American Journal of Engineering Research (AJER)
e-ISSN : 2320-0847 p-ISSN : 2320-0936
Volume-03, Issue-04, pp-251-259
www.ajer.org

Research Paper Open Access

A Two Step Data Mining Approach for Amharic Text Classification

Seffi Gebeyehu[1] Dr.Vuda Sreenivasa Rao[2]
Lecturer, School of Computing and Electrical Engineering, IOT, Bahir Dar University, Ethiopia[1].
Professor, School of Computing and Electrical Engineering, IOT, Bahir Dar University, Ethiopia[2].

Abstract: **-** Traditionally, text classifiers are built from labeled training examples (supervised). Labeling is usually done manually by human experts (or the users), which is a labor intensive and time consuming process. In the past few years, researchers have investigated various forms of semi-supervised learning to reduce the burden of manual labeling. In this paper is aimed to show as the accuracy of learned text classifiers can be improved by augmenting a small number of labeled training documents with a large pool of unlabeled documents. This is important because in many text classification problems obtaining training labels is expensive, while large quantities of unlabeled documents are readily available. In this paper, intended to implement an algorithm for learning from labeled and unlabeled documents based on the combination of Expectation- Maximization (EM) and two classifiers: Naive Bayes (NB) and locally weighted learning (LWL). NB first trains a classifier using the available labeled documents, and probabilistically labels the unlabeled documents while LWL uses a class of function approximation to build a model around the current point of interest. An experiment conducted on a mixture of labeled and unlabeled Amharic text documents showed that the new method achieved a significant performance in comparison with that of a supervised LWL and NB. The result also pointed out that the use of unlabeled data with EM reduces the classification absolute error by 27.6%. In general, since unlabeled documents are much less expensive and easier to collect than labeled documents, this method will be useful for text categorization tasks including online data sources such as web pages, e-mails and news group postings. If one uses this method, building text categorization systems will be significantly faster and less expensive than the supervised learning approach.

Keywords: **-** *Text Classification, Expectation Maximization, Naïve Bayes, LWL.*

I. INTRODUCTION

In recent times, there has been an explosive growth in the amount of data that is being collected in the business and scientific area. Data mining techniques can be used to discover useful patterns that in turn can be used for classifying new instances of data (Cherkassky and Mulier, 1998).

Knowledge discovery as a processis depicted in Figure 1 and consists of an iterative sequence of the following steps:

1. Data cleaning (to remove noise and inconsistent data).

2. Data integration (where multiple data sources may be combined).

3. Data selection (where data relevant to the analysis task are retrieved fromthe database).

4. Data transformation (where data are transformed or consolidated into forms appropriatefor mining by performingsummary or aggregation operations, for instance).

5. Data mining (an essential process where intelligent methods are applied in order toextract data patterns).

6. Pattern evaluation (to identify the truly interesting patterns representing knowledgebased on some interestingness measures).

7. Knowledge presentation (where visualization and knowledge representation techniquesare used to present the mined knowledge to the user).

Steps 1 to 4 are different forms of data preprocessing, where the data are preparedfor mining. The data mining step may interact with the user or a knowledge base. Theinteresting patterns are presented to the user and may be stored as new knowledge inthe knowledge base.

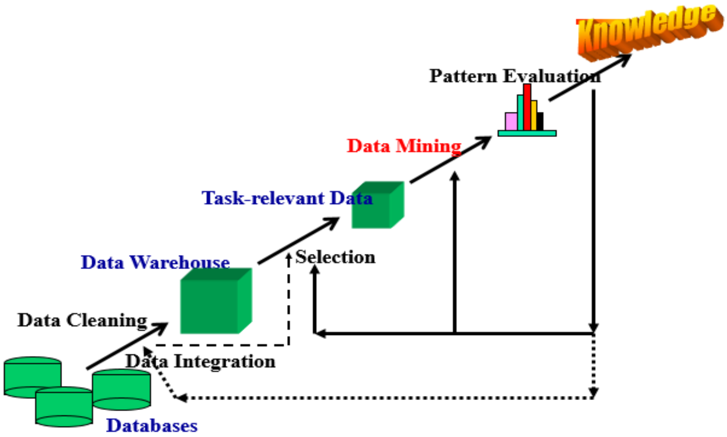

Figure 1. Knowledge Discovery (KDD) Process.

