# Afaan Oromo News Text Categorization using Decision Tree Classifier and Support Vector Machine: A Machine Learning Approach

Kamal Mohammed Jimalo[1] Ramesh Babu P[1] Yaregal Assabie[2]

[1]College of Engineering and Technology, Wollega University, Post Box No: 395, Nekemte, Ethiopia.
[2]School of Information Science, Addis Ababa University, Post Box No: 1176, Addis Ababa, Ethiopia.

**ABSTRACT**

*Afaan Oromo is one of the major African languages that is widely spoken and used in most parts of Ethiopia and some parts of other neighbor countries like Kenya and Somalia. It is used by Oromo people, who are the largest ethnic group in Ethiopia, which amounts to 25.5% of the total population. There are large collections of Afaan Oromo document available in web, in addition to hard copy document in library, and documentation centers. Even though the amount of the document increase, there are challenging tasks to identify the relevant documents related to a specific topic. So, a text categorization mechanism is required for finding, filtering and managing the rapid growth of online information. Text categorization is an important application of machine learning to the field of document information retrieval.*

*The objective of this research is to investigate the application of machine learning techniques to automatic categorization of Afaan Oromo news text. Two machine learning techniques, namely Decision Tree Classifier and Support Vector Machine are used to categorize the Afaan Oromo news texts. Annotated news texts are used to train classifiers with six news categories: sport, business, politics, health, agriculture, and education. To design Afaan Oromo news text categorization system, different techniques, and tools are used for preprocessing, document clustering, and classifier model building. In order to preprocess the Afaan Oromo documents, different text preprocessing techniques such as tokenization, stemming, and stop word removal would be used. 824 news texts were used to do this research. To come up with good results text preparation and preprocessing was done. Stop-word was removed from the collection. The 10 fold cross validation was used for testing purposes.*

*The result of this research indicated that such classifiers are applicable to automatically classify Afaan Oromo news texts. The best result obtained by Decision Tree Classifier and Support Vector Machine is on six categories data (96.58, 84.93%) respectively. This research indicated that Decision Tree Classifier is more applicable to automatic categorization of Afaan Oromo news text.*

**Key Words:** *Afaan Oromo, Text, categorization, Classification and Classifier.*

## 1. INTRODUCTION

It is easy to reach news from various resources like news portals today. In news portals news categorization makes the news articles more accessible.

The globalization era provides a growing amount of information and data coming from different sources. As a result, it becomes more and more difficult for target users to select contents they desire. This has a negative effect for searching the relevant documents from the entire collection. Supporting the target users to access and organize the enormous and widespread amount of information is becoming a primary issue. As a result, several online services have been proposed to find and organize valuable information needed by the target users [**1**]. However, these services are not capable to fully address the users' interest. So, a mechanism is required for finding, filtering and managing the rapid growth of online information. This mechanism is called text categorization [**2**]. Text categorization (TC) can be defined as the task of determining and assigning topical labels to content [**1**]. It is also defined by Berger [**3**] as the task of automatically sorting a set of documents into categories from a predefined set.

The categories can be pre-defined or automatically identified. The text categorization task based on the predefined categories is called text classification. While the text categorization task based on automatically identified categories is called text clustering [**4**]. Manual news categorization (classification) is slow, expensive and inconsistent [**5**]. Therefore automated news text categorization (ANTC) is one of the primary tools of news portal construction.

The aim of news categorization is to assign pre-defined 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 [**6**]. In the news industry metadata is a very important part of a news item. Fast spreading of Internet decreased complexity of news exchange, which resulted in dramatic increase in the number of available news sources and the volume of news items an average recipient received every day. As a paradox, that led to over flooding of consumers with the information and actually decreased its usability [**7**].

On the other hand, speed has always been very important factor in the news industry. Due to the inability to process all the content they receive fast enough, news recipients have to rely on metadata to find out the content they are interested in, which means that it is very important for metadata to be consistent, accurate and comprehensive. It could be even said that news story with inaccurate or insufficient metadata does not exist; because it will rarely reach the consumers no matter how important its content might be [**7**] .

To promote the ease of interchange of news items, The International Press Telecommunication Council (IPTC), an international organization that is primarily focused on developing and publishing Industry Standards for the interchange of news data [**8**], has provided numerous categorization schemes aimed to standardize coding of various aspects of news related metadata. The whole product is called NewsCodesTM [**9**] and at the moment it consists of 28 sets of codes which cover areas like Genre, Confidence, Urgency, Format, Location, Media Type, Scene (for photo and video) etc.

The oldest and the most widely used NewsCodesTM set is called Subject Reference System (SRS) and it provides for coding of subject of the content of a news item. SRS consists of more than 1000 categories divided hierarchically into 3 levels: Subject, Subject Matter and Subject Detail. SRS is constantly updated and maintained by the most important world news providers like Reuters, Associated Press, Agency France Presse and the other members of the IPTC. The codes are language neutral, and descriptions are provided in most important world languages like English, French, German, Japanese, Spanish and many others. English version is the normative one and translations into other languages are provided by the national agencies that are members of the IPTC. The full list of codes is available at the IPTC NewsCodesTM website [**10**].

