Automatic Amharic text news classification: Aneural networks approach

Worku Kelemework

School of Computing and Electrical Engineering, Institute of Technology, Bahir Dar University, Bahir Dar City, P.O.Box: 26, Ethiopia, e-mail: workukelem@gmail.com, workuk@bdu.edu.et

ABSTRACT

The study is on classification of Amharic news automatically using neural networks approach. Learning Vector Quantization (LVQ) algorithm is employed to classify new instance of Amharic news based on classifier developed using training dataset. Two weighting schemes, Term Frequency (TF) and Term Frequency by Inverse Document Frequency (TF*IDF), are used to weight the features or keywords in news documents. Based on the two weighting methods, news by features matrix is generated and fed to LVQ. Using the TF weighting method, 94.81%, 61.61% and 70.08% accuracies are obtained at three, six and nine classes experiments respectively with an average of 75.5% accuracy. For similar experiments, the application of TF*IDF weighting method resulted in 69.63%, 78.22% and 68.03% accuracies with an average of 71.96% accuracy.

Key words - Learning Vector Quantization (LVQ), Text news classification, Term Frequency (TF), Term Frequency by Inverse Document Frequency (TF*IDF)

INTRODUCTION

Text classification is a mapping of text documents to classes (Sebastiani, 2008). In this study, text documents are Amharic news items and classes are the categories each news item belongs.

The automated classification of texts has been flourishing in the last decade or so due to incredible increase in electronic documents on the Internet; this renewed the need for automated text classification (Klein, 2006). When Amharic is considered, electronic documents are increasing that needs automatic classification. This paper describes how to organize massively available Amharic news items into meaningful way by undergoing automatic classification.

Amharic

Amharic is the working language of the Federal Government of Ethiopia. Twenty seven million people speak the Language. It is the second largest Semitic language next to Arabic. Amharic is written from left to right similar to English unlike other Semitic languages such as Arabic and Hebrew (Wapedia, 2009). Amharic has its own writing system taken from Ge'ez alphabet. The Amharic writing system consists of a core of thirty three characters each of which occur in basic form and in six other forms called orders (Bender *et al*., 1976). Table 1 shows three core Amharic characters with their six orders.

A. Amharic Punctuation Marks

Identifying punctuation marks is vital to know word demarcation for natural language processing. According to Tewodros Hailemeskel (2003), the punctuation marks in Amharic are about ten though few of them used in computer writing system. 'Hulet Neteb' (':')-word separator and 'Arat Neteb' ('::')-sentence separator are the major punctuation marks. But, space is mostly used instead of Hulet Neteb (':') specially in computer writing system.

B. Amharic Number System

Amharic number system consists of twenty characters. They represent numbers one to ten, multiples of ten (twenty to ninety), hundred and thousand. The

Table 1. Amharic characters example

1st Order	2nd Order	3rd Order	4th Order	5th Order	6th Order	7th Order
ሀ Hä	ሁ Hu	ሂ Hi	ሃ Ha	ሄ He	ህ H	ሆ Ho
ለ Lä	ሉ Lu	ሊ Li	ላ La	ሌ Le	ል L	ሎ Lo
መ Mä	ሙ Mu	ሚ Mi	ማ Ma	ሜ Me	ም m	ሞ Mo

Table 2. Amharic characters with different forms of the same sound

Character	Other form/s of the character
ሀ (hä)	ሐ and ኀ
ሠ (sä)	ሰ
አ (ä)	ዐ
ጸ (tsä)	ፀ

numbering system is not suitable for arithmetic computation because there is no representation for zero (0) symbol, no place value, no comma and no decimal point. Amharic numbering system is used in dates specially calendar; otherwise western numerals are used in most literature these days (Bender *et al.*, 1976).

C. Problem of Amharic Writing System

There are a number of problems associated with Amharic writing system which are challenging natural language processing of Amharic documents; which are dealt below.