Classification is an important problem for machine learning and data mining research communities. The basic idea of a classification algorithm is to construct a classifier according to a given training set. Once the classifier is constructed, it can predict the class value(s) of unknown test data sample(s).Suppose we work for a web site that maintains a public listing of job openings from many different companies. A user of the web site might find new career opportunities by browsing all openings in a specific job category. However, these job postings are speared from the Web, and do not come with any category label. Instead of reading each job post to manually determine the label, it would be helpful to have a system that automatically examines the text and makes the decision itself. This automatic process is called *text classification*. Text classification systems categorize documents into one (or several) of a set of pre-defined topics of interest.

Text classification is of great practical importance today given the massive volume of online text available. In recent years, there has been an explosion of electronic text from the World Wide Web, electronic mail, corporate databases, chat rooms, and digital libraries. One way of organizing this overwhelming amount of data is to classify it into descriptive or topical taxonomies. For example, Yahoo maintains a large topic hierarchy of web pages. By automatically populating and maintaining these taxonomies, we can aid people in their search for knowledge and information.

The classic approach to build a text classifier is to first (often manually) label a set of training documents, and then apply a learning algorithm to build the classifier. Manual labeling of a large set of training documents is a bottleneck of this approach as it is a time consuming process. To deal with this problem, (Nigam *et al.*, 2000; Blum & Mitchell, 1998) propose the idea of using a small labeled set of every class and a large unlabeled set for classifier building. These research efforts aim to reduce the burden of manual labeling.

This paper uses Expectation-Maximization (EM) to learn classifiers that take advantage of both labeled and unlabeled data. EM is a class of iterative algorithms for maximum likelihood or maximum a posteriori estimation in problems with incomplete data (Dempster, Laird, & Rubin, 1977). In our case, the unlabeled data are considered incomplete because they come without class labels. The algorithm first trains a classifier with only the available labeled documents, and uses the classifier to assign probabilistically weighted class labels to each unlabeled document by calculating the expectation of the missing class labels. It then trains a new classifier using all the documents both the originally labeled and the formerly unlabeled and iterates. In its maximum likelihood formulation, EM-performs hill-climbing in data likelihood space, finding the classifier parameters that locally maximize the likelihood of all the data both the labeled and the unlabeled. We combine EM with Naive Bayes, and EM with LWL classifiers that are commonly used in text classification.

The data pre-processing carried out in this research involved developing and adopting tools for:

- Data cleaning which involves removal of repeated news items, manual classification ofunclassified news items, removal of entry errors, etc.
- Identifying and removing stop words and word affixes.
- Correcting for commonly missing letters in VG2 which sometimes occur during dataconversion.
- Normalizing the different letters of the Amharic script that have the same sound.
- Correcting major spelling variations in words focusing in transliteration problems.
- Analyzing compound words to correct for inconsistent usage of the compound words (the use of compound words sometimes as a single-word and sometimes as two or more words) as well as to give consideration for the semantics of the compound words. Moreover, the processed Amharic documents were collected in their pre-defined categories and the whole data were changed to arff (attribute reference file format) file format, which is suitable for the Weka open source application package used for the automatic classification.

The goal of this paper is to demonstrate that supervised learning algorithms using a small number of labeled examples and a large number of *unlabeled* examples create high accuracy text classifiers. In general, unlabeled examples are much less expensive and easier to come by than labeled examples. This is particularly true for text classification tasks involving online data sources, such as web pages, email, and news stories, where huge amounts of unlabeled text are readily available. Collecting this text can frequently be done automatically, so it is feasible to quickly gather a large set of unlabeled examples. If unlabeled data can be integrated into supervised learning, then building text classification systems will be significantly faster and less expensive than before.

II. PREVIOUS WORKS

Expectation-Maximization is a well-known family of algorithms with a long history and many applications. Its application to classification is not new in the statistics literature. The idea of using an EM-like procedure to improve a classifier by" treating the unclassified data as incomplete" is mentioned by R. J. A. Little among the published responses to the original EM paper (Dempster et al., 1977). A discussion of this "partial classification" paradigm and descriptions of further references are made by McLachlan and Basford (1988: p 29).

2.1. Comparison of the proposed method:

Different researches have been conducted on Amharic text categorization from supervised documents **(**Zelalem, 2001; Surafel, 2003; yohannes, 2007; Worku, 2009; Alemu 2010).

However, this paper is different since it deals with semi supervised learning approach.

Gebrehiwot (2011) conducted research on Tigrigna text categorization from unlabeled documents using repeated bisection and direct k-means for clustering and SVM techniques for classification. He showed that SMO support vector classifiers perform better than j48 and decision tree classifiers. He also pointed out the ambiguity of Tigrigna language which demands further research to apply ontology-based hierarchical text categorization. However, he used feature selection having high discriminative power which is not a simple task. This paper doesn't use any feature selection.

