Multi-Label Text Document Classification using Word Similarity for Amharic

Haimanot Mogesse
HiLCoE, Computer Science Programme, Ethiopia
hmogesse@yahoo.com

Wondwossen Mulugeta
HiLCoE, Ethiopia
School of Information Science, Addis Ababa University, Ethiopia
wondisho@yahoo.com

Abstract

Multi-label classification methods are increasingly required by modern applications. The traditional single label classification methods assume that each instance is associated with one conceptual label from a category set, and classes are mutually exclusive by definition. In real-life scenarios, however, some problems like news and documents should be annotated with multiple labels to reflect the varying ideas they contain. This work addresses the issue of multi-labeling Amharic news using classifier models designed and developed for multi-labeling of text documents, using machine learning approach. It uses two datasets, one with 11 and the other with 10 categories, illuminating one of them. It also addresses word similarity aspect of selected determinant terms. The tool used for this research is the open source MEKA package, a multi-label extension to WEKA and the classifiers used are Binary Classifier (BR) and its variant Classifier Chain (CC) which considers the issue of label correlation. As the tool doesn't directly upload Unicode data, transformation of Unicode to ASCII is done using a self developed Python program. Supervised technique is implemented using cross validation vs single split (66% to 34%) train/test activity. The result shows that CC classifier performs better in Label Accuracy, also known as H_Accuracy, using cross validation technique than single split, achieving 86.7% and 76.8%, respectively. This measurement indicates accuracy for each label calculated and averaged across all labels. BR, on the other hand, achieved 28.5% accuracy using single train/test method, yet performs better than CC on average precision, recall and rank loss.

Keywords: Multi-label; Machine Learning; Label Correlation; Amharic Text Classification

1. Introduction

Categories signify organization of items into groups according to their similarities or shared characteristics and classifying documents into these categories, which are meaningful to users, is one of the most successful approaches adopted to organize information [1].

During the past decade, applying machine learning approach to the problem of multi-labeling is attracting interest from the research community [2]. It has been widely applied to diverse problems like automatic annotation for multimedia contents, bioinformatics, web mining, rule mining, information retrieval, tag recommendation, etc. Traditional classification is concerned with learning from a set of instances that are associated with a single label from a set of disjoint labels L, |L| > 1. In multi-label, instances are associated with a subset of L [3].

1.1 Definition of Text Classification

Some authors use text classification interchangeably with text categorization. [6], and define it as a task of classifying documents into predefined classes. Text categorization is described in [5] as text classification or topic spotting, which is the task of automatically sorting a set of documents into categories from a predefined set. Classification, in contrast to clustering, is defined as the task of assigning instances to pre-defined classes while the later is grouping related objects together or automatic identification of a set of categories and assigns to the identified categories, without labeling them [6].

Thus, classification is a process of finding a model (a function) that describes and distinguishes data classes or concepts. It is the function that determines the closeness of a text document to the already defined concept classes. The goal of classification is, hence, to use the model to predict the class of new objects whose class label or category is unknown [7].

1.2 The Amharic Language in Brief

Amharic is one of several languages in the Semitic group descended from the earlier language Ge`ez. Semitic is one of the six Afroasiatic Super-family, while Amharic is the second-most spoken Semitic language in the world, after Arabic [8]. It is the official working language of the Federal Democratic Republic of Ethiopia, used nationwide. Amharic is written using the Ethiopic script called Fidel (ፊደል). Amharic Fidel grew out of the Ge'ez script, *abugida*. Each character represents a consonant+vowel sequence, but the basic shape of each character is determined by the consonant, which is modified for a vowel. Amharic is written left-to-right unlike other Semitic languages spoken outside of Ethiopia [9].

1.3 Word Similarity Consideration

Authors prefer to use different words to express the same concept. In a word level classification, similar concepts expressed using different terms will be missed out [10]. Thus, considering the similarity of words will have a critical role in the learning process of a classifier.

For this research, more than 300 words are identified as determinant words for the labels. A shallow stemming is also performed on the words using a program written with Python. This program also addresses the concern of similarity in meaning of words though it is done at very shallow level, and normalization of words that use different letters but have same sound.