Redundancy of some characters: sometimes more than one character is used for similar sound in Amharic (Ethiopia Tadesse, 2002; Zelalem Sintayehu, 2001). Though the various forms have their own meaning in Ge'ez, there is no clear cut rule that shows its purpose and use in Amharic according to Bender *et al*. (1976).Table 2 illustrates the different forms of Amharic characters with similar sound.

The problem of the same sound with various characters is not only observed with core characters, but also exhibited in the same order of characters. For example, ሀ and ሃ, ኀ and ኃ; አ and ኣ; etc 9tewodros, 2003). The use of various forms of characters for the same sound poses a problem in the process of feature preparation for the classifier learning since the same word is represented in different forms. For example, the word '**ጸሀይ**' ('sun') can be represented in Amharic as **ጸሀይ**, **ጸሐይ**, **ጸኀይ**, **ፀሀይ**, **ፀሐይ**, **ፀኀይ**, etc.

Compound words: there is no standard way of writing Amharic compound words (Bender *et al.*, 1976). Space or hyphen is used between two words in a compound word; sometimes the words are merged together. According to Tewodros Hailemeskel (2003), there is a meaning difference when compound words separated by space are treated separately. For example, the word '**ሆደ-ሰፊ**' ('tolerant') formed from the words '**ሆደ**' meaning 'stomach' and '**ሰፊ**' meaning 'wide'. One can imagine how the meaning of the original word is diverted to different contexts.

Spelling variation of the same word: the same word is written in various forms (Tewodros Hailemeskel, 2003; Ethiopia, 2002; Zelalem, 2001). For example, the word '**ሰምቶአል**' ('he hears') can be written in Amharic as **ሰምቶአል**, **ሰምቷል**, **ሰምትዋል**, etc. Spelling variation may happen also in the case of translating foreign word to Amharic. For instance, the word '**ቴሌቪዥን**' ('television') can be written as **ቴሌቭዝን**, **ቴሌቭዥን**, **ቴሌቪዥን**, etc.

Abbreviation: no consistency is kept in abbreviating Amharic words (Ethiopia Tadesse, 2002) and Zelalem Sintahyeu, 2001). The word '**ዓመተ ምህረት**', meaning 'AD', can be abbreviated as **ዓም**, **ዓ.ም**, **ዓ.ም.**, **ዓ/ም**, etc.

All the aforementioned problems pose challenges since the same word is treated in different forms in the process of feature preparation for text classifier.

So, care is taken to solve such problems.

Amharic is technologically under resourced language (Solomon Tefera and Menzel, 2007). Only three researches have been tried on the area of text classification till this research is done as to the researcher's knowledge. Zelalem Sintayehu (2001), Surafel (2003) and Yohannes Afework (2007) have done research on Amharic text classification using Statistical method, K-Nearest Neighbor (KNN) and Naïve Bayes algorithms, and Support Vector Machine (SVM) and decision tree algorithms respectively. The major challenge in all the studies is the decrease in accuracy when the number of classes increases. All of these studies apply only TF*IDF weight method.

This research uses one of the neural networks learning algorithm called Learning Vector Quantization (LVQ) to study Amharic text news classification. It tries to answer the following questions:

Is neural network approach using LVQ learning method feasible for automatic Amharic text news classification?

Can we reduce the effect of increasing number of classes and news items on Amharic text news classification performance using LVQ learning method?

What is the effect of TF and TF*IDF weighting methods on Amharic text news classification performance?

Text classification using Learning Vector Quantization (LVQ)

LVQ is supervised version of Kohonon neural network (Martin-Valdivia *et al*., 2007). LVQ network has two layers called competitive layer and linear layer (Demuth and Beale, 2004). The competitive layer learns to classify input vectors. The linear layer transforms the competitive layer's classes into target classes defined by the user. The classes learned by the competitive layer are referred as subclasses and the classes of the linear layer are called target classes.

A text classifier based on a neural network approach is a network of units, where the input units represent terms or features of news, the output units represent the classes of interest (Sebastiani, 2008). Figure 1 indicates the architecture of LVQ according to Demuth and Beale (2004).