Recent work by some of the authors combines active learning with Expectation- Maximization (McCallum & Nigam, 1998). EM is applied to the unlabeled documents both to help inform the algorithm's choice of documents for labeling requests, and also to boost accuracy using the documents that remain unlabeled (as in this paper). Experimental results show that the combination of active learning and EM requires only slightly more than half as many labeled training examples to achieve the same accuracy as either active learning or EM alone. Our method is different since it doesn't need large number of training examples.

Our work is an example of applying EM to fill in missing values for which the missing values are the class labels of the unlabeled training examples. Work by Ghahramani and Jordan (1994) is another example in the machine learning literature of using EM with mixture models to fill in missing values. Whereas we focus on data where the class labels are missing, they focus on data where features other than the class labels are missing.

Recent work by (Basu, Banerjee & Monney, 2002, cited in Yu, et al., 2003) tries to bring clustering closer to classification by using a small number of labeled documents as seeds to initialize *k*-means clustering. The original *k*-means algorithm selects initial seeds randomly. Again, this paper is different. It only uses EM for categorizing classes to clusters and fill in the missing values from large unlabeled documents.

Yu, et al. 2003, propose a method to solve the problem of supervised learning by combining clustering with EM and feature selection. The technique can effectively rank the words in the unlabeled set according to their importance. The user then selects some words from the ranked list for each class. This process requires less effort than providing words with no help or manual labeling of documents. But this paper is different since it doesn't use any feature selection for classifying a mixture of labeled and unlabeled documents.

III. TECHNIQUES FOR AMHARIC TEXT CLASSIFICATION

We show an outline drawing about a whole flow of general text classification process in Fig. 2. A procedure of general text classification is as indicated below.

1. Data Selection: We extract training data and test data necessary for classification process from original data.

2. Preprocessing: In general, since training data and test data which we just extract include noises, we usually remove them. Especially in text classification, we carry out preprocessing for the data, which includes a removal of disabled words, a conversion from a plural form to a singular form, and conversion derivative words into ones with an original form.

3. Feature Selection: We carry out feature selection from preprocessed training data. In this process, there are some methods for feature selection due to difference of ways to express documents.

4. Making Vectors: We convert all the test data into document vectors based on selected features.

5. Text Classification: We apply text classification algorithms to the document-vectorized test data, and classify documents into categories. In this process, there are some methods for text classification due to difference of ways to build up documents and different machine learning algorithms are proposed.

6. Evaluation: We determine classification accuracy by using some measure for evaluation.

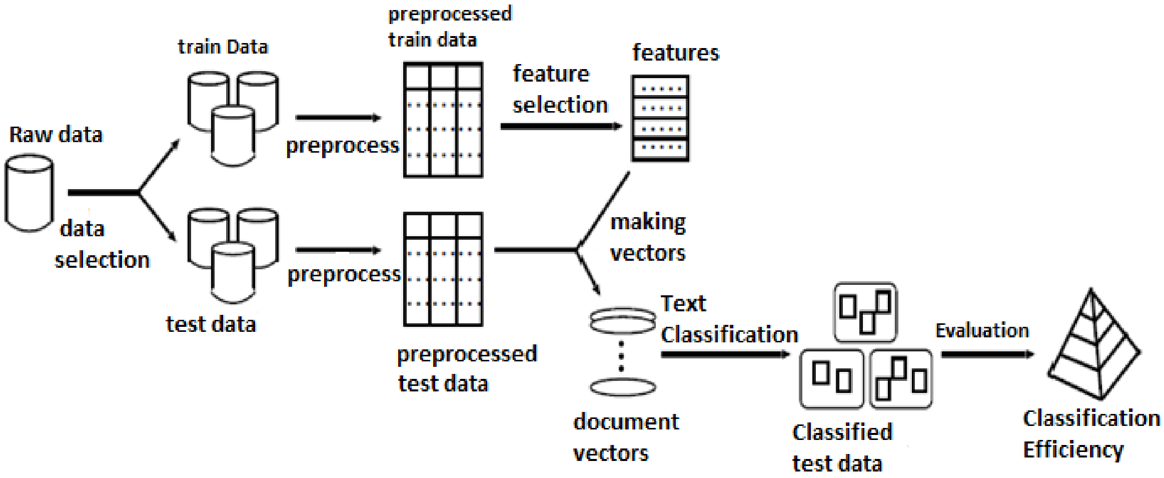

Figure 2. Flow diagram of text classification process.