Until the late 1780s the most popular approach to TC in the operational community was a knowledge engineering (KE) which manually defines a set of rules encoding expert knowledge on how to classify documents under the given categories. However, from the early 1790 a machine learning (ML) approach has become the major research area. Machine learning has considered on a general inductive process that automatically builds an automatic text classifier by learning from predefined set of documents [**11**].

Currently, the categorization task falls at the crossroads of information retrieval (IR) and machine learning (ML). IR researchers believe that it is the user who can say whether a given item of information is relevant to a query issued to a Web search engine, or to a private folder in which documents should be filed according to their contents. Wherever there are predefined classes, documents manually classified by the user are often available. As a result, the predefined data can be used for automatically learning the meaning that the user assigned attributes to the classes. However, reaching levels of classification accuracy that would be impossible if this data were unavailable. In the last few years, more and more ML researchers adopt TC as one of their benchmark applications of choice which are being imported into TC with minimal delay from their original invention [**1**]. In ML approaches; the task that deals with classification is called supervised learning, whereas the task that deals with clustering is called unsupervised learning. In addition, application developers interest mainly due to the enormously increased need to handle larger and larger quantities of documents. This need is emphasized by increased connectivity and availability of document corpus of all types at all levels in the information chain.

Most of the time, the categories are just symbolic labels and their meaning is usually available. However, it is often the case that metadata (such as publication date, document type, and publication source) is unavailable to categorize them. In these cases, categorization must be accomplished only on

the basis of knowledge extracted from the documents themselves [**12**]. For a given application when either external knowledge or metadata is not available, heuristic techniques of any nature may be adopted in order to leverage on these data, either in combination or in isolation from the IR and ML techniques [**13**]. Text categorization can be done manually or automatically. Each of them has advantage and pitfalls to the target users. Automated text categorization is attracting more research these days because it reduces human intervention from manually organizing documents which is too expensive, and error prone [**11**].

Most of automatic text categorization is done by assigning predefined categories to a given text documents. Such concept of text categorization is called text classification [**14**]. Text classification is generally divided in to two main categories. These are flat text classification and hierarchical text classification [**1**]. In flat text classification, categories are treated in isolation of each other and there is no structure defining the relationships among them. The hierarchical text classification addresses this large categorization problem using a divide-and-conquer approach [**15**].

The text classification approach provides a conceptual view of document collections and has important applications in the real world. For example, news stories are organized by subject categories (topics); academic papers are often classified by technical domains; patient reports in health-care organizations are often indexed from multiple aspects such as using taxonomies of disease categories, types of surgical procedures, insurance reimbursement codes and so on. A text categorization approach that uses the unlabeled text collections is called document (text) clustering [**16**].

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. The standard text clustering algorithms can be categorized into partitioning and hierarchical clustering algorithms [**16**]. 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.

### 1.2 Afaan Oromo Language

Afaan Oromo is one of the major African languages that is widely spoken and used in most parts of Ethiopia and some parts of other neighbor countries like Kenya and Somalia [**19**] [**20**]. It is used by Oromo people, who are the largest ethnic group in Ethiopia, which amounts to 25.5% of the total population. Besides first language speakers, a number of members of other ethnicities who are in contact with the Oromo's speak it as a second language, for example, the Omotic speaking Bambassi and the Nilo-Saharan-speaking Kwama in northwestern Oromia [**21**] . Currently, Afaan Oromo is an official language of Oromia regional state (which is the largest Regional State among the current Federal States in Ethiopia). Being the official language, it has been used as medium of instruction for primary and junior secondary schools of the region. Moreover, the language is offered as a subject from grade one throughout the schools of the region. Few literature works, a number of newspapers, magazines, educational resources, official credentials and religious documents are published and available in the language.

Afaan Oromo is Cushitic language which is family of Afro Asiatic languages. It is spoken by more 22 Million peoples and most of native speakers are people living in Ethiopia, Kenya, Somalia and Egypt. It is third largest language in Africa following Kiswahili and Hausa; 4th largest language, if Arabic is counted as Africa language [**26, 27**]. The exact time when the Latin alphabet started being used for Afaan Oromo writing was not well known, but on November 3, 1791 it adopted as official alphabet of Afaan Oromo on. Now it is language of public media, education, social issues, religion, political affairs, and technology.

In general, Afaan Oromo is widely used as written and spoken language in Ethiopia and neighboring courtiers like Kenya and Somalia. With regard to the writing system, "**Qubee**" (a Latin-based alphabet) has been adopted and become the official script of Afaan Oromo since 1791 [**19**] .

