Bangla News Classification using Naive Bayes classifier

Abu Nowshed Chy
Dept. of Computer
Science & Engineering
University of Chittagong
Chittagong - 4331, Bangladesh
Email: nowshed@skeim.org

Md. Hanif Seddiqui
Dept. of Computer Science & Engineering
Bangabandhu Sheikh Mujibur Rahman
Science and Technology University, Gopalganj-8100
and University of Chittagong, Chittagong-4331
Email: hanif@cu.ac.bd

Sowmitra Das
Dept. of Computer
Science & Engineering
University of Chittagong
Chittagong-4331, Bangladesh
Email: sowmitra@skeim.org

***Abstract*—Web is gigantic and being constantly update. Bangla news in web are rapidly grown in the era of information age where each news site has its own different layout and categorization for grouping news. These heterogeneity of layout and categorization can not always satisfy individual user's need. Removing these heterogeneity and classifying the news articles according to user preference is a formidable task. In this paper, we propose an approach that provides a user to find out news articles which are related to a specific classification. We use our own developed web crawler to extract useful text from HTML pages of news article contents to construct a Full-Text-RSS. Each news article contents is tokenized with a modified light-weight Bangla Stemmer. In order to achieve better classification result, we remove the less significant words i.e. $stop-word$ from the document. We apply the naive Bayes classifier for classification of Bangla news article contents based on news code of IPTC. Our experimental result shows the effectiveness of our classification system.**

I. Introduction

The Web is rapidly moving towards a platform for mass collaboration in content production and consumption, and the increasing number of people are turning to online source for daily news. Bangla online newspaper started appearing at about the same time that the internet became public in Bangladesh. Every day, vast amount of Bangla news articles are created by the several news sites that exist in the World-Wide-Web and its rate increases exponentially. An online newspaper has many forms [1]. One form is electronic edition of the printed newspaper. The user can read the online edition similar to a paper edition; there is no categorization, neither with respect to content nor with respect to layout. Another form of online newspaper is news website, which enables the user browsing in menus that are organized in subject categories and sub-categories. Common to most of the above forms of online newspapers is that the user is assumed to read the news from a computer screen, while connected via the Internet to a certain news provider. However, these services might not be sufficient for many readers and reading situations.

A lot of newspaper reader like to view and analyze news from various news sources. Many times, readers are interested only in the news articles of their categories of interest [2]. Therefore,the users have to scan through all the news articles of several news sites in order to get the articles of his/her interest. For example, a user interested in sports related news has to go through all the news articles from various news sites and their time spent in analyzing news from the multiple sources which is tedious. So a reader would prefer a system that would gather news articles from various sources; which is accessible to all the time and anywhere, also from a mobile reading device. A reader would like to visit a unique newspaper that includes the articles from various favorite sources, arranged and presented in an order that best fits her interests and reading habits.

Moreover, International Press Telecommunication Council (IPTC) publishes as standard of news classification, called as news code taxonomies. The IPTC creates and maintains sets of concepts, called a controlled vocabulary or a taxonomy to be assigned as metadata values to news objects like text, photographs, graphics, audio and video files and streams. This allows for a consistent coding of news metadata across news providers and over the course of time [1].

We hardly find the IPTC standard classification at most of the online Bangla news media. Therefore, our attempt is to classify the online news article contents to satisfy users query.

Web pages have their own underlying embedded structure in the HTML language [3]. They typically contain noisy content such as advertisement banner and navigation bar. If a pure-text classification method is directly applied to these pages, it will incur much bias for the classification algorithm, making it possible to lose focus on the main topics and important content. Thus, a critical issue is to design an intelligent preprocessing technique to extract individual news document from a Web page. Our web crawler crawl each news site and extract the individual news document of specific date from that news site as a Full-Text-RSS file. The indidividual document is then tokenized with a light-weight bangla stemmer and the less significant words, which are called stop-words, are removed. Finally, classification is performed to each news item based on IPTC news code taxonomy. The basic news classification problem can be formulated as follows:

Given a training set of documents $D_{train} = (d_1, l_1)....(d_n, l_n)$ of labeled news documents where each document d_i belongs to a document set D and the label $l_i = l_i(d_i)$ of d_i is within a manually predefined set of categories $C = c_1, ..., c_m$ of IPTC news code taxonomy. The goal in news document categorization is to devise a learning algorithm that given the training set D_{train} as input will generate a classifier $h : D- > C$ that will be able to

[1] http://www.iptc.org/site/NewsCodes/

978-1-4799-3497-3/13/$31.00 ©2013 IEEE

accurately classify unseen documents from D [4].