The basic tools and techniques required to classify Amharic documents are text preprocessing, document clustering, and classifier model building. In order to preprocess the Amharic documents, different text preprocessing techniques such as tokenization, stemming, and stop word removal are used.

3.1.Document clustering:

Document clustering is used to discover natural groups in data set without having any background knowledge of the characteristics of the data in the documents. There are different document clustering algorithms. They are mainly divided in to hierarchical and partitioning clustering algorithms (Matthew, 2004). We used an expectation maximization(EM) a partitioning clustering algorithms to cluster a set of Amharic documents directly in to a set of groups (clusters) from a mixture of labeled and unlabeled documents.

3.1.1. The EM Algorithm:

An expectation-maximization (EM) algorithm is used in statistics for finding maximum likelihood estimates of parameters in probabilistic models, where the model depends on unobserved latent variables. EM alternates between performing an expectation (E) step, which computes an expectation of the likelihood by including the latent variables as if they were observed, and maximization (M) step, which computes the maximum likelihood estimates of the parameters by maximizing the expected likelihood found on the E step. The parameters found on the M step are then used to begin another E step, and the process is repeated (Dempster et al. 1997). The EM algorithm is a popular class of iterative algorithms for maximum likelihood estimation for problems involving missing data. It is often used to fill the missing values in the data using existing values (Li et al., 2004).

3.1.2. Assumptions of basic EM.

- The data is produced by a mixture model.
- There is a one-to-one correspondence between mixture components and classes.
- These assumptions are usually violated by real-world textual data.
- The benefits of unlabeled data are less clear when the assumptions don’t hold.

3.2. Construction of Text Classifiers.

Many statistical classification algorithms and machine learning techniques have been successfully applied to text categorization. These include methods such as Naïve Bayes, SVM, Decision Tree, Linear Discriminant Analysis (LDA) (Hull, 1996), Neural Networks, KNN, and the likes (Pan, 2006). Here we introduce only those methods that are directly relevant to this work, as we apply them to our classification task.

The clustered documents using EM are used to classify Amharic documents using the naïve Bayes and locally weighted learning classifiers. As a result, the best classification scheme from the two above constructs a model from the cluster categories of the training collection.

3.2.1. Lazy methods

Lazy learning methods defer processing of training data until a query needs to be answered. This approach usually involves storing the training data in memory, and finding relevant data from the data base to answer a particular query. This type of learning is also referred to as memory-based learning. Relevance is often measured using a distance function with high points having high relevance. Locally Weighted learning that uses locally weighted training to average, interpolate between, extrapolate from, or otherwise combine training data (Vapnik, 1992 cited in Christopher G., 1999).

In most learning methods a single global model is used to fit all of the training data. Since the query to be answered is known during processing of training data, training query specific local models is possible in lazy learning. Local models attempt to fit the training data only in a region around the location of the query (the query point).

Learning process will be started on the stored examples only after when a new query instance is encountered (Mitchell, 1997). Nearest Neighbor algorithm is the one of the most basic instance-based methods. The instance space is defined in terms of Euclidean distance. However, since Euclidean distance is inadequate for many domains, several improvements were proposed to the instance-based nearest neighbor algorithm which are known as IB1 to IB5 (Martin, 1995).

The Locally Weighted Learning algorithm (LWL) is similar to other lazy learning methods; however it behaves differently when classifying a new instance. LWL algorithm constructs a new Naïve Bayes model using a weighted set of training instances (Frank et al., 2003). It empirically outperforms both standard Naïve Bayes as well as nearest-neighbor methods on most data sets tested by those authors. The study uses LWL as a classifier on the Amharic text data.

IV. THE PROPOSED METHOD

The specific approach we described here is based on a combination of three well-known learning algorithms: the Naive Bayes classifier (Lewis & Ringuette 1994; McCallum & Nigam 1998) and the Expectation-Maximization (EM) algorithm (Dempster, Laird, & Rubin 1977). The Naive Bayes algorithm is one of a class of statistical text classifiers that uses word frequencies as features. In addition to Naive Bayes, a lazy method called locally weighted learning (LWL) is also applied for classifying large unlabeled and few labeled documents.