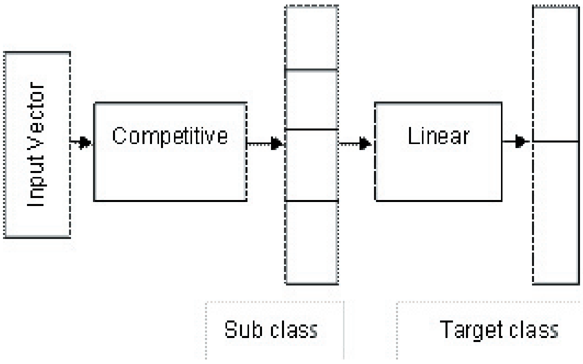

Figure 1. Architecture of LVQ

The construction of the classifier has been done using Learning Vector Quantization (LVQ) in a supervised manner. Hence, the algorithm demands training and test datasets which are pre-classified by the experts. LVQ algorithm uses training dataset for classifier construction; and test dataset for the evaluation of the classifier constructed. LVQ uses different parameters to experiment one of which is epoch, which is the learning step. In this study, different levels of epoch are experimented. Nine epoch levels are used for training: 100, 500, 1000, 1500, 2000, 2500, 3000, 3500 and 4000. 100 is the default epoch for LVQ algorithm. epoch lower than 100 are not selected based on preliminary trial. Thus, experiment is made from the default epoch level up to 4000 increasing at interval of 500 (except the first). Interval of 500 is selected to see the impact of higher epoch levels because if smaller interval is chosen it takes long time to reach to 4000.

Architecture of automatic Amharic text news classification

Amharic news items are used as an input to the system. Then preprocessing tasks like normalization (changing varying Amharic characters with similar sound to one common form, changing punctuation marks to space), tokenization, stop word and number removal, stemming, weighting terms and dimension reduction are done. After all these preprocesses, datasets are prepared in a matrix form, from which training and test datasets are prepared and used as training and testing purpose respectively. From the training dataset, model (classifier) is constructed. The model is tested using the test dataset. The testing outcome is the assignment of classes for news items that are not encountered during training. Finally, evaluation is made based on the test result using accuracy. Figure 2 shows architecture of automatic Amharic text news classification.

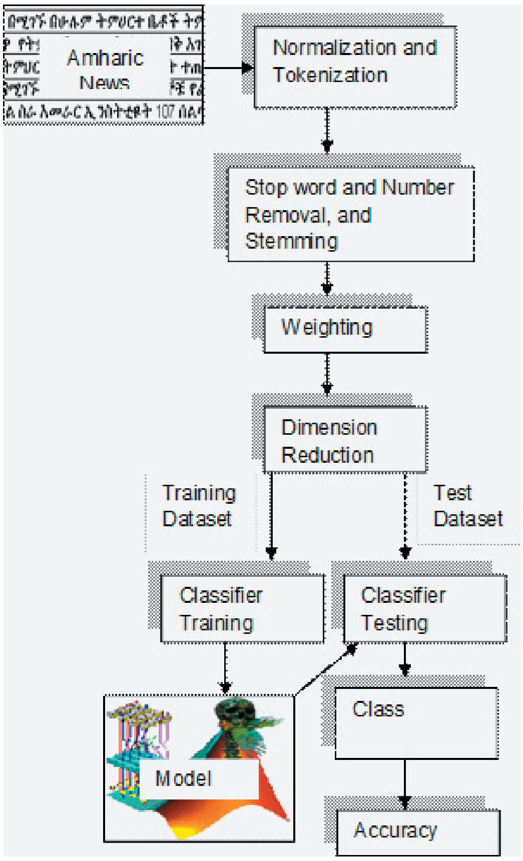

Figure 2. Architecture of automatic Amharic text news classification

THE DATA SET

The data source for this study is news of Ethiopian News Agency (ENA). The news data are classified in accordance with 13 major classifications and 103 sub classifications in the Agency. For the purpose of this study, nine classes are taken into consideration with a total of 1, 538 news items. The nine classes have been selected based on random sampling, which are Bank and insurance, Tourism development, Mines and energy, ICT, Art, Educational coverage, Weather forecast, Religious assemblies and reports, and Creativity work. The number of news items in each class and the total number of news items are shown in Table 3.

