A New Text Mining Approach Based on HMM-SVM for Web News Classification

Krishnalal G
Senior Lecturer, CSE
SJCET
Palai, Kottayam, Kerala, India

S Babu Rengarajan
Professor & Head of IT
PET Engineering College, Vallioor,
Tamil Nadu, India.

K G Srinivasagan
Associate Professor, CSE
National Engineering College
Kovilpatti, Tamil Nadu, India

ABSTRACT

Since the emergence of WWW, it is essential to handle a very large amount of electronic data of which the majority is in the form of text. This scenario can be effectively handled by various Data Mining techniques. This paper proposes an intelligent system for online news classification based on Hidden Markov Model (HMM) and Support Vector Machine (SVM). An intelligent system is designed to extract the keywords from the online news paper content and classify it according to the pre defined categories. Three different stages are designed to classify the content of online newspapers such as (1) Text pre-processing (2) HMM based Feature Extraction and (3) Classification using SVM. Data have been collected for experimentation from The Hindu, The New Indian Express, Times of India, Business Line, and The Economic Times. The experimental results are based on the news categories such as sports, finance and politics and their accuracies in percentage are 92.45, 96.34 and 90.76 respectively. These results are very good compared to that of other text classification methods.

Keywords

Feature Extraction, HMM, kNN, POS, SVM.

1. INTRODUCTION

News articles on topical issues are helpful for company managers and other decision-makers. However, due to the sheer number of news articles published, it is a time-consuming task to select the most interesting one. Therefore, a method of news-article categorization is essential to obtain the relevant information quickly.

In order to develop such a text-classification system, many researchers have devoted their work for automating the text classification task. A news-story categorization system is developed, where a rule base is generated by human expertise. Building such a system requires huge efforts from the indexing experts, taking more than a couple of months because of the enormous number of rules.

On the other hand, a statistical approach based on keyword extraction from training texts is a popular method of generating a knowledge base [3]. This has the advantage that the knowledge base can be quickly generated without much cost. However, we need guidelines on how to gather a large quantity of training texts. As for automated text categorization, our results are interesting. It shows that relatively knowledge-poor machine learning algorithm outperforms human beings in a text classification task. This suggests that automated text categorization techniques are reaching a level of performance at which they can compete with humans not only in terms of cost-effectiveness and speed, but also in terms of accuracy of classification.

The rest of the paper is organized as follows: section 2 describes related works, section 3 provides the basics of text classification procedure, section 4 presents the proposed Multi class classifier based on HMM-SVM, and section 5 displays the experimental results followed by concluding remarks.

2. RELATED WORKS

Before going into the new intelligent classification system, it is essential to have an overview of the various existing methodologies.

2.1 Manual and Fuzzy Text Classification

Much useful information is in the form of text: This ranges from emails, web pages, newspaper articles, market research reports, through to CVs, complaint letters from customers, and internally generated reports [18]. As far as the online news papers is concerned, the system is supposed to provide news under various categories like national, international, regional, politics, finance, sports, entertainment etc.

In earlier days of online newspapers, the classification and indexing of news is done manually. Classifying and indexing news reports by hand were found to be an expensive, slow and labor-intensive activity.

Consistent accuracy was difficult to obtain with human indexers, and the work tended to cause high staff turnover. With these issues in mind, Carnegie Group, based in Pittsburgh, worked with Reuters [27] to develop the Construe system, an automated news categorization system based on a fuzzy rule-based text categorization.

2.2 Automated Text Categorization

In the last 15 years or so, substantial research has been conducted on text classification through supervised machine learning techniques [13], [15]. The vast majority of studies in this area focus on classification by topic, where bag-of-content-word models turn out to be very effective. Recently, there has also been increasing interest in automated categorization by overall

sentiment, degree of subjectivity, authorship and along other dimensions that can be grouped together under the cover terms of "genre" and "style" [4].

Genre and style classification tasks cannot be tackled using only the simple lexical cues that have been proven so effective in topic detection. For instance, an objective report and a subjective editorial about the Iraq war will probably share many of the same content words; *vice versa*, objective reports about Iraq and soccer will share very few interesting content words. Thus, categorization by genre and style must rely on more abstract topic-independent features. Popular choices of features [12] [14] have been function words (that are usually discarded or down-weighted in topic based categorization), textual statistics (e.g., average sentence length, lexical richness measures), n-grams and part-of-speech (POS) information. There are some other techniques like k-Nearest Neighbour (kNN) for text classification. But when considering the accuracy factor of classification, these algorithms are not enough to be used for news classification in domains such as online newspapers or online journals since the accuracy is having prime importance in these scenarios.