The rest of this paper is structured as follows: **Section II** describes the related work while the foundamental concepts of news code of IPTC is articulated in **Section III**. We introduce our approach in **Section IV**. **Section V** includes experiments and evaluation to show the effectiveness of our proposed approach. Concluded remarks and some future directions of our work is described in **Section VI**.

II. Related Work

The task of news classification is to automatically classify news documents into predefined classes based on their content. A number of statistical and machine learning techniques has been developed for news classification in different language. In [5], KNN classification algorithm is used to predict the test samples category according to the K training samples which are the nearest neighbors to the test sample and judge it to that category which has the largest category probability. [6] propose Weight Adjusted k-Nearest Neighbor (WAKNN) classification algorithm that is based on the k-NN classification paradigm. In WAKNN, the importance of each word in the classification of a training document set is learned and the weight vector reflecting this importance is maintained. The weight vector is used in the similarity measure computation such that important words contribute more in the similarity measure.

Naive Bayes (NaiveBayes) probabilistic classifier is often used in text classification applications and experiments because of its simplicity and effectiveness which basic idea is to use the joint probabilities of words and categories to estimate the probabilities of categories given a document [7]. The naive part of such a model is the assumption of word independence. However, its performance is often degraded because it does not model text well, and by inappropriate feature selection and the lack of reliable confidence scores. In [8], Karl-Michael Schneider address these problems and show that they can be solved by some simple corrections. The paper [9] shows that the accuracy of a NaiveBayes classifier can be significantly improved by taking advantage of a hierarchy of classes. They adopt an established statistical technique called shrinkage that smoothes parameter estimates of a data-sparse child with its parent in order to obtain more robust parameter estimates and deleted interpolation that smoothes n-grams in language modeling.

In text classification, Support Vector Machine (SVM) classifier transforms the original data in a higher dimension, from where it can find a hyperplane for separation of the data using essential training instances called support vectors and thus classify the document [10]. But SVM classifier provide excellent precision with poor recall. In paper [11] James G. Shanahan and Norbert Roma describe an automatic process for adjusting the thresholds of generic SVM to improve recall. [12] used n-gram (n character slice of a longer string)based classifier algorithm for Bangla text categorization.

An n-gram is a subsequence of n-items in any given sequence. In computational linguistics n-gram models are used most commonly in predicting words (in word level n-gram) or predicting characters (in character level n-gram). Typically, n is fixed for a particular corpus of documents and the queries made against that corpus where corpus is a huge text. In n-gram based text classification, Johannes Furnkranz [13] shows that after the removal of stop words, word sequences of length 2 or 3 are most useful.

Using longer sequences reduces classification performance. [14] used Linear Least Squares Fit (LLSF) classifier that use a large collection of human-assigned matches between texts and categories as a ”training set”, to compute word-to-category connections that enables us to weight these connections in such a way as to optimally fit the training set and to probabilistically capture the correct categories for arbitrary texts. [2] used Vector-Space-Model classifier which basic idea is to construct a prototype vector per category using a training set of documents. Given a category, the vectors of documents belonging to this category are given a positive weight, and the vectors of remaining documents are given a negative weight. By summing up these positively and negatively weighted vectors, the prototype vector of this category is obtained. Rodriguez et. al. described an approach that integrates WordNet with two training approaches (Rocchio - relevance feedback and Widrow Hoff - machine learning) through the Vector Space Model which improve their performance [15].

Decision tree (DTree) algorithms select informative words based on an information gain criterion, and predict categories of each document according to the occurrence of word combinations in the document [16]. [17] described a fast decision tree induction algorithm especially suited to text data and a rule simplification method that converts the decision tree into a logically equivalent rule set.