Table 3. Number of Amharic News dataset

Class No.	Class	News No.
1	Bank and insurance	297
2	Tourism development	253
3	Mines and energy	251
4	ICT	167
5	Art	152
6	Educational coverage	138
7	Weather forecast	132
8	Religious assemblies and reports	103
9	Creativity work	45
	Total	1, 538

Amharic dataset preprocessing

The tasks that are done for preprocessing of Amharic news includes tokenization, stop words and number removal, stemming, index term weight, dimension reduction and matrix generation. After the accomplishment of preprocesses, the classifier is constructed by Learning Vector Quantization (LVQ) learning method using MATLAB as a tool. Finally, the system is evaluated based on the results obtained using accuracy. The subsequent sections discuss the methods used for preprocessing the data to make it ready for classification task.

A. Tokenization

Tokenization is the process by which tokens are identified as candidates to be used as features (Baeza-Yates and Ribeiro-Neto, 1999). Candidates in the sense that stop words and numbers are removed from tokens. And tokens which do not satisfy Document Frequency (DF) thresholding are not considered.

In this study, words are taken as tokens. All punctuation marks are converted to space and space is used as a word demarcation. Hence, if a sequence of characters is followed by space, that sequence is identified as a word.

B. Stop Word and Number Removal

Stop words are non content bearing words, which are less discriminating among documents since they appear in most of them features (Baeza-Yates and Ribeiro-Neto, 1999).

There are common stop words in Amharic which are used for grammatical purposes like **ነው**, **ነበር**, **ሆኖም**, **እና**, **ነገርግን**, etc, which are non informative to identify documents. In addition to the common stop words, there are also news specific stop words like **ገለፁ**, **ዘግበዋል**, **አስታወቀ**, etc; their use is for elaboration and common to all news in accordance with the reporters of ENA. Because of the unavailability of standard stop list done by previous researchers, the researcher of this study is obligated to develop stop list.

Since stop words are highly frequent words, total frequency of terms aided by manual inspection, is the method employed in the process of identification of stop words. Stop list is prepared after identifying stop words; the list that contains, words which have to be removed from tokens generated during the tokenization process. The need of manual inspection is, because of frequently occurring keywords. For example, the word '**ቱሪዝም**' ('tourism') is the most frequent word in the class 'Tourism development', which is crucial in discriminating the class. Hence, such words are not included in the stop list.

The purpose of identifying stop words is, to remove such words from the list of index terms. Index terms are features and believed to represent news or discriminate one news item from the others; whereas, stop words are not. Hence, using those words in the list of index terms is unimportant. That is why their exclusion from index term list is vital.

In most cases, numbers are less discriminating among documents (Baeza-Yates and Ribeiro-Neto, 1999). In this study also, numbers are not considered as index terms. So, index terms list does not contain any number.

C. Stemming

Stemming is changing varying words, due to grammatical reasons, to the root form of the word (Frakes and Baeza-Yates, 2002). Stemming is one of the preprocessing made on Amharic text news for this study. Stemmer that can remove common Amharic prefixes and suffices is developed. Table 4 shows an example of the prefixes and suffices removed and an example under each affix.

The stemmer developed for this study is based on Nega and Willett (2002). In such case, rules are

Table 4. Example of affix removed during stemming

Type	Affix	Example	
		Word	Translated to
Prefix	ለ	ለጂማ	ጂማ
	ስለ	ስለጂማ	ጂማ
	በ	በጂማ	ጂማ
Suffix	ም	ጂማም	ጂማ
	ና	ጂማና	ጂማ
	ን	ጂማን	ጂማ

applied to find the stem of Amharic words. The rules to remove prefix or suffix from a given word may not hold true always. For instance, removing **‘ው’** (‘wu’) from the word **‘ሰው’** (‘person’) would give **‘ሰ’** (‘se’), which is meaningless; and removing **‘በ’** (‘be’) from **‘በልግ’** (‘autumn’) gives **‘ልግ’** (‘lg’),which does not represent the original meaning. Hence, two exception lists are prepared for which affix removal rules do not applied;

List of words that prefix removal rule does not hold true and list of words from which suffix removal rule is not applied.