2. Related Work

There are significant numbers of works done worldwide on text document classification for different languages using machine learning approach. Read [11] worked towards multi-labeling using a total of 12 multi-label datasets, with dimensions ranging from 6 to 983 labels, and from less than 1,000 examples to almost 100,000. Four algorithms are used for each of the BM-based and Ensemble algorithms and evaluated all algorithms under a WEKA-based framework. Regarding results for predictive performance, accuracy ranges from 31.9%, for Reuters data using Binary Method (BM) classifier, up to 75.1% for Medical data using Chaining Method (CM) are obtained. On the Ensemble algorithm, the research achieved 36% by Ensembles of Binary Method (EBM) on Reuters data while observed 77.6% by Ensembles of Classifier Chains (ECC) on Medical data. The author summarized the finding as CM-based methods and other methods that intensively model label correlations obviously have a place in multi-label classification, especially for datasets of relatively small dimensions.

Read [12] added significant value on a previous work mentioned above using 16 datasets of which 6 are large datasets. The attribute spaces are binary and numeric attribute spaces. On the CC against BR experiment, CC produces similar and better predictive performance in all Accuracy, Exact Match and F1-MACRO measurements on the different datasets.

Tsoumakas and Katakis [15] performed comparative experiments of multi-label classification methods. The authors introduced the concepts of label cardinality which is the average number of labels of the examples in a document and label density which is the average number of labels of the examples in a document divided by the number of labels. These measures will help as the possible number of labels has greater impact on multi-label classification methods. The research also shows the performance of the methods differs according to the dataset used and the learning algorithms applied.

Text document classification for Amharic has been done by different authors using different classification techniques most of them on news articles of different sources [14, 15, 16, 17]. However, they address the problem of classifying documents to the appropriate

single label. Surafel Teklu [14] achieved 95.8% accuracy using Naïve Bayes algorithm for 3 categories and the least 64.5% for 16 categories using KNN classifier. Yohaness Afework [15] did a bit further pre-processing and achieved 79.72% and 81.15% accuracy, using Logic Model Tree (LMT) and the Library of SVM (LibSVM) classifiers, respectively. Concept based classification of Amharic document was done by Meron Sahlemariam [18].

In general, previous studies on Amharic addressed a number of issues such as, classification, text categorizer, automatic indexing, etc. However, none of these works use machine learning approach for multi-label classification of text documents and evaluate the impact of considering similarity of words towards the performance of a classifier.

3. The Proposed Solution

3.1 Data Sets

The Amharic version of Reporter newspaper is used as a corpus for this research. The news articles are documented by the newspaper representing a single category. Two datasets are used: using 800 examples for 11 labels and using 741 examples excluding one of the labels. One of the labels, **የታክሲ ገጠመኝ** - which means 'Taxi Discussions', is excluded because of its use of determinant words of other category for informal way of expression, or being exclamation clause for this category.

The site (www.ethiopianreporter.com) has seven main news item sections, namely '**ፖለቲካ** - Politics', '**ቢዝነስና ኢኮኖሚ** - Business & Economy', '**ቆይታ** - Interview', '**ኪንና ባህል** - Art & Culture', '**ክቡር ሚኒስትር** - Talk to the Minister', '**ስፖርት** - Sport' and '**ሌሎች ዓምዶች** - Other Columns'. The last one has three sub categories under it namely, '**ህግ** - Legal issues', '**ማህበራዊ** - Social' and '**ልዩ ልዩ** - Others'; and these three have four, eight and nine subsections, respectively.

The columns used for this research are selected by their suitable presentation and probability of interdependence with each other, so that multi-labeling of a document that is tagged with single label would be possible. Manual labeling is done by two persons and one mediator, after developing a template that defines and describes what each category is and is not. Topic relevance is the main factor of labeling an article to a particular label.

Table 1 shows how many can be multi-libeled and how many are left being single label from the available single label articles. For example, from 101 politics articles, 57 of them can be multi-labelled.