Corner classification (CC) network is a kind of feed forward neural network for instantly document classification. To classify text object instantly, new training algorithm-TextCC, for feed forward neural network is presented in [18]. since textural data has a very high-dimension feature space text classification model using an artificial neural network as the text classifier scalability is poor if the neural network is trained using the raw feature space.

Another paper propsed four dimensionality reduction techniques and show that Principal Component Analysis (PCA) was found to be the most effective in reducing the dimensionality of the feature space [19].

Some researchers comparing the performance of two or more indidividual classifier on the same data set to show that which classifier perform well on what kind of data set [4] [20]. In the context of combining multiple classifiers for text clssification, a number of researchers have shown that combining multiple classifiers can improve classification accuracy [21] [22]. An Evaluation of number of Statistical Approaches to Text Categorization is given in [23].

III. News Code of IPTC

To promote the ease of interchange of news items, The International Press Telecommunication Council (IPTC), an international organization that is primarily focused on developing and publishing Industry Standards for the interchange of news data, has provided numerous categorization schemes aimed to standardize coding of various aspects of news related metadata. The whole product is called NewsCodes. A NewsCode is a

978-1-4799-3497-3/13/$31.00 ©2013 IEEE

single code representing a concept which is used to categorize news content. Many of these codes can make a set for a specific use, such a set is branded as NewsCodes by the IPTC.

Typically NewsCodes are language agnostic and as each code has an explicit and comprehensive definition so one can easily share not only the codes but also their semantics. More generic terms used are vocavulary, taxonomy, topicset and so on. Vocabulary is a set of codes. Can be either controlled or uncontrolled. Controlled Vocabulary managed by some authority (e.g. a person or an organisation), employing some mechanism (e.g. an XML Schema, IPTC G2 KnowledgeItem) to maintain this set. Each code in a controlled vocabulary represents a concept. Concept means anything that one may wish to refer to and may be represented by one or more codes.

Collection of concept with associated code is called taxonomy.A taxonomy may support typed relationships between concepts. Such a taxonomy is sometimes known as an ontology or thesaurus. The IPTC groups the taxonomies into four main areas: **Descriptive NewsCodes** - is a group of taxonomies to describe the content of news items properly. **Administrative NewsCodes** - is a group of taxonomies for proper administration of news items. **Transmission NewsCodes** is a group of taxonomies with controlled values for the transmission process. **Exchange Format NewsCodes** - is a group of taxonomies with values to support specific functionalities of the different IPTC news exchange format standards.

IV. Our Approach

The goal of news classification is to assign categories to a news document according to the content of the document. In this regard, we perform a number of steps as follows:

1. A number of online news sites are selected based on type and popularity.
2. By using our own RSS crawler we crawl each sites and generate Full-Text-RSS that contains title, link, description and date.
3. Then we use a RSS parser to extract individual news document from the RSS feed.
4. After getting idividual document we process the documents text by several techniques such as tokenization, punctuation removal, digit removal and so on.
5. We build a light bengali lexicon i.e. begali dictionary that help us in stemming.
6. We establish a light bengali stemmer to identify the stem of each word.
7. We then perform some advanced text processing task i.e, single letter words removal, stop word removal.
8. Finally, we consider each document as a word-vector and supervised learning is conducted to assign multiple relevant classes.
9. Results of the supervised learning task are applied to new documents and we categorize the documents based on Naive-Bayes classifier.

The overview of our approach is depicted in Fig. 1.