A mixture of both labeled and unlabeled text documents collected from ENA are used to conduct the experiment on Amharic text classification. A two-step approach is used to classify Amharic documents. We first use EM clustering algorithm to group classes to clusters of the mixture document so that both labeled and unlabeled documents will be clustered to the predefined classes. The second step uses different text classifying algorithms to predict the documents to their predefined categories. The clustered data sets are used for training the text classifiers. Hence, the text classification model is developed from the training data sets using Naive Bayes (NB) and locally Weighted learner (LWL) classifying algorithms. Using the cross validation sampling method the model will be evaluated whether the unlabeled documents are correctly classified to their predefined classes or not.

4.1. The proposed System Architecture:

In the first stage the preprocessing makes the raw data ready for the experiment by removing irrelevant terms and stop words from the document. In the next stage, The EM technique applied to the case of labeled and unlabeled data with NB or LWL yields a straightforward and appealing algorithm. A schematic of this algorithm is shown in Figure 3. A NB (LWL) classifier is built in the standard supervised fashion from the limited amount of labeled training data. Then, we perform classification of the unlabeled data with the NB or LWL model.

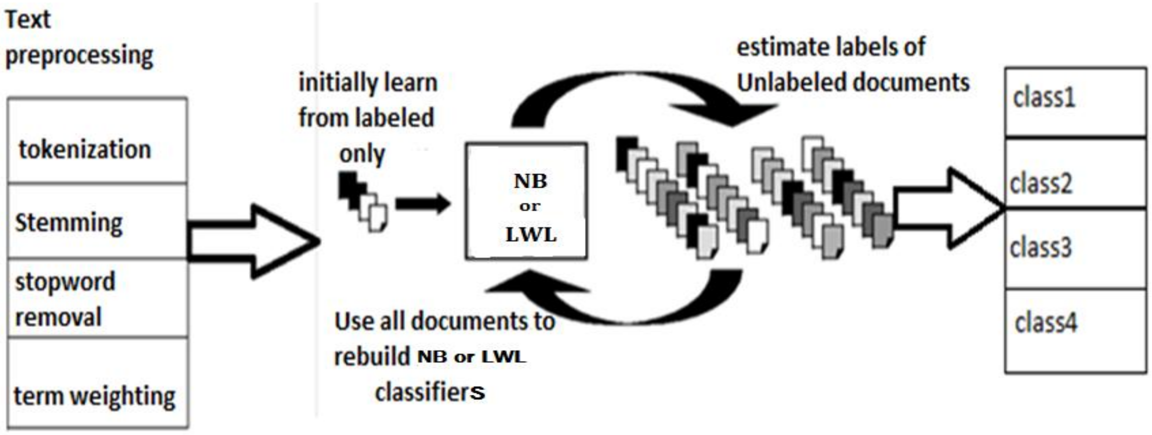
Text
preprocessing
tokenization
Stemming
stopword removal
term weighting
initially learn from labeled only
NB or LWL
estimate labels of Unlabeled documents
Use all documents to rebuild NB or LWL classifiers
class1
class2
class3
class4

Figure 3. Text classifier from labeled and unlabeled data using EM.

4.2. Data Set for Text classification:

The data source for the study was the Ethiopian News Agency (ENA). The total number of Amharic news items randomly collected for this experiment is 1,952. Out of which 300 items are labeled by the expertise in ENA and 1,652 of them are unlabeled.

4.3. Clustering:

An expectation-maximization (EM) algorithm is used in statistics for finding maximum likelihood estimates of parameters in probabilistic models, where the model depends on unobserved latent variables. . EM alternates between performing an expectation (E) step, which computes an expectation of the likelihood by including the latent variables as if they were observed, and maximization (M) step, which computes the maximum likelihood estimates of the parameters by maximizing the expected likelihood found on the E step. The parameters found on the M step are then used to begin another E step, and the process is repeated (Dempster et al., 1997). The EM algorithm is a popular class of iterative algorithms for maximum likelihood estimation for problems involving missing data. It is often used to fill the missing values in the data using existing values (Li Xiaoli et al. 2004).

4.4. Classification:

Text classification is the process of assigning predefined category labels to new documents based on the classifier learnt from training examples, in which document classifier is first trained using documents with reassigned labels or classes picked from a set of labels, which we call the taxonomy or catalog. Once the classifier is trained, it is offered test documents for which it must guess the best labels.

In this paper two classifiers namely NB and LWL are used for classification purpose. An experiment is conducted on both traditional (without unlabeled documents) and with unlabeled documents (EM) as tabulated below.

V. RESULTS

The results of the classifiers on 120 labeled and 1,702 unlabeled documents can be summarized below.