The stemmer developed takes words as an input and removes prefix of the word. After the prefix is removed, the word is again checked if it lasts with suffix in the suffix list, if so, the suffix is removed from the word. Table 5 shows an example.

Table 5. Example of stemming

	Example 1	**Example 2**	**Example 3**	**Example 4**
Input:	ጂማ	ጂማን	የጂማ	የጂማን
Prefix	No	No	የ	የ
Output1	ጂማ	ጂማን	ጂማ	ጂማን
Suffix:	No	ን	No	ን
Final output:	ጂማ	ጂማ	ጂማ	ጂማ

D. Index Term Weight

All index terms are not equally important in representing and discriminating a document; it is thus, required to measure how important a term is with regard to representation and discrimination of a document (Giorgino, 2008; Liao *et al*., 2003). Term Frequency (TF) and Term Frequency by Inverse Document Frequency (TF*IDF) are the weighing schemes used in this study. TF, IDF and TF*IDF are explained below based on Baeza-Yates and Ribeiro-Neto (1999) and (Manning *et al*., 2008).

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

$$F \quad = FREQ_{ik} \qquad (1)$$

In (1), FREQik is the number of occurrence of term k, in document i. TF is zero if the term does not appear in document i.

IDF is a measure of the general importance of the term. (2) depicts IDF of a term.

$$IDF = \log_2^{\frac{N}{d_k}} \qquad (2)$$

In (2), N is the total number of documents in the collection, dk the number of documents in which term k occurs. TF*IDF is the combination of TF and IDF weighting methods. TF*IDF incorporates two intuitions:

If an index term occurs more frequently in a document, the index term is more important for that document, the Term Frequency intuition. If more number of documents contain the index term, the index term is less discriminating between the documents, the Inverse Document Frequency intuition.

$$F \quad * IDF = FREQ_{ik} * \log_2^{\frac{N}{d_k}} \qquad (3)$$

In (3), FREQik is the number of occurrence of term k in document i, N is the total number of documents in the collection, dk the number of documents in which term k occurs.

E. Dimension Reduction

The feature space comprises one new dimension for each unique term that occurs in the text documents, which can lead to tens of thousands of dimensions for even a small-sized text collection, so, there is a need to integrate dimension reduction phase in text classification (Skarmeta *et al*., 2000; Yi and Beheshti. 2008).).

After identifying the number of tokens generated during tokenization, stop words and numbers removal and stemming are applied to reduce the number of tokens to be used as features. But still the dimension has to be reduced so that the most important features of each class is identified. The need to reduce the dimension is: Irrelevant features are removed which may affect performance badly and for convenient computational complexity.

In this study, Document Frequency (DF) thresholding is used to reduce the dimension of features generated. DF is the number of documents that contain a certain feature (Krishnakumar, 2006). The system is supported by manual observation; whether stop words which are not eliminated during stop word removal are mixed and if there are important features which do not satisfy the threshold.After all the preprocessing done on the dataset, 80 features are identified to represent news.

Matrix

The input to the learning algorithm is a matrix generated with the value of term weights using TF and TF*IDF. Table 6 and Table 7 show an example of matrix generated for the nine classes experiment using TF and TF*IDF weight methods respectively; the rows and columns are reduced for viewing purpose. That means all 80 terms (features) and all 9 classes are not shown in the Tables.