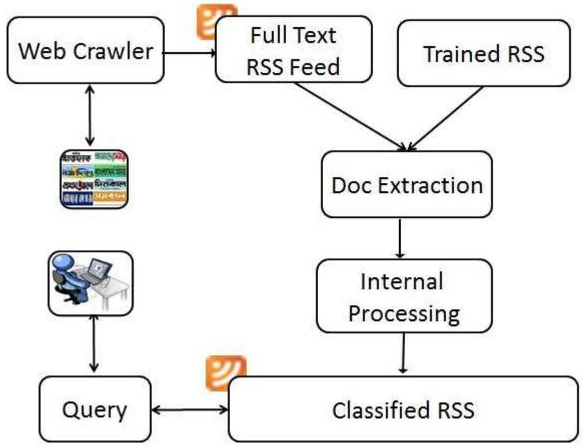

Fig. 1. Our News Classification System Overview

A. Construction of Full-Text-RSS

To create Full-Text-RSS, we extract the Web news article contents from news sites. Usually, the extraction process includes two steps. Firstly, the news sites are crawled on Breadth-First-Search method to collect the news pages. Secondly, the news article contents are extracted from news pages and generate RSS file. As the news sites comprise different kinds of Web pages, there are many non-news pages such as the advertisement, related stories and comments. In order to recognize and extract the parts of news article contents from the full text of news pages, wrappers are generated based on the analysis of layout of news pages by many extraction methods. Web page layout is the style of graphic design in which text or pictures are set out on a Web page. The different news sites use the different news page layout, and each news site uses more than one layout. Inspite of the heterogeneous layouts we analyze the news pages to detect an actual news url and the title, news-body and publication date is extracted. We observe that a page containing smaller number of links is likely to be a detail of news article content. Once we find an actual news url, extracting title, news-body and the publication date is relatively easier to extract.

B. News Data Preprocessing

An RSS parser extracts individual news document from the Full-Text-RSS feed. After the extraction of news document the preprocessing step is initiated with tokenization.

1) Tokenization: Tokenization is the process that aims to fracture the stream of characters into tokens delimited by white space, tab, new line and so on. As we are working on Bengali news document, there are lots of Bengali punctuations in the document. Moreover news document may contain Bengali as well as English digits. As meaningful Bengali words do not contain these characters we remove these. We also eliminte the single letter word from the document in this stage. Then the documents are nothing but a bag-of-words. As Bengali is a highly inflected language with relatively free or pragmatically free word order i.e., Bengali (verb, noun, adjective) words are inflected forms of roots we perform stemming on the tokenized words.

978-1-4799-3497-3/13/$31.00 ©2013 IEEE

```
Algo. bangla_stemmer (word)

for each $word \in document$
   $dictionary\_checkers(word)$
   if $word \in dictionary$
      $stem{=}word$
   else $stem1{=}stemming(word)$
      if $stem1 \in dictionary$
         $stem{=}stem1$
      else $stem{=}stemming(stem1)$
```

Fig. 2. Pseudo code of how our bengali stemmer works.

2) Stemming: Stemming is an operation that splits a word into the constituent root part and affix without doing complete morphological analysis. Terms with common stems tend to have similar meaning, which makes stemming an attractive option to increase the performance of spelling checkers and other information retrieval applications where morphological analysis would be too computationally expensive. Another advantage of stemming is that it can drastically reduce the dictionary sized used in various NLP applications, especially for highly inflected languages [24].

The simplicity of languages like English, and other English-Like Languages has enabled the development of stemmers for them. Moreover, these languages being web-resource enriched, make the work of the researchers easy. However, the languages which are highly inflectional and resource poor at the same time have always been treated with negligence, especially Bangla. Though Bangla is the fourth most widely spoken language in the world. In the written form of Bangla there are 11 vowels and 39 consonants. Moreover, there are 10 short forms of vowels called vowel modifiers (i.e. Kar), 7 short forms of consonants called consonant modifiers (i.e.Fala). The last two decades has witnessed an immense escalation of Bangla web and digital text contents and is having an exponential growth rate. This has enhanced the need for the development of highly efficient Information Retrieval (IR) systems and consequently good stemmers [25].

By considering the above facts we design a Bangla stemmer to improve the performance of our system. Moreover we create a light Bangla dictionary that help us stemming a word. To achieve this we write a parsing program that extract Bangla word from a online bilingual dictionary. There are much inflected word on this dictionary so we remove these inflection from the dictionary. This dictionary contains about 15000 words that frequently used in online Bangla news. The dictionary plays an important role while stemming a word. The algorithm in Fig. 2 demonstrate how our Bangla stemmer works with the help of light Bangla dictionary.