Table 1: Category and Labeling Statistics of Corpus

Dataset1				
Item	*Categories of Articles*	*No. of Articles*	*Single Label*	*Multi Label*
1.	Politics	101	44	57
2.	Alem	56	0	56
3.	Business	102	49	53
4.	Sport	103	14	89
5.	Mahiberawi	44	21	23
6.	Shemach	47	28	19
7.	Tena	99	46	53
8.	Technology	17	17	0
9.	Diaspora	75	0	75
10.	Yetaxi Getemegn	59	59	0
11.	Art	98	42	56

3.2 Method Used

Corpus-based approach, also called statistical approach, is one of the major methods of classification which uses automated learning techniques over corpora of natural language examples in an attempt to automatically induce suitable language processing models [6].

Supervised approach is applied in the classification task of this work. Accordingly, training and testing have been done which required some examples from the corpus for training purpose so as to conduct prediction on the previously unseen examples. From the known statistical methods used to automatically build classification models, we have used Support Vector Machine (SVM) classifier.

Machine learning provides algorithms that automatically discover patterns in data and make intelligent decisions. Since text may be modeled as quantitative data with frequencies on the word attributes, it is possible to use most of the methods for quantitative data directly on text. However, text is a particular kind of data in which the word attributes are sparse, and high dimensional, with low frequencies on most of the words [19].

3.3 The Classification Model

The open source MEKA package, a multi-label extension to WEKA, is used as a tool for the automatic classification modeling. MEKA is based upon the WEKA framework, providing support for development, running and evaluation of multi-label and multi-target classifiers. MEKA comes bundled with WEKA's weka.jar, and also Mulan's mulan.jar for running classifiers from this framework. MEKA allows datasets to be stored as .ARFF file.

In single label, an instance belongs to → A certain 'Category X' or 'No Category'

Its Attribute Relation Format representation is the following where @ Attribute [1] and [2] depicted below indicate the data type expected in the @ data presentation part.

Relation @ Attribute Text [1] string @ Attribute Class [2] {Sport, Business, Politics, … }@data 'a certain text1, say ['**የኢትዮጵያ እግር ኳስ ፌዴሬሽን የፍጻሜ ጨዋታ መኖሩን ገለጸ**' which has related meaning as 'Ethiopian football federation indicates there is final match, ...], sport, 'a certain text2', Business, 'a certain text3', Politics, …

Accordingly, Attribute [1] indicates that the expected data is in string format, and Attribute [2] is predefined classes, from which label can be tagged to the text.

In multi-label classification, however, an instance belongs to → categories 'X' or 'XY' or 'XYZ, …'. Hence, if we follow the same representation as single label, the possible outcomes will reach up to 2^n -1 where n is the number of individual labels. Its ARFF format representation will look like:

Relation @ Attribute Text string [1] *@ Attribute* Class [2] {Sport, Business, Politics, sport-politics, sport-business, business-politics, sport-politics-business, …} @ data 'a certain text1, e.g., '**የኢትዮጵያ እግር ኳስ ፌዴሬሽን የፍጻሜ ጨዋታ መኖሩን ገለጸ**', sport, 'a certain text2, e.g., **መንግሥት የአበባ ኤክስፖርተሮች በኢትዮጵያ አለም አቀፍ ግንኙነት ላይ ላበረከቱት አስተዋጽኦ ሽልማት አበረከተ**' which has related meaning as 'The government has awarded flower exporters recognizing their contribution to Ethio-Global relationship', Business-politics, etc.

In this case, the complexity will increase especially when the number of labels keeps growing. The binary relevance classifier of MEKA uses the following binary representation:

Relation @Attribute Text string *@Attribute* Sport {0,1} *@Attribute* Business {0,1} *@ Attribute* Politics {0,1}, … *@data* 'a certain text', 1,0,0, …, 'a certain text2', 0,1,1, …

The binary representation indicates whether the text belongs to a particular category, i.e., 1, or not, i.e., 0. This representation enhances identification of labels which commonly appear together.

3.4 Modeling Multi-Label using Machine Learning Method

A particular area of machine learning, called inductive learning, consists of techniques that induce models by using a set of previously known instances or examples. As conducting preprocessing will boost up the modeling capability, some preprocessing was done and the procedure followed is depicted in Figure 1 below.