3. TEXT CLASSIFICATION PROCEDURE

The following steps are essential for the successful development of an intelligent classification system.

Goal definition: The goals of the classification system should adhere to the following requirements:

- *a.* To be a fully automated system or a classification supporting system.
- *b.* A number of categories are to be assigned to each incoming text (one or plural).
- *c.* Classification correctness.
- *d.* Computing time needs be limited when generating a knowledge base and when classifying each text.

Category definition: The categories to be classified should be defined beforehand. Ideally each category should be exclusive to each other in the keyword distribution.

Text collection: A large number of texts must be prepared, some of which are used for training the system, while others are used for evaluation purpose. Each text should be assigned to the relevant categories beforehand.

Statistical analysis of texts: Statistical analysis of the texts is very important, for example text length, the number of texts in each category, and keyword distribution in each text or in all texts.

Knowledge base generation: A knowledge base is generated using training texts. Weighted keywords [30], which define the character of each category, are extracted and stored in the knowledge base.

Evaluation of classification results: One or more categories are assigned to each text. Then, the "classification correctness" ("recall" and "precision") is calculated [7, 8].

Results analysis and knowledge base refinement: If the degree of correctness is not efficient enough to achieve the system's goals, then the knowledge base will require refinement.

4. PROPOSED METHODOLOGY

The proposed intelligent news classifier is designed and developed for extracting the full potential of HMM and SVM methods. The general architecture of the proposed system is as shown in Figure 1.

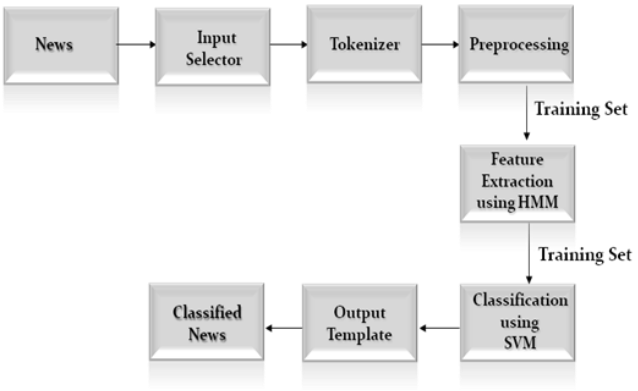

Fig.1. General Architecture

4.1 Pre-Processing

News articles in the form of text files are fed into the system using JFileChooser API. It then undergoes Parsing. Parsing is carried out using the Java StringTokenizer. Parsing is the division of text into a set of discrete parts, or tokens, which in a certain sequence can convey a semantic meaning [29]. The String Tokenizer class provides the first step in this parsing process, often called the lexer (lexical analyzer) or scanner. StringTokenizer implements the Enumeration interface. Therefore, given an input string, it can enumerate the individual tokens contained in it.

To use String Tokenizer, specify an input string that contains delimiters. Delimiters are characters that separate tokens. For example, ", ; :" sets the delimiters to a comma, semicolon, and colon. The default set of delimiters consists of the whitespace characters: space, tab, newline, and carriage return.

The words obtained from the tokenizer form the basis for the feature space of the training data. However, a fair amount of pre-processing not only prunes down the training size, but also makes the data more 'clean' and capable of training the classifier more effectively [22].

The conversion of the entire text to lower-case and removal of non-alphanumeric contents are followed by Stop-word elimination and grammatical stemming. Many text classification systems use

stopword list to delete "noise" words [33] before going to the classification algorithm. The system examined two types of stopwords: (a) the words appearing in every category (these words depend on text domains, the number of categories, and the volume of the training texts), and (b) the words which do not apparently characterize any category, examples of which are 'the', 'is', 'doing', 'have', etc that appear very frequently in all the documents, but almost always carry no useful information.

The main interest is in words that most frequently characterize the news into any one of the classes and virtually assure that a given news article belongs to a particular category only.

The next step is to go in for grammatical stemming [25]. A stemmer is an algorithm that determines a stem form of a given word. The stem need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root. The purpose of stemming is to map different forms of an identically implying word to the same feature in the classifier's training space.