3) Stop Word Removal: Statistical analysis through documents shows that some words have quite low frequency, while some others act just the opposite. The characteristic of these common words is that they carry almost no significant information to the document. Instead, they are used just because of grammar. We usually refer to this set of words as stop words [26].

For the purpose of this research we define a general stop word list. These non-significant words represent noise, and may actually damage the classification performance because they do not discriminate between relevant and non-relevant documents. Secondly, we expect to reduce the size of the considerable vector elements.

In establishing a general stop word list for Bangla, we follow several steps. At first, with the help of our RSS crawler we collect 3142 distinct news document as RSS file from a popular online Bangla newspaper prothom-alo.com. Each document is then tokenized and performed stemming. Finally, a simple feature selection method, "inverse document frequency (IDF)", is applied to determine stop-words, the less significant words in identifying documents. We sorted all the word according to their IDF value in ascending order and we consider the large stop word list (300 most frequently words) as a trend in information retrieval system over time. We inspect our system by considering other values of stop word list. We consider the first 300 words as stop words. Moreover, we included some other widely used non-information-bearing words in the stop word list manually.

C. Classification Algorithms

In Bangla news classification, we use the Naive Bayes Classifier.

1) Naive Bayes classifier: Naive Bayes (NB) probabilistic classifier algorithm was first proposed and used for text categorization task by D.Lewis [7]. NB is based on the bayes theorem in the probabilistic frame work. The basic idea is to use the joint probabilities of words and categories to estimate the probabilities of categories given a document. The naive part of such a model is the assumption of word independence. The simplicity of the assumption of word independence makes the computation of the Naive Bayes classifier far more efficient than the exponential complexity of non-naive Bayes approaches because it does not use word combinations as predictors.

When the NB classifier is applied in Text Classification problem we use the following equation:

$$p(class \mid doc.) = \frac{p(class).p(doc. \mid class)}{p(doc.)} \tag{1}$$

where $p(class \mid doc.)$ is the probability that a given document D belongs to a given class C. $P(doc.)$ is the probability of a document, we can notice that $p(doc.)$ is a Constance divider to every calculation, so we can ignore it. $P(class)$ is the probability of a class (or category), we can compute it from the number of documents in the category divided by documents number in all categories. $p(doc. \mid class)$ represents the probability of document given class and documents can be modelled as sets of words, thus the $p(doc. \mid class)$ can be written like:

$$p(doc. \mid class) = \prod_i p(word_i \mid class) \tag{2}$$

So, we can rewrite the equation(2) as:

$$p(class|doc.) = p(class).\prod_i p(word_i|class) \tag{3}$$

Where $P(word_i \mid class)$ is the probability that the i-th word of a given document occurs in a document from class, When the size of the training set is small, the relative frequency estimates of probabilities, $p(word_i \mid class)$, will not be reasonable; if a word never appears in the given training data, its relative frequency estimate will be zero. Instead, we applied the Laplace law of succession to estimate $p(word_i \mid class)$. The estimate of the probability $p(word_i \mid class)$ is therefore written as:

$$p(word_i \mid class) = \frac{T_c t + \lambda}{N_c + \lambda V} \tag{4}$$

Where $T_c t$ is the number of times the word occurs in that category C. N_c is the number of words in category C. V is the vocabulary size. λ is the positive constant, usually 1, or 0.5 to avoid zero probability [27] [28].

V. Experiments and Evaluation

in this section, we present the evaluation of our aproach and explain the implementation of our system

A. Evaluation Metrics

To evaluate the systems classification accuracy, i.e. the proportion of correctly classified news document, we used common Information Retrieval (IR) performance measures, i.e. precision and recall to evaluate the system.