### 1.3 Afaan Oromo Alphabets and Writing System

According to Taha [**25**], Afaan Oromo is a phonetic language, which means that it is spoken in the way it is written. The writing system of the language is straightforward which is designed based on the Latin script. Unlike English or other Latin based languages there are no skipped or unpronounced sounds/alphabets in the language. Every alphabet is to be pronounced in a

clear short/quick or long /stretched sounds. In a word where consonant is doubled the sounds are more emphasized. Besides, in a word where the vowels are doubled the sounds are stretched or elongated.

Like in English, Afaan Oromo has vowels and consonants. Afaan Oromo vowels are represented by the five basic letters such as a, e, i, o, u. Besides, it has the typical Eastern Cushitic set of five short and five long vowels by doubling the five vowel letters: 'aa', 'ee', 'ii', 'oo', 'uu' [**19**]. Consonants, on the other hand , do not differ greatly from English, but there are few special combinations such as "**ch**" and "**sh**" (same sound as English),"**dh**" in Afaan Oromo is like an English "*d*" produced with the tongue curled back slightly and with the air drawn in so that a glottal stop is heard before the following vowel begins. Another Afaan Oromo consonant is "**ph**" made when with a smack of the lips toward the outside "**ny**" closely resembles the English sound of "gn". We commonly use these few special combination letters to form words. For instance, **ch** used in **barbaachisaa** '*important'*, **sh** used in **shamarree '***girl'*, **dh** use in **dhadhaa** '*butter'* , **ph** used in **buuphaa '***egg',* and **ny** used in **nyaata** '*food'* . In general, Afaan Oromo has 36 letters (26 consonants and 10 vowels) called "**Qubee".** All letters in English language are also in Afaan Oromo except the way it is written. Table 2 shows Afaan Oromo alphabet.

**1.4 Afaan Oromo Consonants**

| | | Bilabial/ Labiodentals | Alveolar/ Retroflex | Palato-alveolar/ Palatal | Velar/Glottal | | | |
|---|---|---|---|---|---|---|---|---|
| Stops | Voiceless | (p) | t | K | | | | |
| | Voiced | b | d | G | | | | |
| | Ejective | ph | x | Q | | | | |
| | Implosive | dh | | | | | | |
| Affricates | Voiceless | ch | | | | | | |
| | Voiced | j | | | | | | |
| | Ejective | c | | | | | | |
| Fricatives | Voiceless | f | s | Sh | h | | | |
| | Voiced | (v) | - | Nasals | | m | n | ny |
| Approximants | | w | 1 | Y | | | | |
| Flap/Trill | | R | | | | | | |

**Table 1: Afaan Oromo vowels**

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

**Table 2: Afaan Oromo Alphabet**

Afaan Oromo punctuation mark is placed in text to make meaning clear and reading easier. Analysis of Afaan Oromo texts reveals that different punctuation marks follow the same punctuation pattern used in English and other languages that follow Latin Writing System. Similar to English, the following are some of the most commonly used punctuation marks in Afaan Oromo:

## 2. Materials & Methods

Text categorization is the task of automatically assigning input text to a set of categories. Text categorization can be divided in to text classification and text clustering based on the category it uses [**11**]. Text clustering is the automatic identification of a set of categories and assigns set documents under the automatically identified categories.

### 2.1 Text Categorization Techniques

Text categorization techniques can be supervised, unsupervised and semi supervised learning [**21**]. Supervised learning is the search for algorithms that reasons from externally supplied class to produce general

hypotheses, which then make predictions about future instances. In other words, the goal of supervised learning is to build a concise model of the distribution of class labels in terms of predictor features [**21**].

The unsupervised learning does not require externally supplied knowledge for categorizing instances. The main aim of this learning algorithm is to generate a group of features that have similar properties. So the grouping is done without any supervision of human beings. Text clustering is an example of unsupervised learning.

In semi supervised learning, parts of the documents are required external knowledge and other does not require external knowledge in the categorization process.

### 2.2Text classification

Text classification is an example of supervised learning where a given document is assigned to predefined categories based on the similarities of the labeled documents in the training set [**11**]. Text classification can be done manually or automatically. Traditionally, text classification has been performed manually [**13**]. The manual text classification uses expert (human being) to categorize the document in to predefined categories. However, as the number of documents explosively increases, the task becomes no longer amenable to the manual categorization, and it requires a vast amount of time and cost. This has lead to numerous researches for automatic document classification. The automatic text categorization is generally divided in to two main categories [**27**] . These are flat text categorization and hierarchical text categorization.

### 2.3 Text Clustering

It is easy to collect unlabeled documents for text categorization purposes. As a result, a text categorization mechanism is required for categorizing the unlabeled documents. This mechanism is called text clustering [**29**]. Text clustering is an unsupervised learning which does not require pre-defined categories and labeled documents [**30**]. The main aim of text clustering is to determine the intrinsic grouping in a set of unlabeled data. The intrinsic groups have high intra-group similarities and low intergroup similarities. Text clustering algorithms are categorized into two main groups such as hierarchical clustering and partitioning clustering algorithms. Partitioning clustering algorithms create a cluster by splitting the data into k partition where each partition represents a cluster.