Since the knowledge base plays an important role in this type of classification system, the method of training set selection, generation and refinement of knowledge base are crucial problems [20]. In this paper, for building a knowledge base, the following parameters are considered.

4.1.1 Treatment of stopword

Though the necessity of stopwords has rather been discussed [5], [6], the creation of training set also involves the removal of stopwords.

4.1.2 Keyword weighting

A set of noun words, verb words and unit symbols appearing in an article as 'keywords' are defined here [19], [26]. The system adopts the frequency of the keyword [28,38] as its assigned weight [30]. These weights may be normalized between categories in order to adjust to any difference in the quantity of training articles assigned to each category.

4.1.3 Scoping for keywords in a defined area

As it is well known, most keywords are included in the headline and within the first paragraph of any news article [37]. Therefore, defining which areas in an article are to be checked for extracting keywords may be more effective.

Selection of training articles plays a vital role in the over all performance of the classifier. If training articles are not selected appropriately, the classification system will be of no use even if the classification algorithm is of excellent quality. In this paper, the following two kinds of parameters are considered in selecting training articles:

a. *Quantity:* To generate a knowledge base from a large quantity of classified training articles takes much computing time [10]. However, the fewer the training articles the worse the classification correctness becomes.
b. *Publication date time-lapse between training and evaluation articles:* The topics of the news articles are continually changing and new words are constantly appearing. We assume that the classification correctness would be better if the time elapsed between the publication dates of the training and the evaluation articles were closer. This parameter is also relevant to continually maintaining the quality of the knowledge base.

4.2 Hidden Markov Model

A method of feature selection in texts and an effective technique of classifying them are described in this part [40]. When classifying texts, words included in them are used as classification features [21]. Undoubtedly, Markovian models are now regarded as one of the most significant state-of-the-art approaches for sequence learning. Besides several applications in pattern recognition and molecular biology, HMMs have also been applied to text related tasks, including natural language processing [7] and, more recently, information retrieval and extraction [14, 36]. The recent view of the HMM as a particular case of Bayesian networks [11], has helped their theoretical understanding and the ability to conceive extensions to the standard model in a sound and formally elegant framework. HMMs are used in this project for the feature extraction and primary classification of the given input news.

We consider that a text is a sequence of observations $O=(O_1,\ldots,O_t)$. The observations O_t correspond to the tokens of the text. Technically, each token is a vector of attributes generated by a collection of NLP tools. We should attach a semantic tag X_i to some of the tokens O_t. An extraction algorithm maps an observation sequence $O_1,\ldots, O_t$ to a single sequence of tags $(\tau_1,\ldots\ldots,\tau_T)$ where $\tau_i \in \{X_1,\ldots,X_k,\Lambda\}$.

An HMM $\lambda=(\pi,A,B)$ consists of finitely many states $\{S_1,\ldots,S_n\}$ with probabilities $\pi_I = P(q_1 = S_i)$, the probability of starting in state S_i and $a_{ij} = P(q_{t+1} = S_j \mid q_t = S_i)$, the probability of transition from state S_i to S_j . Each state is characterized by a probability distribution $b_i(O_t) = P(O_t \mid q_t = S_i)$ over observations. Given an observation sequence $O=(O_1,\ldots,O_T)$, and according to Bayes' principle, for each observation O_t, we have to return the tag X_i which maximizes the probability $P(\tau_t = X_i \mid O)$, which means that we should identify a sequence of states $q_1,\ldots,q_T$ which maximize $P(q_t = S_i \mid O, \lambda)$ and return that tag X_i that corresponds to the state S_i for each token O_t.

The forward-backward algorithm of HMM is used further.

The algorithm comprises three steps:

1. Computing forward probabilities
2. Computing backward probabilities

3. Computing smoothed values

$\alpha_t(i) = P(q_t = S_i, O_1,\ldots, O_t \mid \lambda)$ is the forward variable. It quantifies the probability of reaching state S_i at time t and observing the initial part $O_{1,\ldots,} O_t$ of the observation sequence. . $\beta_t(i) = P(O_{t+1},\ldots, O_T \mid q_t = S_i, \lambda)$ is the backward variable and quantifies the chance of observing the rest sequence $O_{t+1},\ldots, O_T$ when in state S_i at time t. Of course, $\alpha_t(i)$ and $\beta_t(i)$ can be computed and then we can express the probability of being in state S_i at time t given observation sequence O, it is $\gamma_t(i) = \alpha_t(i)\beta_t(i) / P(O \mid \lambda)$

The extraction using HMM can be described as follows.