Suppose that the number of the documents which are in C_j category in fact and also the classifier judge them to C_j category is a; the number of the documents which are not in C_j category in fact, however the classifier judge them to C_j category is b; the number of the documents which are in C_j category in fact, however the classifier do not judge them to C_j category is c; the number of the documents which are not in C_j category in fact and also the classifier do not judge them to C_j category is d [5]. That is, once we get contingency table, we can define precision and recall as follows:

Precision, P: Precision is the ratio of the number of documents which judge correctly by classifiers to the number of documents which classifiers judged to this category, so the precision of C_j is defined as following:

$$Precision, P = \frac{a}{a+b} \tag{5}$$

Recall, R: Recall rate is the ratio of the number of documents which judge correctly by classifiers to the number of documents which are this category in fact, so the recall rate of C_j is defined as following:

$$Recall, R = \frac{a}{a+c} \tag{6}$$

B. Quantitative Analysis

In the taxonomy of IPTC, there are several hundreds of concepts for news subject codes. However, we consider only the upper level 34 concepts such as 'religion and belief', 'social problem', 'agriculture', 'crime, law and justice', 'sport event', 'politics', 'election', 'government', 'transport accident', 'environmental pollution', 'natural disaster', 'health' and so on. We have more than seven thousand news topics extracted from the Daily Prothom Alo. However, six hundred twenty eight (628) news items are manually tagged with multiple classification codes.

Our approach starts learning with 208 news items to classify the other 416 news items. We evaluate our approach with the TREC [29] evaluation technique to produce Recall-Precision graph depicted in Fig. 3.

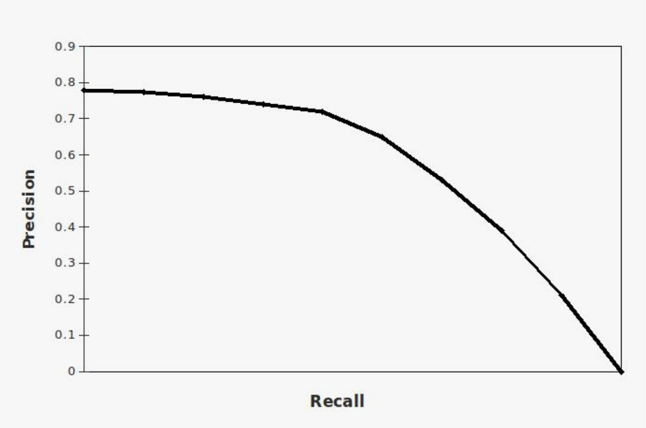

Fig. 3. Recall-Precision Graph of Our Bangla News Classification Approach

We are still on the way of evaluating news document classification with more lower level classification concepts those have some sort of overlapping characteristics.

VI. Conclusion and Future Direction

The growing domain of online newspaper presents a rich area, which can benefit immensely from automatic classification approach. In this paper we present a system of automatically classifying Bangla News documents. This system provides users with efficient and reliable access to classified news from different sources. It achieves a high accuracy of classifications with the possibility that one story be classified into more than one category. We used the Naive Bayes algorithm which is based on probabilistic framework to handle our classification problem. Bangla stemmer and stop word removal play an important role to achieve higher accuracy in Bangla news document classification.

Our future target is to consider the full set of IPTC news subject code for classification and to compare with other techniques such as KNN,SVM on Bangla news document classification) on our system.

References

[1] L. Tenenboim, B. Shapira, and P. Shoval, "Ontology-based classification of news in an electronic newspaper," 2008.

[2] B. Pendharkar, P. Ambekar, P. Godbole, S. Joshi, and S. Abhyankar, "Topic categorization of rss news feeds," *Group*, vol. 4, p. 1, 2007.

[3] D. Shen, Z. Chen, Q. Yang, H. Zeng, B. Zhang, Y. Lu, and W. Ma, "Web-page classification through summarization," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 2004, pp. 242–249.

[4] A. Kumar and M. Jain, "Text classification of news articles."

[5] Y. Zhou, Y. Li, and S. Xia, "An improved knn text classification algorithm based on clustering," *Journal of Computers*, vol. 4, no. 3, pp. 230–237, 2009.

[6] E. Han, G. Karypis, and V. Kumar, "Text categorization using weight adjusted k-nearest neighbor classification," *Advances in Knowledge Discovery and Data Mining*, pp. 53–65, 2001.

[7] D. Lewis, "Naive (bayes) at forty: The independence assumption in information retrieval," *Machine Learning: ECML-98*, pp. 4–15, 1998.

[8] K. Schneider, "Techniques for improving the performance of naive bayes for text classification," *Computational Linguistics and Intelligent Text Processing*, pp. 682–693, 2005.