### 2.4 Hybrid Approaches for Text Categorization

In the real world, there is available unlabeled data but labeling them is expensive because it requires human expert. As a result, there is a limited number of labeled data [**21**]. Using the unlabeled data for text categorization is not preferred due to its time complexity and interpretation problems. So a mechanism is required to combine clustering and classification algorithms for text categorization process [**23**].

First, clustering is used with text classification prior to a classifier to reduce feature dimensionality by grouping similar features into a much smaller number of feature clusters. These clusters are used to the original feature space [**24**]. Second, clustering is used with text classification as feature clustering and document clustering. In this way a reduction for both dimensions is attained. Feature clustering generates coarser pseudo features, which reduce noise and sparseness that might be exhibited in the original feature space. In the second stage, documents are clustered as distributions over the "distilled" pseudo features, and therefore generate more accurate document clusters [**21**]. Third, clustering is used in semi-supervised classification as a method to extract information from the unlabelled data in order to boost the classification task. Particularly clustering is used to create a training set from the unlabelled data and to augment the training set with new documents from the unlabelled data [**16**].

### 2.5 Text Categorization Phases

Text documents represented in a natural language are not easily used by the classifier for building algorithms. In order to solve this problem mapping a text document into a schematic representation of its content is required. As a result, the categorization algorithm transforms each document into a vector of weights corresponding to an automatically chosen set of keywords. This transformation has two main steps [**27**].

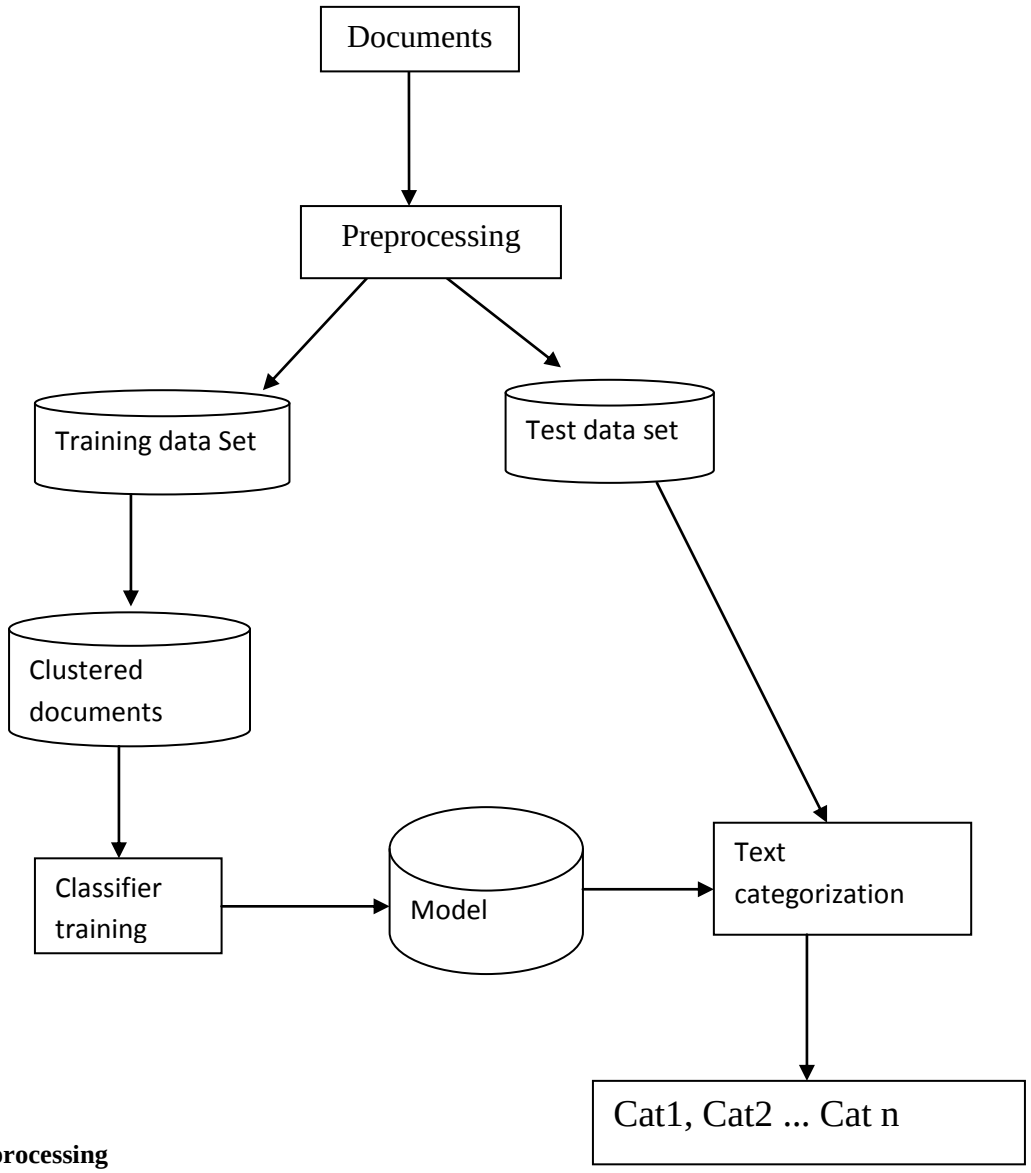
First, suitable representation of the document has to be chosen. This representation is used for all documents to be indexed and it has all necessary words that can characterize the documents. The information retrieval and machine learning researchers have different views in representing strings [**11**].