Table 6. Matrix using TF weight method

ቱሪዝም 'turizm'	0	0	0	0	0	0	0
ትምህርት 'tmhrt'	0	0	0	0	0	0	0
ባንክ 'bank'	0	1	1	0	3	0	0
ምንዛሪ 'mnzari'	2	1	0	1	1	2	2
ማእድን 'maldn'	0	0	0	0	0	0	0
Class	1	1	1	1	1	1	1

Table 7. Matrix using TF*IDF weight method

ቱሪዝም 'turizm'	0	0	0	0	0	0	0
ትምህርት 'tmhrt'	0	0	0	0	0	0	0
ባንክ 'bank'	0	3.46	3.46	0	10.38	0	0
ምንዛሪ 'mnzari'	6.64	3.32	0	3.32	3.32	6.64	6.64
ማእድን 'maldn'	0	0	0	0	0	0	0
Class	1	1	1	1	1	1	1

In Table 6 and Table 7, zero indicates that the feature does not occur in that class; otherwise, its weight value is used to show its importance. In both Tables, class 1 indicates "Bank and Insurance" as depicted in Table 3.

Amharic text news classification performance

Classifier is constructed using training dataset using 66.67% of the total dataset. The remaining 33.33% of the total dataset is used to test the accuracy of the classifier. Fifty four experiments were carried out for the nine epoch levels in the three, six and nine news classes using both TF and TF*IDF weight methods excluding preprocessing experiments. The experiment was carried out by considering increasing number of classes. Three classes: ICT, Art and Educational coverage.

Six classes: ICT, Art, Educational coverage Weather forecast, Religious assemblies and reports, and Creativity work. Nine classes: Bank and insurance, Tourism development, Mines and energy, ICT, Art, Educational coverage, Weather forecast, Religious assemblies and reports, and Creativity work.

Table 8 and Table 9 show accuracy result of the classifier evaluated by the test dataset using TF and TF*IDF weight methods respectively.

The best accuracy obtained for the three, six and nine classes using the two weight methods are indicated in Table 10.

Table 8. Accuracy using TF weighting Scheme at 3, 6 and 9 classes at various epoch levels

Class	3 Classes	6 Classes	9 Classes
Epoch			
100	69.63%	61.61%	62.09%
500	68.89%	61.61%	62.30%
1000	68.89%	61.61%	61.48%
1500	69.63%	61.61%	70.08%
2000	94.81%	61.61%	62.30%
2500	68.89%	61.61%	61.07%
3000	69.63%	61.61%	61.48%
3500	68.89%	61.61%	61.89%
4000	68.89%	61.61%	61.68%

Table 9. Accuracy using TF*IDF weighting Scheme at 3, 6 and 9 classes at various epoch levels

Class	3 Classes	6 Classes	9 Classes
Epoch			
100	69.63%	57.78%	62.09%
500	62.22%	70.22%	62.30%
1000	65.19%	78.22%	61.48%
1500	69.63%	57.78%	68.03%
2000	65.19%	57.78%	62.30%
2500	65.19%	58.22%	62.30%
3000	65.19%	72.89%	62.30%
3500	69.63%	57.78%	54.51%
4000	62.22%	57.78%	62.30%

Table 10.Best accuracy at increasing No. of classes and news using TF and TF*IDF weight methods

Classes	Accuracy	
	TF	**TF*IDF**
Three	94.81%	69.63%
Six	61.61%	78.22%
Nine	70.08%	68.03%
Average	75.50%	71.96%

DISCUSSION

For the three classes, TF weight method is better than TF*IDF weight method by 25.18%. But for the six classes experiment TF*IDF weight method is better than TF weight method by 16.61% than TF weight method. The result of nine classes experiment testifies that TF weight method scored better accuracy than TF*IDF weight method by 2.05%. The average of all the experiments indicates that TF weight method registered better accuracy by 3.54% than TF*IDF weight method.

The main performance difference between the two weighting schemes happens because of the range of values in the weighting schemes. In the datasets, TF weight value is between 0 and 5 and TF*IDF weight value is in the range of 0 and 45. This affects the classifier accuracy. Because, according to [22], it is recommended to have maximum value of 1 and minimum value of -1 for the input pattern of LVQ algorithm. This seems plausible for the greater accuracy result of TF weight method than TF*IDF weight method.