The experimental result of selected classifiers is also analyzed in two scenarios: one without considering word similarity and the shallow stemming and the other that considers both processes. Flat file in ARFF format is used to load the texts with respective categories. As MEKA works only with Latin representation, Python programming and its inbuilt functions Unicode strings as a stream of bytes; UTF-8 encoding is used for this research.

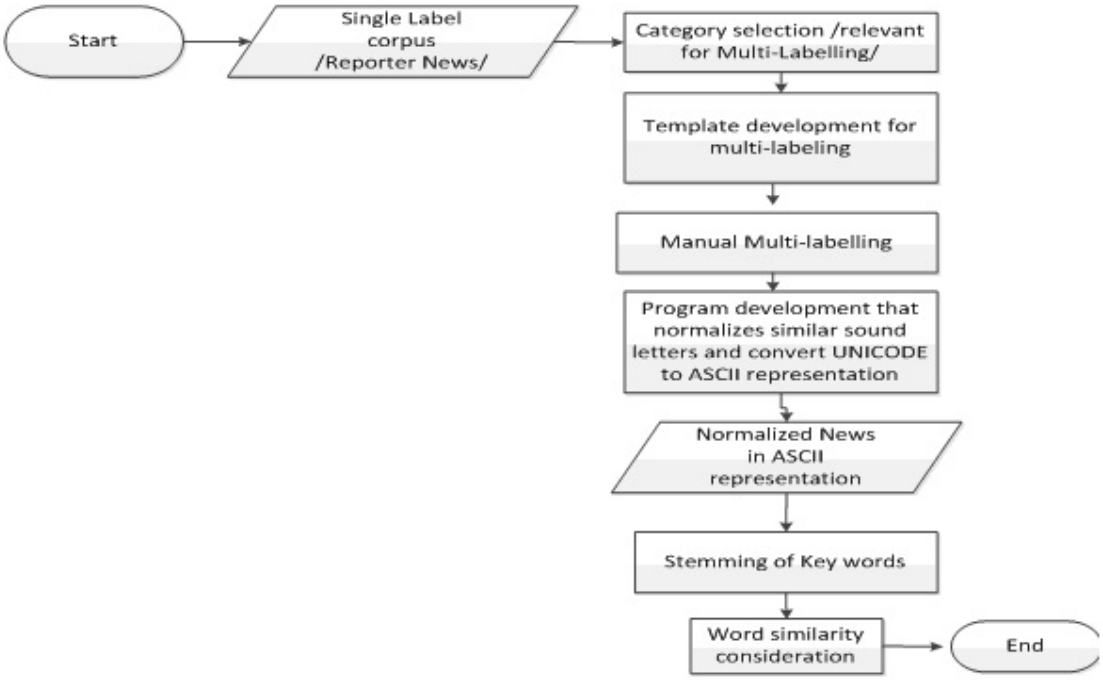

Figure 1: The Pre-processing for Multi-Labeling

The mapping of Unicode to ASCII adopts SERA. The conversion to ASCII addresses concerns regarding word spelling variation also; e.g., '**ሀይል**', '**ኀይል**', '**ሃይል**', '**ኃይል**', '**ሐይል**', '**ሓይል**' all are converted to 'hayl' as per the letter normalization procedure since all mean 'energy'. This process helps to enhance the performance of the classifier as it will not be affected by the different letters that Amharic has for the same sound. The experiment was done in two scenarios. In Scenario I, both Datasets are transformed to ARFF and loaded to the classifier just after the manual multi-labeling and conversion to ASCII is done. In Scenario II, about 300 determinant words are identified by manual multi-labeling, and using a Python program common prefixes like ' **የ** - ye', '**በ** - be', '**ቀ** - qe', and '**ለ** - le' were removed after conversion to ASCII. This is done to work with the stem only. Since different terms are used for the same concept, word similarity is also considered. Some use the term '**ሱቅ** - suq' to refer to a shop, while others use 'medebr - **መደብር**', same for '**ታክስ** -taks', '**ቀረጥ** - qereT' and '**ግብር** - gbr' to refer 'local tax'. Such variations are converted to similar expressions. Figure 2 shows the process map.