Second, it assigns weights to each representation term which shows the frequency of occurrence of the term in the indexed document. Even though the document indexing is performed, the results obtained has high dimension and take a large amount of storage space.

## 3. Proposed System Architecture

In this study, the proposed text categorization system is developed in three stages as shown in figure1. These are preprocessing, clustering, classifier training, and testing the text categorization system. The preprocessing makes the raw data ready for the experiment. In this stage the irrelevant terms are removed from the documents and words of the same context with different forms are converted into the same word. The final goal of this stage is to convert the collection of text in to matrix of index terms with their tf×idf weight values.

In the experimentation, cluster of documents are created for training data set. The resulting clustering solution is used to train a text classification model. The developed model is tested using test data set. Finally, the system comes up with categories of Afaan Oromo news text documents.




### 3.1 Text Preprocessing

Text preprocessing is crucial step for the subsequent text clustering and classification tasks. During text preprocessing, there are a sequence of steps applied to generate content-bearing terms and assigns weights that shows their importance for representing the document they identified from.

First, tokenization is performed which attempts to identify words in the document corpus. The common method of representing the document text is using the bag of words approaches where the word from the document corresponds to a feature and the documents are considered as a feature vector. This indicates that the words can only discriminate the given documents in

the categorization process. So, the punctuation marks and digits are irrelevant component of the text documents. In the present study, the punctuation marks and digits are removed and replaced with space. Tokenization is the process of chopping character streams in to tokens, while linguistic preprocessing then deals with building equivalence classes of tokens which are the set of terms that are indexed. Tokenization in this work also used for splitting document in to tokens and detaching certain characters such as punctuation marks.

```
For file in corpus
        Define word delimiter to space
        Read files
                For file in read
                If there is word delimiter
                        Put each terms as separate token
```

**Algorithm 3.1:** Tokenization

Normalization involves process of handling different writing system. Primarily every term in the document should be converted in to similar case format in this study lower case. For instance 'INFORMATION', 'Information', 'information' is all normalized to understandable as lower case 'information' the system.

| Number | Stop word | Meaning |
|---|---|---|
| 1 | kana | This |
| 2 | Sun | That |
| 3 | Inni | He |
| 4 | Ani | Me? |
| 5 | Isaan | They |
| 6 | Ishee | She |
| 7 | Akka | Like |
| 8 | Ana | Me |
| 9 | Fi | And |

**Table 3.1:** Sample Stop Word lists of Afaan Oromo documents

In the present study, stop words are identified manually by consulting books, dictionaries and different research articles of the Afaan Oromo languages. The consulted books and dictionaries help to identify preposition, conjunction, articles, pronoun and auxiliary verbs of the Afaan Oromo language. After identifying the stop words in the Afaan Oromo documents, algorithm 3.1.2 remove the stop words from the document corpus.

```
: Stop word removal
Read stop word list file
Open the file for processing
Do
   Read the content of the file line by line
    Assign the content to string
      For word in string split by space
         If word in stop word list
          Remove word from the index term
        Else
           Continue
       End if
   End for
While end file
```

**Algorithm 3.1.1: Stop word removals**

**Stemming**: stemming is process used in most search engines and information retrieval systems. It is core natural language processing technique for efficient and effective IR system. Generally stemming transforms inflated words in to their most basic form. There are different stemming algorithms but the most common one is that of Porter, called 'Porter Stemmer'. Even if stemming is very similar to lemmatization in most of indexing process stemming is used. Stemming is language dependent process in similar way to other natural language processing techniques. It is often removing inflectional and derivational morphology. E.g. automate, automatic, automation = *automat*.