As depicted in Table 8, using TF weight method the best accuracy is registered at three classes experiment. The least accuracy is recorded at the six classes experiment. The nine classes experiment, resulted accuracy lower than the three classes but higher than the six classes experiment. Hence, we can say that the increase in the number of classes and news are not the determinant factor for the decrease of performance with regard to the LVQ algorithm.

Based on Table 9, least accuracy for the TF*IDF weight method is scored in the nine classes experiment with less (1.6%) difference in the three classes experiment. The best accuracy for this weighting method is recorded in the six classes experiment. The three classes experiment resulted in the second best accuracy. Like the TF weight method, it can be said that the increase in the number of classes and news items are not the major factors in the reduction of performance using LVQ algorithm for the Amharic text classifier.

CONCLUSION

This study tried to see the potential application of Learning Vector Quantization to automatic classification of Amharic text news. Under this umbrella, effectiveness of text news classifier at increasing level of classes and news items has been investigated using TF and TF*IDF weighting methods. The concluding remarks are described below:

The best accuracy using TF weighting scheme has been obtained at three classes, which is 94.81%. The least accuracy recorded for TF weighting scheme has been scored at six classes, which is 61.61%. On the other hand, for TF*IDF weighting scheme, the best accuracy recorded at six classes that is about 78.22% and the least accuracy has been obtained at nine classes that accounts to 68.03%. TF weighting scheme is better in accuracy than TF*IDF weighting scheme by 3.54% on average from the three, six and nine classes experiments.

Increase in class does not affect the performance of the classifier unlike previous studies, which show consistent decrease in accuracy as the number of classes increase. In the course of training using LVQ algorithm, it is found that computational time increases as the number of news items, features and classes increases.

Generally, the study shows that Learning Vector Quantization can be employed to automatic Amharic text classification but an integration of standard preprocessing techniques is crucial. Classifier is constructed using LVQ-neural network algorithm. But the accuracy obtained has to be improved. The recommendations given revolve around improving accuracy, facilitating Amharic text classification, or untouched problems related to text classification for Amharic context and recommendations for the agency, ENA.

Stemmer: standard stemmer that can be applied on Amharic is vital to decrease feature size. The result obtained on this study is encouraging in reducing the size of features by applying stemmer. If standard stemmer is available, it can play great role in the reduction of features as a result computational complexity can be decreased and effectives can be enhanced. Hence, there is a need to investigate on Amharic stemmer. Prior to stemming, actually there is a need to use stop word removal system. In this study, very few news specific and language specific stop words are identified, using these stop words good result is obtained. If complete stop word list is prepared for the domain considered, that again have an important positive effect on reducing the size of the features.

Spell checker: spelling error multiplies features by forming different features for the same word. As a result, spell checker researches are essential for guiding the development of Amharic spell checkers. In addition to spelling errors, some Amharic words can take varying forms due to the presence of varying Amharic characters but with similar sound. Here, the recommendation is extended to Amharic language experts; to devise a clear rule on the use of those characters. Compound words and abbreviations are also written in various forms; hence, the language experts again recommended on standardizing the way compound words and abbreviations are written.

Dimension reduction: in this study, Document Frequency (DF) thresholding is used as a method of dimension reduction of features; other dimension reduction techniques like Information Gain (IG), x ^{2}-test (CHI), etc can be used for reducing the size of features. Corpus preparation: in other languages, it is very common to prepare corpus for research purpose; unfortunately, we are not lucky for not having standard corpus for Amharic text classification, as to the researcher's knowledge. Researchers can devote much time on their work if standard corpus is prepared for Amharic classification experiments like 'Reuters-21578' for English. Feature preparation: the features that can represent classes are selected using words in this study. But confusion occurred when the words are common across classes that resulted in misclassifications. Hence, there is a need to undergo research on text classification that considers features selected based on phrases or using ontology.

Classification types: the data on ENA SQL server exhibit hierarchical in nature. As far as the researcher's knowledge, this problem is not researched for Amharic. So, this is potential area of research. And some news items in ENA reveal the characteristics of more than one class. But ENA uses only single-label classification scheme. So, it is recommended for ENA to start implementation of multi-label classification scheme so that the true characteristics of news items are exhibited. This also helps researchers to undergo study on multi-label classification of Amharic news.