This paper approaches the task of multi-label classification with problem of transformation using Classifier Chain (CC). CC is a model of multi-label classification with problem transformation. The basic idea of this algorithm is to transform the multi-label learning problem into a chain of binary classification problems where subsequent binary classifiers in the chain are built upon the predictions of preceding ones.

CC involves |L| binary classifiers as in BM. Classifiers are linked along a chain where each classifier deals with the binary relevance problem associated with label lj € L. The feature space of each link in the chain is extended with the 0/1 label associations of all previous links.

4. Experimental Results

Two datasets are used in this research: using 800 examples for 11 labels and using 741 examples excluding one of the labels. The first dataset, Dataset1, includes label **የታክሲ ገጠመኝ** - 'Taxi Discussion'. These two datasets, are analyzed using a single train-test using 66% to 34% proportion and cross validation techniques.

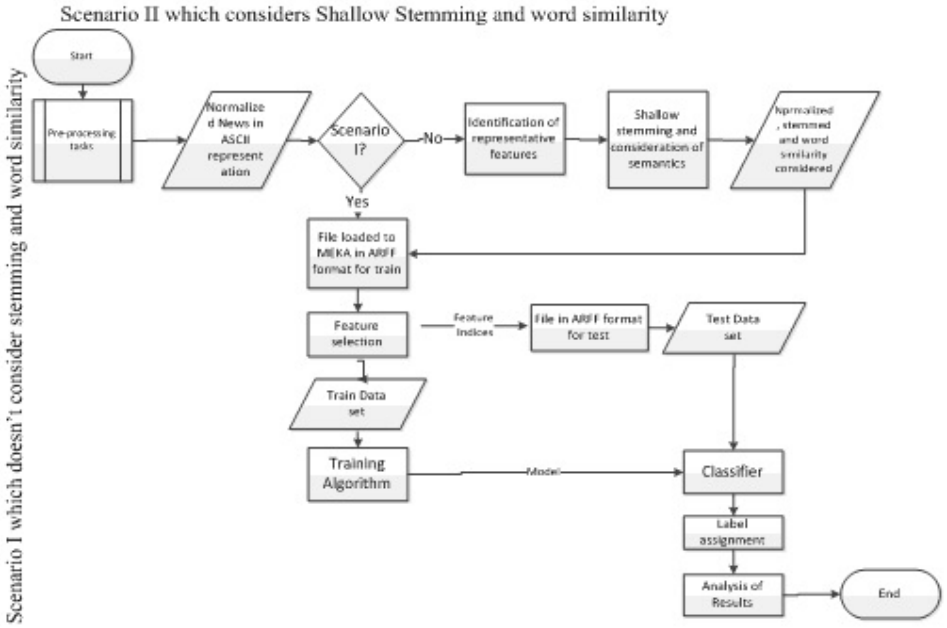

Figure 2: The training and testing process flow in two scenarios: Scenario I which doesn't consider stemming and word similarity and Scenario II which considers stemming and word similarity

4.1 Using Single Train/Test Split Percentage, Scenario I

In this approach, the classifier uses 66% of examples for training and 34% for prediction. The results are shown in Table 2.

Exclusion of 'Taxi Discussion' has resulted in slight dwindling of performance as H_loss (Hamming loss) and Rank loss are a bit higher than Dataset1 in both classifiers. As depicted in Table 1, examples of this group are single labels or the label cardinality is 1; thus the classifier's possibility of labeling them correctly is high. This is reflected on the result of hamming loss, as Dataset1 has lower hamming loss than Dataset2.

Table 2: Experimental Result for the two Datasets

Measurement	*Dataset1*		*Dataset2*	
	CC	*BR*	*CC*	*BR*
Accuracy	0.055	0.156	0.056	0.195
H_Accuracy	0.769	0.285	0.73	0.27
H_loss	0.231	0.715	0.27	0.73
Exact_match (1 – ZeroOne Loss)	0.055	0	0.056	0
One_Error	0.945	0.941	0.925	0.937
Rank_Loss	0.397	0.389	0.546	0.395
AvgPrecision	0.478	0.542	0.278	0.579
LogLoss	9.316	5.905	10.006	5.647
Recall	0.035	0.843	0.031	0.969
F-Measure	0.044	0.263	0.04	0.325
L Card_train	1.46	1.46	1.45	1.45
L Card_test	1.66	1.66	1.81	1.81

The average numbers of labels are 1.66 and 1.81 for Dataset1 and Dataset2, respectively. This is because Dataset1 maintains more single label as explained above. Label densities, can be calculated from label cardinality which are 0.15 and 0.18 which also indicate that labels are dispersed.