*1. READ the next word to be stemmed*
2. OPEN stop word file
Read a word from the file until match occurs or End of File reached
IF word exists in the stop word list
Go to 5
3. If word matches with one of the rules
Remove the suffix and do the necessary adjustments
Go back to 3
ELSE
Go to 6
4. Return the word and RECORD it in stem dictionary
5. IF end of file not reached
Go to 1
ELSE
Stop processing
6. IF there is no applicable condition and action exist
Remove vowel and return the result

**Algorithm 3.1.2: Stemming Algorithm**

### 3.2. Index Construction

Indexing involves tokenizing, normalization, stop word removal and stemming. The code 3.2.1 is fragment code that tokenizes and normalizes the terms in the document.

```
Characters = ".,!#$%^&*();:\n\t\\\"?!{}[]<>0122556789"
def tokenize (document):
    terms = document.lower ().split()
    return [term.strip (characters) for term in terms]
```

**Code 3.2: Tokenization and Normalization**

Here this fragment code splits document based on space between each words, then convert every words in document in to lower case if it is not in lower case primarily. The documents in lower case are checked as if it is not punctuation mark or number. Finally normalized and tokenized document will be returned for next process. The index terms selected in this study are content bearing terms which are not part of stop list. Primarily there are identified list of stop words which are not content bearing, just used for grammatical purpose only. The following code is used for the stop word removal from the documents.

```
s=open ('stopword.txt','r')
os.chdir ('corpus')
Stop list=s.read ()
s.close ()
for i in token:
        if i not in stoplist: # Stop list removal
        Stemmed=thestemmer (i)
```

**Code 3.2.1 : Stop word removal**

### 3.3 Classification

The clustered documents are used to classify Afaan Oromo news text documents using the support vector machine (SVM), Naive Bayes Classifier, Bayes Network Classifier and decision tree classifiers. As a result, the support vector machine and decision tree classification scheme constructs a model from the cluster categories of the training collection.

#### 3.3.1 Decision Tree Classifier

A decision tree classifier predicts the target value of the class based on various attributes of the data set. The decision tree classifier has internal and terminal (leaf) nodes. The internal node indicates the different attributes of the text classification, and the leaf nodes show the classification of the attributes. The researcher uses the Decision Tree decision tree classifier for classifying the Afaan Oromo news text documents data set. The decision tree classifier classifies a given data set using algorithm shown below.

Algorithm 3.4: Decision tree classifier algorithm

1) Create a decision tree based on the attribute values of the available training data.

2) Identify the attribute that discriminates the training data sets.

3) Based on the identified attributes, classify the given data sets in to different classes (categories).

4) If there is any value for which the data instances are falling within its category that has the same value for the given class, then it terminates that branch and assigns to it the class until all the data instances are assigned to their class.

**Algorithm:** Decision tree classifier algorithm

#### 3.3.2 Support Vector Machine

The core idea behind the SVM classifier is to fit the linear model to the mapped training data by maximizing the margin of the classifier [**28**]. This shows that the value of the given parameter of hyper plane to the nearest training patterns from given classes is maximized as many training patterns as possible. In this study, the researcher classifies the Afaan Oromo documents using sequential minimal optimization supports Support Vector machine (SVM) classifiers. It has advantage over the other support vector machine classifiers [**14**]. First, the amount of memory required for SVM is linear in SVM breaks the large quadratic programming problem into a series of smallest possible quadratic programming problems. The training set size which allows SVM to handle very large training sets. Second, it avoids matrix computation and this makes SVM to scale somewhere between linear and quadratic in the training set size for various test problems.

Finally, it is the fastest classifier for linear SVM and sparse data sets. Generally, the SVM algorithm involves three important components. These are an analytic solution to the optimization problem for the two chosen multipliers; a heuristic for choosing the two multipliers to optimize at a given step; and a step to compute the offset.

## 4. RESULTS

### 4.1 Classification using decision tree classifier

Weka (A Data mining Tool) supports different decision tree classifiers that are used for classifying numeric and nominal attributes. In the present study, the DECISION TREE decision tree classifier is used for experimentation. This classifier requires a small amount of memory and time. The Decision Tree classifies the 96.58 %of 584 instances correctly within 0.02 seconds as illustrated in table 4.1.1 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 below for the Decision Tree Classifier.

| Sport | Business | Politics | Health | Agriculture | Education | |
|---|---|---|---|---|---|---|
| 98 | 3 | 0 | 0 | 3 | 0 | Sport |
| 3 | 99 | O | 2 | 0 | 0 | Business |
| 1 | 0 | 103 | 0 | 0 | 0 | Politics |
| 0 | 2 | 1 | 101 | 0 | 0 | Health |
| 2 | 0 | 0 | 0 | 101 | 1 | Agriculture |
| 0 | 2 | 0 | 0 | 0 | 62 | Education |

**Table 4.1.1: confusion matrix of the Decision Tree classifier**

Based on the above confusion matrix, the performance of the classifier is shown in table 4.1.2 using precision, recall, F-measure, and ROC-Area.