	Traditional				Expectation Maximization			
Instances	NB		LWL		NB		LWL	
Correctly classified	54	45%	37	30.8%	1089	59.7%	1380	75.7%
Incorrectly classified	66	55%	83	69.1%	733	40.3%	442	24.3
Relative Absolute error	81.3%		81.4%		59.6%		53.8%	

Table 1. Summary table of the classifiers.

Disadvantage of supervised approach with expectation maximization algorithm is that, first of all the data should be labeled and hence increases the efforts and time consuming while the unlabeled data is useless here. In case the document fails to lie in pre-defined classes the approach is not able to classify that document and hence leading to decrease in efficiency of classifier.

In semi supervised approach unlabeled data is not useless; it has been used to train the classifier and in case the document fails to lie in pre-defined classes it leads to dynamically generation of new class to categorize that document and the database is updated automatically. This can be proved from the result of our experiment given in table1. As we can understand from Table 1, EM decreases the absolute classification error for LWL (supervised) by 27.6% while for Naive Bayes by 21.7%.

On top of this, the LWL classifier with EM classifies 1,380 instances (75.7%) out of 1,822 correctly while NB-EM classifies 1,089 (59.7%) correctly. We can see that EM increases the accuracy of both supervised NB and LWL. On the other side, LWL with EM performs better than NB-EM for this Amharic text data. Hence the study uses LWL as the main classifier.

5.1 **Classification with and without unlabeled documents**

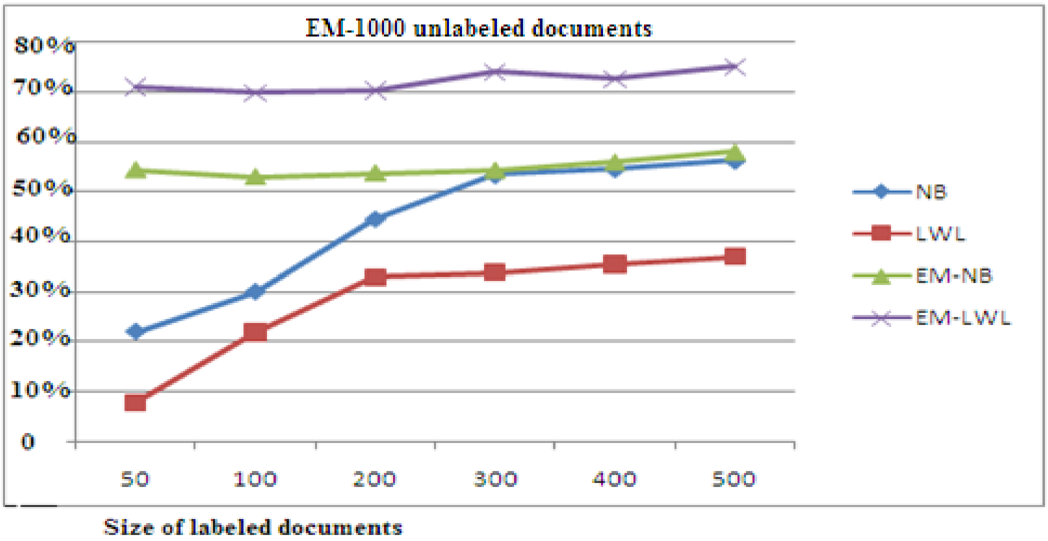

Figure 4. Classification accuracy with and with out 1,000 unlabeled documents

With small amounts of training data, using EM yields more accurate classifiers. With large amounts of labeled training data, accurate parameter estimates can be obtained without the use of unlabeled data, and the two methods begin to converge as can be seen in figure 4. Between NB and EM-NB. EM performs significantly better than traditional Naive Bayes and LWL as shown in figure 4. For example, with 200 labeled documents (20 documents per class), Naive Bayes reaches 44.5% accuracy while EM achieves 53.4%. For the same 200 labeled data, LWL reaches an accuracy of 33% while EM achieves 70.25%. Note that EM also performs well even with a very small number of labeled documents; with only 50 documents (five labeled document per class), Naive Bayes obtains an accuracy of 22%, while EM 53.8%. As expected, when there is a lot of labeled data, having unlabeled data does not help nearly as much, because there is already enough labeled data to accurately estimate the classifier parameters.

4.5. Testing Amharic Text Classification System:

Since the classes of the test documents are known, the researchers compare actual class label of each document with that of the predicted from the result shown in figure 4. as summarized in table 2.