Input text T = $(W_1,\ldots,W_t)$; HMM λ ; set of tags $X_1,\ldots,X_n$ corresponding to target HMM states $S1,\ldots,S_n$

Generate sequence O = $(O_1,\ldots,O_t)$, where O_t is a vector containing word w_t.

Call Forward-Backward algorithm and calculate $q_t^* = \max s_i\gamma_t(i), 1 \le t \le T$

If $q_t^* = S_t \in \{S_1,\ldots S_n\}$ then output "<Xi>wt<Xi>";

else output w_t.

After the feature extraction process, the output can be normalized into a new feature vector, and then the trained SVM classifier is ready to be used for classifying a new text.

Consider K classes l = $\{l_1,\ldots, l_k\}$ with their respective HMM set λ′ = $\{\lambda_1\ldots, \lambda_k\}$, where $\lambda^i = \{\lambda^i_1\ldots, \lambda^i_{Ni}\}$ So the total number of HMM $N = \sum_{i=1}^{k} N_i$.

Compute probability, P (O| λ) for a group of HMM $\lambda^i = \{\lambda^i_1\ldots, \lambda^i_{Ni}\}$.

Suppose,

$$P_{\max_i} = \max_{1 \le j \le N_t} P(O \mid \lambda_j^i)$$

Where the label l_i corresponds the maximum probability in the group. Of course, we get the feature vector g= $\{w_1\ldots w_t\ldots w_m\}$ with the HMM λij . Then l_i and g are combined into a new feature vector $\vec{g} = \{l_i, g\}$. Normalize the obtained new feature vector by $\tilde{\vec{g}} = \frac{\vec{g}}{\|g\|^2}$

and this plays an important role in SVM classification for pre-processing.

The above section described a clear idea about the working of HMM as a feature Extractor. Next comes the learning of parameters λ of an HMM from a set $O = \{O^1, O^2, \ldots, O^K\}$ of example sequences [41]. Applying another algorithm called Baum- Welch algorithm solves it, which estimate the parameters using the following formulae

$$\overline{a_{ij}} = \frac{\sum_{k=1}^{k}\sum_{t=1}^{T_{k-1}} \varepsilon_t^k(i,j)}{\sum_{k=1}^{k}\sum_{t=1}^{T_{k-1}} \gamma_t^k(i)} \quad (1)$$

$$\overline{b_j}(i) = \frac{\sum_{k=1}^{k}\sum_{t=1,O_t^k=v_n}^{T_k} \gamma_t^k(i)}{\sum_{k=1}^{k}\sum_{t=1}^{T_k} \gamma_t^k(j)} \quad (2)$$

Where $\varepsilon_t^k(i,j)$ and $\gamma_t^k(i)$ are the joint event and the state variable associated with the kth observation sequence respectively [34]. The feature extraction using HMM is clearly depicted in Figure 2.

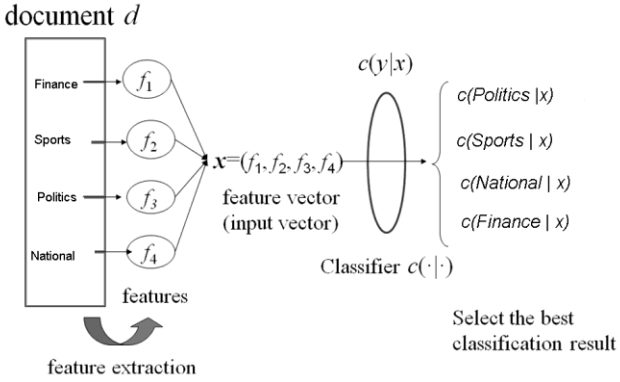

Figure 2. HMM Based Feature Extraction

4.3 Support Vector Machine.

Support Vector Machine (SVM) is a classification technique that was first applied to text categorization by Joachims [1, 9]. It is a powerful supervised learning paradigm based on the structured risk minimization principle. During training, this algorithm constructs a hyperplane that maximally separates the positive and negative instances in the training set. Classification of new instances is then performed by determining which side of the hyperplane they fall on [16].