| Class | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area |
|---|---|---|---|---|---|---|
| Sport | 0.927 | 0.013 | 0.927 | 0.927 | 0.927 | 0.986 |
| Business | 0.952 | 0.015 | 0.925 | 0.952 | 0.943 | 0.986 |
| Politics | 0.99 | 0.002 | 0.99 | 0.99 | 0.99 | 0.994 |
| Health | 0.971 | 0.004 | 0.981 | 0.971 | 0.976 | 0.981 |
| Agriculture | 0.971 | 0.006 | 0.971 | 0.971 | 0.971 | 0.979 |
| Education | 0.969 | 0.002 | 0.984 | 0.969 | 0.976 | 0.983 |
| Weighted Avg. | 0.966 | 0.007 | 0.966 | 0.966 | 0.966 | 0.985 |

**Table 4.1.2: Detail accuracy of Decision Tree classifier by class**

The performance of each class is computed from the confusion matrix as shown in table 4.3 by identifying the correctly classified instances diagonally against the actual number of instances in a category row-wise. As we understand from the confusion matrix in table 4.1.1, "politics" category registered the best accuracy of 99%, followed by "health" and "agriculture" with 97.12% and "education" 96.88% accuracy. On the other hand, "sport" category has least performance because this category has low internal similarity. The performance of the classifier is also measured using F-measure value as shown in table 4.1.2. Based on the F-measure value, the DECISION TREE classifier in table 4.1.2 shows that the category "sport" is classified less accurate than the other categories but the "politics", "health", and "agriculture" are classified more accurately than other categories while the "sport" category has low performance which is 94.2% F-measure. This indicates that the instances cluster in to the corresponding cluster of the "sport" category have fewer words in common than others. Due to this, the "sport" category is classified less accurate than other categories.

4.2 Classification using Support Vector Machine

The Weka version 3.6.4 used for the experiment has different SVM (Support vector machine) classifiers. In the present study, SVM classifier has higher performance than other support vector machine classifiers because it correctly classifies 496 (84.93 %) out of 584 instances as shown in table 4.2.1 below

| Sport | Business | Politics | Health | Agriculture | Education | |
|---|---|---|---|---|---|---|
| 94 | 6 | 0 | 0 | 4 | 0 | Sport |
| 2 | 99 | O | 2 | 1 | 0 | Business |
| 2 | 0 | 102 | 0 | 0 | 0 | Politics |
| 1 | 2 | 1 | 100 | 0 | 0 | Health |
| 1 | 0 | 0 | 0 | 96 | 7 | Agriculture |
| 59 | 0 | 0 | 0 | 0 | 5 | Education |

**Table 4.2.1: Confusion Matrix of SVM classifier**

Based on the above confusion matrix, the performance of the classifier is shown in table 4.2.2 using precision, recall, F-measure and ROC-Area.

| Class | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area |
|---|---|---|---|---|---|---|
| Sport | 0.904 | 0.135 | 0.591 | 0.904 | 0.715 | 0.909 |
| Business | 0.952 | 0.017 | 0.925 | 0.952 | 0.938 | 0.973 |
| Politics | 0.981 | 0.002 | 0.99 | 0.981 | 0.986 | 0.985 |
| Health | 0.962 | 0.004 | 0.98 | 0.962 | 0.971 | 0.984 |
| Agriculture | 0.923 | 0.01 | 0.95 | 0.923 | 0.937 | 0.966 |
| Education | 0.078 | 0.013 | 0.267 | 0.078 | 0.124 | 0.935 |
| Weighted Avg. | 0.849 | 0.024 | 0.836 | 0.849 | 0.824 | 0.96 |

**Table 4.2.2: Detail accuracy of SVM classifier by class**

The performance of each class is computed from the confusion matrix as shown in table 4.5 by identifying the correctly classified instances diagonally against the actual number of instances in a category row-wise. As we understand from the confusion matrix in table 4.5, "politics" category registered the best accuracy of 98.1%, followed by health and business with 96.15% and 95.17% accuracy. On the other hand, "education" category has least performance because it has low internal similarity.

The performance of the classifier is also measured using F-measure value as shown in table 4.6. Based on the F-measure value, the SVM classifier in table 4.6 shows that the category "education" is classified less accurate than the other categories but the "politics", "health", and "business" are classified more accurately than other categories while the "education" category has low performance which is 14.2% F-measure. This indicates that the instances cluster in to the corresponding cluster of the "education" category have fewer words in common than others. Due to this, the "education" category is classified less accurate than other categories.

## 5. Conclusion

The explosion of the World Wide Web provides a growing amount of information and data coming from different sources. Therefore, a text categorization mechanism is required for finding, filtering and managing the rapid growth of online information.

This research has presented an automatic news items categorization for Afaan Oromo news text documents using machine learning techniques: Decision tree Classifier and Support Vector Machine. This research also compares two known classification techniques using Afaan Oromo news text documents which lie into six classes. The comparison is based on two main aspects for the selected classifiers, accuracy and time. In terms of accuracy, results show that the Decision Tree classifier achieves the highest accuracy.

