicha" and etc.) which results the same root or stem according to the stemmer used

IV. EXPERIMENTAL RESULT AND DISCUSSION

Classification using Bayes Network Classifier Weka supports Bayes Network Classifier that is used for classifying numeric and nominal attributes. The Bayes Network classifier is used for experimentation. This classifier requires a small amount of memory and time. The Bayes Network classifies the 97.15 % of 923 instances correctly within 0.02 seconds as illustrated in table 6.

```
Correctly Classified Instances        896    97.15 %
Incorrectly Classified Instances       27     2.85 %
Total Number of Instances             923
```

Weka has a number of options for measuring the performance of a classifier out of them detailed accuracy by class and confusion matrix is shown by table 5 below for the Bayes Network classifier.

Table 5 confusion matrix using Bayes Network Classifier

```
=== Confusion Matrix ===
 a   b   c   d   e   f   g   h   i   <-- classified as
 61   0   0   0   0   1   0   0   0 |  a = accident
  0 135   1   0   0   0   0   1   0 |  b = agriculture
  0   4  89   0   0   1   0   0   0 |  c = economy
  0   0   5  89   0   3   0   0   0 |  d = education
  0   0   0   1  43   0   0   0   0 |  e = gadaa
  0   0   1   2   1 123   3   0   1 |  f = health
  0   0   0   0   0   1  87   0   0 |  g = politics
  0   0   1   0   0   0   1  91   1 |  h = religion
  0   0   4   0   0   4   0   1   167 |  i = sport
```

Based on the above confusion matrix, the performance of the classifier is shown in table 6 using precision, recall, F-measure and ROC-Area.

Table 6: performance of the classifier using precision, recall, F-measure and ROC-Area.

```
=== Detailed Accuracy by Class ===
             TP Rate   FP Rate   Precision   Recall  F-
Measure   ROC Area Class
              0.984     0            1          0.984
0.992     0.995    acident
              0.985     0.005     0.971     0.985     0.978
0.996    agriculture
              0.947     0.014     0.881     0.947     0.913
0.981    economy
              0.958     0.004     0.967     0.918     0.942
0.988    education
              0.977     0.001     0.977     0.977     0.977
0.989    gadaa
              0.999     0.013     0.925     0.939     0.932
0.982    health
              0.989     0.005     0.956     0.989     0.972
0.995    politics
              0.968     0.002     0.978     0.968     0.973
0.994    religion
              0.949     0.003     0.988     0.949     0.968
0.995    sport
Weighted Avg.    0.9715      0.006       0.96       0.959
0.959     0.991
```

Classification using Naïve bayes classifier

The testing on the data is done using 10-fold stratified cross validation. A total of 923 documents are used in the experiment. Out of this document 95.666% are correctly classified.

```
Correctly Classified Instances         883
95.6663 %
Incorrectly Classified Instances        40
4.3337 %
Total Number of Instances              923
```

Table 7 shows the comparison between Bayes net, and Naïve base networking used in this research.

Classifier	classified documents	No of documents	Their accuracy	Time taken
Bayes networking	Correctly classified documents	896	97.15 %	0.02 seconds
	In correctly classified documents	27	2.85	
Naïve base networking	Correctly classified documents	883	95.666 %	0.02 seconds
	In correctly classified documents	40	4.3337	

Generally, the results of the Naïve Bayes and Bayes Net, classifier algorithms were implemented for training text classification model depending on 9 main classes of documents. The performances of these classifiers are analyzed by applying various performance factors. Among those classifications algorithms, Bayes networking algorithm shows higher performance 97.15% and hence it was utilized for constructing classification model. From this work it was possible to conclude that machine learning techniques based on itemset method can be applied for Afaan Oromo text categorization.

V. CONCLUSIONS

This research attempts to develop Afaan Oromo text document categorization system based on itemset method. The developed prototype involves tokenizing, stop word removal and stemming. Indexing Afaan Oromo text document has common ground with that of the English langauge because both the language uses Latin alphabets. But Afaan Oromo indexing varies in many different ways. As it has been identified by the study Afaan Oromo has its own grammar (which is called Seerluga') from English. Tokenization of Afaan Oromo is almost similar to the English one except apostrophe (') is not punctuation mark in Afaan Oromo, rather it is part of words. The Afaan Oromo stemmer was the only component adopted from previously conducted research. Stemming Afaan Oromo documents has its own unique procedures, totally different from the english one.

In this research, the classification algorithms namely Bayesian and Lazy classifier are used for classifying afaan Oromo text documents. The Bayesian Algorithm includes two techniques namely Bayes Net, and Naïve Bayes techniques. By analyzing the experimental results it is observed that the functions classifiers Bayes networking (97.15%) classification technique has yields better result than other techniques.

VI. REFERENCES

[1] Teferi Degeneh, 2015, The Development of Oromo Writing System, Doctor of Philosophy (PhD) thesis, University of Kent.

[2] Meron Sahlemariam, "Concept-based automatic Amharic Document categorization, "MSc Thesis, 2009

[3] Brussels, "Ontologies - Introduction and Overview", Unpublished MSc Thesis Vrije Universiteit Brussel, 2004

[4] Maron M. and Kuhns J., "Probabilist Indexing and Information Retrieval.," London ACM, pp. PP 22-35, 1760.

[5] Sebastiani, F.: Text Categorization. In Alessandro Zanasi (ed.), Text Mining and its Applications, WIT Press, Southampton, UK, 2005, ppt 109-129.

[6] Kamal, et al "Afaan Oromo News Text Categorization using Decision Tree Classifier and Support Vector Machine: A Machine Learning Approach", published May, 2017

[7] C agri Toraman, ""Text Categorization and Ensemble Pruning in Turkish News Portals"," August, 1811.

[8] A.,Nigam, K., Thrun, S. and Mitchell, T. McCallum, "Text Classification from Labeled and Unlabeled Documents Using EM. Boston: ," Kluwer Academic Publishers, 39(2), pp. pp.103–125, 1800.

[9] Birmingham, Python Text Processing with NLTK 2.0 Cookbook.pdf(August 24, 2010)

[10] Zelalem Sintayehu, Automatic Amharic news 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

[11] [11].Frédéric Flouvat · Fabien De Marchi · Jean-Marc Petit, A new categorization of datasets for frequent itemsets, J Intell Inf Syst (2010) 34:1–19.

[12] [12].Darko Zelenika, Janez Povh, Andrej Dobrovoljc, Document classification, 2012

[13] CLIFTON Phua, Vincent Lee, Kate Smith & Ross Gayler, A Comprehensive Survey of Data Mining-based Fraud Detection Research

[14] Parks B., 1999 "Basic News Writing"," united states. http://www.ohlone.edu/people/bparks/./basicnewswriting.pdf.

[15] Jiri Hynek and Karel jezek, Use of Text Mining Methods in a Digital Library, elpub2002