Most of the previous studies that apply SVM to text categorization use all the words in the document collection without any attempt to identify the important keywords [17]. On the other hand, there are various remarkable studies on keyword selection for text categorization in the literature [23]. As stated above, these studies mainly focus on keyword selection metrics and employ either the corpus-based or the class-based keyword selection approach, do not use standard datasets, and mostly lack a time complexity analysis of the proposed methods [32, 39]. In addition, most studies do not use SVM as the classification algorithm. For

instance, Yang [12] and Pedersen [21] use kNN, and Mladenic and Grobelnic [22] use Naive Bayes [31] in their studies on keyword selection metrics. Later studies reveal that SVM performs consistently better than these classification algorithms.

Assume a group of examples $\{(x_1, y1), (x_2, y2),.., (x_k, y_k)\}$ where $x_k \in R^*$ and $y_k \in \{-1,+1\}$. We consider decision functions of the form sgn ((w.x) + b), where (w.x) denotes the inner product of w and x. So a decision function $f_{w,b}$ should be found with the properties

$$y_i((w \cdot x)+b) \geq 1, 1 \leq i \leq k \tag{3}$$

But in many cases, the separating hyper plane does not exit. To allow for possibilities for violating Equation (3), slack variables like

$$Minimize\ \phi(w,\xi) = (w \cdot w) + C\sum_{i=1}^{k} \xi_i \tag{4}$$

$$Subject\ to\ y_i((w \cdot x)+b) \geq 1-\xi_i, 1 \leq i \leq k$$

are applied.

The above minimization problem is a constrained quadratic programming (QP) problem, which can be formulated as the following convex QP problem:

$$Maximize \sum_{i=1}^{k} a_i - \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k} a_i a_j y_i y_j (x_i \cdot x_j) \tag{5}$$

$$Subject\ to \sum_{i=1}^{k} y_i \alpha_i = 0, 0 \leq \alpha_i \leq C (i = 1,..., k)$$

Where α_i are Lagrange multipliers, C is a parameter that assigns penalty cost to misclassification of samples. By solving the above QP problem, the solution gives rise to a decision function of the form:

$$f(x) = \mathrm{sgn}\left\{\sum_{i=1}^{k} \alpha_i y_i (x_i \cdot x) + b\right\} \tag{6}$$

Where b is a bias term. Only a small fraction of the coefficient α_i is nonzero. The corresponding pairs of entries are known as support vectors and they fully define the decision function [5, 6].

Thus, the above decision function is expressed as an inner product of the data. This observation leads to the generalization to the nonlinear case, which is achieved by mapping the problem data to a higher dimensional space H(feature space) through a transformation of the form $x_i.x_j = Ø(x_i)\ Ø(x_j)$, the mapping function is implicitly defined through a symmetric positive definite kernel function $K(x_i, x_j) = Ø(x_i)\ Ø(x_j)$. Then the decision function can be rewritten as:

$$f(x) = \mathrm{sgn}\left\{\sum_{i=1}^{k} \alpha_i y_i K(x_i, x) + b\right\} \tag{7}$$

A training algorithm of multi-class SVM can be described as the task of constructing a number of binary SVMs; one classifier C_{ij} for every pair of distinct classes i and j [35]. Each classifier C_{ij} is trained with samples in the i^{th} class with positive labels and the samples in the j^{th} class with negative labels. The decision function used in this classification is given by

$$f_{ij}(\tilde{\vec{g}}) = \sum_{n}^{\Lambda} y_n^{l_i l_j} \alpha_n^{l_i l_j} \phi(\tilde{\vec{l}}_i l_j, \tilde{\vec{g}}) + b^{l_i l_j} \tag{8}$$

$$i \neq j, i = 1,...,M$$

Where Λ is the total number of the i^{th} and j^{th} classes from the training data? We can see that SVM method described above is one-against-one method [2]. When given an unknown sample, if the decision function predicted that the sample belongs to class i, the classifier C_{ij} attributes one vote for that class, otherwise the vote is attributed to class j. When all the votes from all the binary classifiers are obtained, the unknown sample belongs to the class having the highest votes.