Regarding Rank loss, which evaluates whether irrelevant label is ranked higher than the relevant label, both classifiers didn't that much reversely order the label pairs; meaning the labels' order is somehow correct. However, concerning One-Error, they failed to rank relevant labels top, as their achievement is 0.945 and 0.925 for CC and 0.941 and 0.937 for BR for Dataset1 and Dataset2, respectively. One Error measures errors that occur when a classifier indicates a certain article 'x' is multi-label, with 'sport' at top and next 'politics' while the article really is single-

label politics. Dataset2, which is without 'Taxi Discussions' shows one error is reduced with both classifiers. This may result due to the fact that articles in this category have a tendency to fall in many other categories as well.

Hamming Loss (H loss) and Exact Match are among the most common accuracy measurements mentioned in the literature [3, 12]. Hamming Loss evaluates the fraction of misclassified instance-label pairs, i.e., a relevant label is missed or an irrelevant is predicted which is similar to the accuracy for each label, averaged across all labels. Exact Match is the accuracy of each example where all label relevance must match exactly for an example to be correct. This means even a single false positive or false negative label makes the example incorrect. Obtaining smaller score, even zero, for exact match is not unusual for string based datasets. In these measures, particularly Hamming loss, CC shows remarkable performance in both datasets which is 0.231 for the first dataset and 0.27 for the second.

Considering the level of confidence that the classifier has upon classifying instances to a certain category, Logloss result is an indication. Categorizing false positive (which is false negative to the correct label) with high confidence is not treated equally as with low confidence. BR has better performance in both datasets as its score is much lesser.

Using Cross-Validation Technique

Cross-Validation is one of the most useful techniques to evaluate different combinations of feature selection, dimensionality reduction, and learning algorithms [20]. In this paper, the test is done using 10 fold, which means 10 fold Cross Validation sets of size n/10, proportionally train on 90% of the dataset and test on 10% of the dataset. This procedure is repeated ten times and a mean accuracy is taken.

The Cross validation method performs better in H_Accuracy, One error, log loss and Recall in both datasets. One of the significant measures for multi-labeling, which is H_Accuracy, shows improvement from the previous method and reached 86.7%.

4.2 Experimental Result for Scenario II

Stemming and meaning of words consideration result is presented in Table 3.

Table 3: Experimental Result Considering Concept of Words and Stemming to a Shallow Level

Measurement	*Dataset1*		*Dataset 2*	
	CC	*BR*	*CC*	*BR*
Rank_Loss	0.411	0.37	0.543	0.393
Avg_Precision	0.472	0.516	0.284	0.579
LogLoss	9.336	5.858	9.984	5.625
Recall	0.161	0.839	0.029	0.969

The effect of Scenario II is a reflected on Rank loss, Average Precision, Log loss and Recall, while the remaining yield equivalent score to that of Scenario I. Rank loss shows a slight increase while the log loss score decreased. Recall also shows some improvement on CC classifier of Dataset1.

5. Conclusion and Future Work

Supervised classification algorithms allow computers to learn associations between examples and class labels. In this paper, two datasets composed of 800 and 741 news with 11 and 10 categories, respectively, were used from Reporter newspaper. Problem transformation model is adopted using BR and CC, where the later considers label correlations.

The complexity of number of label sets in this work goes to the maximum of 3 labels and the average is 2 labels which makes the number of possible label sets or label correlation to be less than 100 including the single labels.

Consideration of word similarity was also done by taking keywords from each domain and it resulted in some improvement on few of the measurement parameters. Regarding the performance, the tool shows promising outcome that could further be improved with availability of quality dataset.