ENA: manual classification is used in ENA till now. The results of Amharic text news classification researches are promising. Hence, the company shall start to think the implementation on automatic classification for Amharic news.Other domains**:** as to the knowledge of the researcher, Amharic text classification is still tried on news items text only. Other areas, like classifying 'research papers', can be researched for Amharic documents.

REFERENCES

Baeza-Yates, R and Ribeiro-Neto B. (1999). **Modern information Retrieval**. Addison-Wesley: New York,

Bender, M., Bowen, J., Cooper, R. and Ferguson C. (1976). **Language in Ethiopia**. Oxford University Press, London.

Demuth, H and Beale, M. (2004). Neural Network Toolbox for Use With Matlla. The Math Works inc. Natica.

Ethiopia Tadesse. (2002). Application of case-based reasoning for Amharic legal precedent retrieval: a case study with the Ethiopian labor law. M Sc Thesis, Addis Ababa University, Ethiopia.

Frakes, W and Baeza-Yates R. (2002). Information retrieval: data structures and algorithms, Prentice Hall.

Giorgino, T. (2004). An introduction to text classification. www.cs.cornell.edu/people/tj/publications/joachims_98a.pdf Retrieved on October 13, 2008.

Klein B. (2004). Text categorization or classification. http://www.bklein.de/text_classification.php, 2004. Retrieved on October 12, 2008.

Krishnakumar, A. (2006). Text categorization: building a KNN classifier for the Reuters-21578 collection. http://en.scientificcommons.org/42606011. Retrieved on October 12, 2008.

Liao, C., Alpha, S and Dixon, P. (2003). Feature preparation in text categorization.

Manning, C., Raghavan, P and Schütze, H. (2008). Introduction to information retrieval Cambridge University Press, Cambridge.

Martín-Valdivia, M., Ureña-López, L. And García-Vega, M. (2007). The Learning Vector Quantization algorithm applied to automatic text classification tasks. *Neural Networks*. **20**: 748-756.

Nega, Alemayehu and Willett P. (2002). Stemming of Amharic words for information retrieval. *Literary and Linguistic Computing*. 17:1-17.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*. **34**: 1-47.

Skarmeta, A., Bensaid, A and Tazi, N. (2000). Data mining for text categorization with semi-supervised Agglomerative: Hierarchical Clustering. *International Journal of Intelligent*

Systems. 15:63-646.

Solomon Tefera and Menzel, W. (2007). Syllable-based speech recognition for Amharic. In **Proceedings of the 5th Workshop on Important Unresolved Matters**, pp. 33–40. Addis Ababa Ethiopia.

Surafel Teklu (2003). Automatic categorization of Amharic news text: a machine learning approach. MSc Thesis, Addis Ababa University, Ethiopia.

Tewodros Hailemeskel. (2003). Amharic text retrieval: An Experiment Using Latent Semantic Indexing (LSI) with Singular value Decomposition (SVD). M Sc Thesis, Addis Ababa University, Ethiopia.

Thulasiraman, P. (2005). Semantic classification of rural and urban images using Learning Vector Quantization. MSc Thesis, Madras University. India.

ttp//www.oracle.com/technology/products/text/pdf/feature_preparation.pdf. Retrieved on October 12, 2008.

Wapedia (2009). አማርኛ, http://wapedia.mobi/am/" Retrieved on April 24, 2009.

Yi, K. and Beheshti, J. (2004). A comparative study on feature selection of text categorization for Hidden Markov Models. http://www.jsbi.org/journal/GIW02/GIW02F006.pdf. Retrieved on September 24.

Yohannes Afework (2007). Automatic Amharic text categorization. MSc Thesis, Addis Ababa University, Ethiopia.

Zelalem Sintayehu. (2001). Automatic classification of Amharic news items: The case of Ethiopian News Agency. MSc Thesis, Addis Ababa University: Ethiopia.