Researches which have been done in the area of machine learning in text categorization indicate good results. Our research showed promising results. The best result obtained by Decision Tree Classifier and Support Vector Machine is on six categories data (96.58, 84.93%) respectively. This research indicated that Decision Tree Classifier is more applicable to Afaan Oromo news text than the other classifiers. Moreover, it is learnt that considering categories with equal number of news items increases the performance of the classifiers. In other words, insufficient examples in one class can affect the classifier as a whole. It was also observed that the classification of Afaan Oromo news text is possible without using the sophisticated feature reduction techniques such as information gain and odds ratio.

**6. References**


[1] Addis A., *Study and Development of Novel Techniques for Hierarchical Text Categorization.* Italy: University of Cagliari, 1810.

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

[3] Berger H., "A Comparison of Tex Categorization Methods Applied to N-Gram Frequency Statistics.proceedings of the 17th Australian Joint conference on rtificial Intelligence cairns Australia," *:Springer* , pp. PP 4-10, 1804.

[4] Barker D. and Kachites A., "Distributional clustering of Words for Text Classification.," *ACM SIGIR*, pp. PP 96-102, 1798.

[5] L.E.Knecht and M.J. Cellio P.J.Hayes, ""A New Story Categorization System."In proceedings of the second Conference on Applied Natural Language Processing ," *ANLC Strouds Burg,PA ,USA*, pp. PP9-17, 1788.

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

[7] Pandzic I.S,Gulija D: Bacan H., ""Automated News Item Categorization "," *Faculty of Electirical Engineering and Computing University of Egreb*.

[8] (1814, April) [Online]. http://www.iptic.org

[9] (1814, January 25) [Online]. http://wwww.iptc.org/NewsCodes/nc ts.table01php?`

[10] [Online]. http://www.iptc.org/NewsCodes/nc ts.table01PHP

[11] Sebstiani F., "Machine Learning in Automated Text Categorization.ACM computing Surveys.Consiglio Nazionale delle Ricerche, Italy," *ACM*, pp. PP 10-15.

[12] Sebastiane F., "A Tutorial on Automated Text Categorization.Consiglio Nazionale delle Ricerche,Italy," *:Istituto di Elaboratione dell'Informazion*, 1800.

[13] F. Sebastiani, "Text Categorization in Text Mining and Its Applications to Intelligence, CRM and Knowledge Management. ," *South Hampton, UK: WIT Press.*, 1805.

[14] C. John, "Sequential Minimal Optimization: A Fast Algorithm for TrainingSupport Vector Machines. USA:," *Morgan Kaufmann*, 1798.

[15] D., Biro G. and Yang, J. Tikk, "A Hierarchical Text Categorization Approach and Its Application to FRT Expansion. ," *Hungary: Elsevier.*, 1801.

[16] 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.

[17] Abera N., ""Long vowels in Afaan Oromo: A generic approach"," *, Master's thesis , School of graduate studies, Addis Ababa University, Ethiopia.*, 1788.

[18] Grage G. & Kumsa T., ""Oromo dictionary", ," *African studies center. Michigan state University*, 1782.

[19] Tilahun G, "" Qubee Afaan Oromo : Reasons for choosing the Latin script for developing an Afaan Oromo Alphabet".," *Journal of Oromo studies*, 1793.

[22] C. Dawson, "" Practical Research Methods." New Delhi:," *UBS Publishers*, 1802.

[20] I., Zeitouni, K., Gardarin, G., Nakache, D. and Metais, E. Popa, "Text Categorization for Multi-Label Documents and Many Categories. ," in *Washington DC, USA: IEEE.*, 1807.

[21] A. Ozgur, "Supervised and Unsupervised Machine Learning Techniques for Text Document Categorization. MSc Thesis. Bogazin University, Turkey.," (1804).

[22] I. Dhillon, "A Divisive Information Theoretic Feature Clustering Algorithm for Text Classification. ," *Journal of Machine Learning Research, 3(27),* , pp. pp.1265-1287., 1803.

[23] N. Slonim, "The Power of Word Clustering for Text Classification. European Colloquium on IR Research:," *ECIR*, pp. pp.22-45, (1801).

[24] Y. Zhao, "Comparison of Agglomerative and Partitioning Document Clustering Algorithms.," *Washington DC: ACM Press.*, (1802).

[25] Show Language,online edition, Etnologue. (1809.) [Online]. Available: http://www.ethnologue.com/web.asp. [Accessed: 21-january-1814].

[26] Parks B., "BASIC NEWS WRITING"," *united states.* , Available at http://www.ohlone.edu/people/bparks/./basicnewswriting.pdf accessed on February 18, 1814 1809.

[27] Duwairi R., ""Arabic Text Categorization" ," *The International Arab Journal of Information Technology , , Jordan University of Science and Technology, Jordan*, vol. Vo.4,No.2, April 1807.`

[28] G., Steinbach, M. and Kumar, V. Karypis, "A Comparison of Document Clustering Techniques. New York, USA:," *ACM Press/Addison-Wesley Publishing Co.*, 1804.