Though local languages that use Latin, like 'Oromiffa', can directly use MEKA tool and verify the level of functionality, the main challenge that comes forth is unavailability of a tool that treats the Unicode representation as it is. Hence, the following are areas for future work.

- Larger Datasets: Achievement of better prediction performance at reasonable memory space and time complexity will be tested.
- Exhaustive pre-processing activities including in-depth stemming and concept of words recognition and their impact will be analyzed.
- Increase the average number of multi-labels as compared to the total number of labels.
- Development of multi-labelling tool that accepts Unicode can be one workable area.

References

[1] S. Niharika, V. Sneha Latha, and D. R. Lavanya, "A Survey on Text Categorization", International Journal of Computer Trends and Technology, Vol.3, Issue 1, 2012.

[2] Min-Ling Zhang and Zhi-Hua Zhou, "A Review on Multi-Label Learning Algorithms", 2012.

[3] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas, "Effective and Efficient Multi-Label Classification in Domains with Large Number of Labels", Available at: talos. csd.auth.gr/tsoumakas/publications/C24.pdf, Last accessed on August 2014.

[4] Yutaka Sasaki, "Automatic Text Classification", University of Manchester, 2008.

[5] Wen Zhang, Taketoshi Yoshida, and Xijin Tang, "Text Classification Based on Multi-Word With Support Vector Machine", Elsevier, Knowledge-Based Systems 21, 2008, pp. 879-886.

[6] Fabrizio Sebastiani, "Text Categorization", University of di Padova, Italy. Available at: http://nmis.isti.cnr.it/sebastiani/Publications/TM05.pdf, Last accessed on January 2014.

[7] Tarek Helmy and Abdirahman Duad, "Intelligent Agent for Information Extraction from Arabic Text without Machine Translation", Available at: http://ceur-ws.org/Vol687/paper2.pdf, Last accessed on November 2014.

[8] Wikipedia, the free encyclopedia, Available at: www.en.wikipedia.org/wiki/Amharic, Last accessed on August, 2014.

[9] Samuel Eyassu and Björn Gamback, "Classifying Amharic News Text Using Self-Organizing Maps", June 2005, Available at: http://eprints.sics.se/173/01/acl05-semlang.pdf.

[10] Shih-Hung Wu, Tzong-Han Tsai, and Wen-Lian Hsu, "Text Categorization Using Automatically Acquired Domain Ontology", in Proceedings of the Sixth International Workshop, July 2003, pp. 138-145.

[11] Jesse Read, "Classifier Chains for Multi-label Classification", University of Waikato, Switzerland, 2009.

[12] Jesse Read, "Scalable Multi-label Classification", University of Waikato, Switzerland, 2010.

[13] Grigorios Tsoumakas and Ioannis Katakis, "Multi-Label Classification: An Overview", International Journal of Data Warehousing & Mining, 3(3), pp. 1-13, July-September 2007.

[14] Surafel Teklu, "Automatic Catagorization of Amharic News Text, A Machine Learning Approach", Addis Ababa.

[15] Yohannes Afework, "Automatic Classification of Amharic News Text, Unpublished MSC Thesis, Addis Ababa University, 2007.

[16] Atelach Alemu, Lars Asker, Björn Gamback, and Magnus Sahlgren, "Applying Machine Learning to Amharic Text Classification", December 2006, Available at: https://www.sics.se/~gamback/publications/wocal07.pdf, Last accessed on June 13, 2015.

[17] Betelhem Mengistu, "N-gram-Based Automatic Indexing for Amharic Text", Addis Ababa University, 2002.

[18] Meron Sahlemariam, Concept Based Automatic Amharic Document Categorization, Unpublished Masters Thesis, Addis Ababa University, 2009.

[19] Charu C. Aggarwal and Chengxiang zhai, "Mining Text Data", Kluwer Academic Publishers. pp. 77-213, Available at: http://www.charuaggarwal.net/text-content.pdf, Last accessed on July 8, 2014.

[20] Predictive Modeling, Supervised Machine Learning and Pattern Classification, http://Sebastianraschka.com/Articles/2014_intro_supervised_learning.html, Last accessed on December 2014.