Class name	Actual	Predicted	Accuracy
Adega	21	16	76%
Economy	11	7	63.6%
Bahlnaturism	8	4	50%
Heg	7	2	28.5%
Tena	14	10	71.2
Mahberawi	14	9	64.2%
Sport	14	10	71.2%
Temhrt	8	5	62.5%
Poletica	20	13	65%
Science	13	8	61.5%
Average Accuracy	**130**	**84**	**65%**

Table 2. Testing result of the Amharic text classification system

The current study predicts 84(65%) out of 130 instances as illustrated in table 2.

VI. CONCLUSION

In this paper, an attempt is made to design a system by which a collection of labeled and unlabeled documents could be categorized in a manner they are easily, quickly and systematically accessible to the users or news experts. This is important because in the 21st century there is a vast flowing of information across the globe. As a result, it deems necessary to classify growing information, particularly news text documents for the case of making them readily and quickly manageable as well as accessible to the experts. In this paper, experimented on how to propose an application of data mining techniques to Amharic text classification from labeled and unlabeled documents using EM.

One way to reduce the amount of labeled data required is to develop an algorithm that can learn from a small number of labeled examples augmented with a large number of unlabeled examples.The Web contains a huge amount of text data that can serve as unlabeled data for many classification tasks. Collecting this text is often cheap since web spiders and crawlers can be programmed to automatically do this task. Hence the need for classifiers that can learn from unlabeled examples is required. The second step uses two known classifiers: Naive Bayes and the lazy learner LWL to build the Amharic text classification system. As it is discussed on chapter five, the LWL classifier classifies 1,380 instances correctly (75.7%) decreasing the classification absolute error by 27.6% whereas the probabilistic classifier NB classifies 1,089 instances correctly (59.7%) decreasing the absolute error from the traditional by 21.7%. As a result, LWL with EM out performs better than NB in classifying Amharic text documents. Though, the performance of EM-NB is low in comparison to EM-LWL for 120 labeled and 1,702 unlabeled text documents, EM-NB shows good improvement in performance when the unlabeled document increases for a fixed size of labeled documents. As a result, for a few labeled and very large unlabeled documents we recommend to use EM-NB for classifying mixture text documents.A major concern with supervised learning techniques for text classification is that they often require a large number of labeled examples to learn accurately. Collecting a large number of labeled examples can be a very expensive process, thus emphasizing the need for algorithm that can provide accurate classifications after getting only a few labeled examples.

REFERENCES

[1] Addis, A. (2010*). Study and Development of Novel Techniques for Hierarchical TextCategorization*. Italy: University of Cagliari.

[2] Araki, J. (2003). *Text Classification with a polysemy Considered Feature Set*. MSc Thesis. University of Tokyo, Japan.

[3] Atelach Alemu and Lars Asker (2005). *Dictionary based Amharic French Information Retrieval.* Department of Computer and Systems Sciences, Stockholm University,Sweden.

[4] Atkeson, C.G., Moore, A.W., Schall, S. (1997). *Locally weighted learning*. Artificial Intelligence review, 11(1):11–73,

[5] Baharati, A. (2002). *A document Space Model for Automated Text Classification Based on Frequency Distribution across Categories*. Mumbai: ICON.

[6] Baker, D. and Kachites, A. (1998). *Distributional clustering of words for text Classification:ACM SIGIR*.pp.96-102.

[7] Beletu Reda (1982).*A Graphemic Analysis of the Writing System of Amharic*. Paper for the Requirement of the Degree of BA in Linguistics. Addis Ababa University, AddisAbaba.

[8] Bender, M. L. et al. (1976). *Language of Ethiopia*. London: Oxford University Press, 1976.

[9] Bilmes, Jeff A. (1998*). A gentle tutorial of the EM algorithm and its application toparameter estimation for Gaussian mixture and hidden Markov Models.* TechnicalReport tr-97-021.

[10] Blum, A., and Mitchell, T. (1998). *Combining labeled and unlabeled data with co-training.*COLT-98.

[11] Buckley, C., Hersh, W., Leone, T. and Hickarn, D. (1998). An interactive retrievalEvaluation and New Large Text Collection for Research. *Proceedings of 17th annualinternational Conference on research and Development in Information retrieval*. Dublin, Ireland:ACM: Springer. pp. 127-140.

[12] Cheng, C., Tang, T., FU, A. and King, I. (2001*). Hierarchical classification of documentswith error control.* Singapore: IEEE, 20(35), pp.10-19.

[13] Cherkassky, V. and Mulier, F. (1998). *Learning from Data: Concepts, Theory, and Methods*. Wiley-Interscience.