[9] A. McCallum, R. Rosenfeld, T. Mitchell, and A. Ng, "Improving text classification by shrinkage in a hierarchy of classes," in *Proceedings of the Fifteenth International Conference on Machine Learning*, 1998, pp. 359–367.

[10] C. Ee and P. Lim, "Automated online news classification with personalization."

[11] J. Shanahan and N. Roma, "Improving svm text classification performance through threshold adjustment," *Machine Learning: ECML 2003*, pp. 361–372, 2003.

[12] M. Mansur, N. UzZaman, and M. Khan, "Analysis of n-gram based text categorization for bangla in a newspaper corpus," *Center for Research on Bangla Language Processing, BRAC University, Dhaka, Bangladesh*, 2005.

[13] J. Fürnkranz, "A study using n-gram features for text categorization," *Austrian Research Institute for Artifical Intelligence*, 1998.

[14] Y. Yang and C. Chute, "An example-based mapping method for text categorization and retrieval," *ACM Transactions on Information Systems (TOIS)*, vol. 12, no. 3, pp. 252–277, 1994.

[15] M. Rodriguez, J. Hidalgo, and B. Agudo, "Using wordnet to complement training information in text categorization," in *Proceedings of 2nd International Conference on Recent Advances in Natural Language Processing II: Selected Papers from RANLP*, vol. 97, 2000, pp. 353–364.

[16] D. Lewis and M. Ringuette, "A comparison of two learning algorithms for text categorization," in *Third annual symposium on document analysis and information retrieval*, vol. 33, 1994, pp. 81–93.

[17] D. Johnson, F. Oles, T. Zhang, and T. Goetz, "A decision-tree-based symbolic rule induction system for text categorization," *IBM Systems Journal*, vol. 41, no. 3, pp. 428–437, 2002.

[18] Z. Zhang, S. Zhang, E. Chen, X. Wang, and H. Cheng, "Textcc: new feed forward neural network for classifying documents instantly," *Advances in Neural Networks–ISNN 2005*, pp. 811–811, 2005.

[19] S. Lam and D. Lee, "Feature reduction for neural network based text categorization," in *Database Systems for Advanced Applications, 1999. Proceedings., 6th International Conference on*. IEEE, 1999, pp. 195–202.

[20] A. Babu and P. Kumar, "Comparing neural network approach with n-gram approach for text categorization," *International Journal on Computer Science and Engineering*, vol. 2, no. 1, pp. 80–83, 2010.

[21] Y. Bao and N. Ishii, "Combining multiple k-nearest neighbor classifiers for text classification by reducts," in *Discovery Science*. Springer, 2002, pp. 313–348.

[22] S. Cho and J. Lee, "Learning neural network ensemble for practical text classification," *Intelligent Data Engineering and Automated Learning*, pp. 1032–1036, 2003.

[23] Y. Yang, "An evaluation of statistical approaches to text categorization," *Information retrieval*, vol. 1, no. 1, pp. 69–90, 1999.

[24] M. Islam, M. Uddin, M. Khan *et al.*, "A light weight stemmer for bengali and its use in spelling checker," 2007.

[25] B. Das, "Development of bengali language stemmer," Ph.D. dissertation, Indian Institute of Technology.

[26] F. Zou, F. Wang, X. Deng, and S. Han, "Automatic identification of chinese stop words," *Research on Computing Science*, vol. 18, pp. 151–162, 2006.

[27] F. Thabtah, M. Eljinini, M. Zamzeer, and W. Hadi, "Naïve bayesian based on chi square to categorize arabic data," in *proceedings of The 11th International Business Information Management Association Conference (IBIMA) Conference on Innovation and Knowledge Management in Twin Track Economies, Cairo, Egypt*, 2009, pp. 4–6.

[28] E. Frank and R. Bouckaert, "Naive bayes for text classification with unbalanced classes," *Knowledge Discovery in Databases: PKDD 2006*, pp. 503–510, 2006.

[29] E. Voorhees and D. Harman, *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press Cambridge, 2005, vol. 63.

978-1-4799-3497-3/13/$31.00 ©2013 IEEE