[14] Dempster, A. P., Laird, N. M., & Rubin, D. B. (1997). *Maximum likelihood from incompletedata via the EM algorithm.* Journal of the Royal Statistical Society, Series B, 39(1),1-38.

[15] Deng, Z., Tang, S., Yang, D.Zhang, M. and Wu, X. (2002).Two Odds-Radio based Text Classification Algorithms. *Proceedings of the third International Conference on Web Information Systems Engineering Workshop*. Singapore: IEEE, pp.20-25.

[16] Dhillon, I. (2003). *A Divisive Information Theoretic Feature Clustering algorithmfor textClassification*. Journal of Machine learning Research, 3(42), pp.1265-1287.

[17] Gebrehiwot Assefa (2010). *A two Step Approach: For Tigrigna Text Categorization*. MScThesis.Addis Ababa University, Addis Ababa, Ethiopia.

[18] Ghahramani, Z., & Jordan, M. I. (1994). *Supervised learning from incomplete data via anEM approach*. Advances in Neural Information Processing Systems 6, pp. 120-127.

[19] Girma Berhe (2001). *A stemming Algorithm Development for Tigrigna Language TextDocuments*. MSc Thesis. Addis Ababa University, Addis Ababa, Ethiopia.

[20] Han, J.andKamber, M.(2006). *Data Mining: Concepts and techniques (2nd ed.)*.MorganKaufmann Publishers.

[21] Kotsiantis, S. B. (2007). *Supervised machine learning*: A review of classification techniques,Informatica 31(3): 249–268.

[22] Lewis, D., and Ringutte, D. (1994). *A Comparison of Two learning algorithms for text Categorization.* Third Annual Symposium on Document Analysis and Information retrieval, pp.81-93.

[23] Li Xiaoli, Liu, B., Lee, W.S. and Philip S. Yu (2004). Text Classification by LabelingWords. *Proceedings of the national conference on AI.*

[24] McCallum, A., Nigam, K., Thrun, S. and Mitchell, T. (2000). *Text Classification fromlabeled and Unlabeled Documents using EM*. Boston: Kluwer Academic publisher,39(2), P.103- 134.

[25] Worku (2009). *Automatic Amharic News Classification using learning vector Quantization*.MSc Thesis. Addis Ababa University, Addis Ababa, Ethiopia.

[26] Yang, Y., Slattery S. and Ghani, R. (2001).A Study of Approaches to HypertextCategorization. *Proceedings of the fourteenth international Conference on Machine learning*, pp.412-420.

[27] Yang, Yiming, and Xin Liu (1999). *A re-examination of Text Categorization Methods*.Conference on Research and Development in Information Retrieval. pp 42-49.

[28] Yohannes Afework (2007). *Automatic Classification of Amharic News Text*. Msc Thesis.Addis Ababa University, Addis Ababa, Ethiopia.

[29] Yu, P., Lee, W., Li, X., Liu, B. (2003).*Text Classification by Labeling Words*. Department of Computer Science. University of Illinois, Chicago.

[30] Zelalem Sintayehu (2001). *Automatic Classification of Amharic News Items: The Case of Ethiopian News Agency*. Master's Thesis. Addis Ababa University, Addis Ababa.

AUTHOR INFORMATION

Seffi Gebeyehu received his B.Sc. Degree in Computer Science from Bahir Dar University. He received his M.Sc. Degree in Information Technology from Adama Science and Technology University, Adama. Currently working as Lecturer in School of Computing and Electrical Engineering, IOT, Bahir Dar University, Ethiopia.His main research interest is Data Mining, Mobile Networking.

Dr. Vuda Sreenivasa Rao received his M.Tech degree in computer science and engineering from Sathyabama University from 2007.He received PhD degree in computer science and engineering from Singhania University, Rajasthan, India from 2010. Currently working as Professor in School of Computing and Electrical Engineering, IOT, Bahir Dar University, Ethiopia. His main research interests are Data mining, Fuzzy logic, Mobile communication, cloud computing and Network Security. He has got 14 years of teaching experience. He has published 38 research papers in various international journals and one Springer international conference paper. He has Editor-in-Chief in 3 international journals and 129 Editorial Board / Reviewer Board memberships in various international journals. He has Technical committee member in various international Conferences. He is a life member of various professional societies like IEEE, ACM, MAIRCC, MCSI, SMIACSIT, MIAENG, MCSTA, MAPSMS, MSDIWC, SMSCIEI, SNMUACEE and MISTE.