5. EXPERIMENTAL RESULTS

To evaluate the effectiveness of the new classification method proposed in this paper, we choose a text set, which consists of 1200 news, taken from The Hindu [42], The New Indian Express [43], Business Line [44], The Economic Times [45], and Times of India [46], distributed among three major categories such as sports, finance, and politics. In our experiment, 800 texts are used as training set, and the rest 400 texts are used as testing set. The distribution of the test news from various news sites is in Table 1.

Table 1. Test news distribution among various classes

Source / Classes	The Hindu	The New Indian Express	Business Line	The Economic Times	Times of India	Total
Sports	58	30	0	0	18	106
Finance	30	22	40	45	27	164
Politics	48	31	5	5	41	130
Total	136	83	45	50	86	400

HMMs are trained to extract features from texts of every class, and multi class SVMs are trained to find the separating decision hyper plane that maximizes the margin of the classified categories. Out of the 106 sports news input, 98 were classified correctly. Among the 164 finance news input, 158 news were correctly classified and out of the 130 politics test news 118 were classified as politics news.

Up on analyzing the misclassified news we found that there exists some sort of ambiguity in the text features in the news input. For instance the terror attack on Sri Lankan cricket players at Lahore contains some features that can mislead our classifier to classify the news as sports news, even though it belongs to other (International) category. The classification accuracy of the proposed system for categories Sports, Finance and Politics are 92.45%, 96.34% and 90.76% respectively. We compared the

classification accuracy of this method with that of *k*NN and SVM, and the test results are shown in Table 2.

Table 2. News Classification accuracy of three methods

Method / Category	KNN (%)	SVM (%)	HMM-SVM (%)
Sports	83.25	87.67	92.45
Finance	80.22	82.57	96.34
Politics	82.26	86.55	90.76

The proposed method's classification accuracy is plotted with that of *k*NN and SVM classifiers in Figure 3.

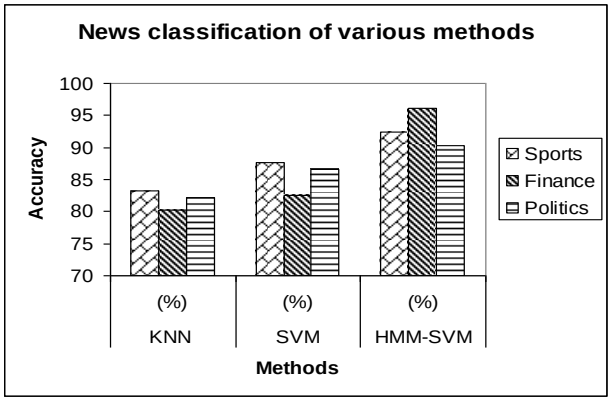

Figure 3. Graphical representation of classification accuracy of various methods

6. CONCLUSION

The intelligent News Classifier is developed and experimented with online news from web for the category Sports, Finance and Politics. The novel approach combining two powerful algorithms, Hidden Markov Model and Support Vector Machine, in the online news classification domain provides extremely good result compared to existing methodologies. By the introduction of several preprocessing techniques and the application of filters we reduced the noise to a great extent, which in turn improved the classification accuracy. Preprocessing in the training data set significantly reduced the training computational time. The experimental result shows the performance of this new approach compared to the existing techniques.

7. REFERENCES

[1] T. Joachims, Learning to Classify Text using Support Vector Machines, Kluwer, 2002.

[2] R. Yan, Y. Liu, A. Hauptmann. On Predicting Rare Classes with SVM Ensembles in Scene Classification. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP03), April 6-10 2003.

[3] D. Tao, X. Tang, X. Li, and X. Wu. Asymmetric Bagging and Random Subspacing for Support Vector Machines-based Relevance Feedback in Image Retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006.

[4] Lei Tang, Huan Liu. Bias Analysis in Text Classification for Highly Skewed Data, Proceedings of Fifth IEEE International Conference on Data Mining (ICDM'05), 2005, pp. 781-784.

[5] J. Brank & M. Grobelnik, N. Milic-Frayling, D. Mladenic. Training text classifiers with SVM on very few positive examples. Microsoft Research technical report MSR-TR-2003-34. 2003.

[6] D. Lin & P. Pantel. Discovery of Inference Rules for Question Answering. Natural Language Engineering 2001 7(4):343-360.

[7] D. Lin & P. Pantel. Induction of Semantic Classes from Natural Language Text. In Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2001. pp. 317-322.

[8] E.Voorhees. Query expansion using lexical-semantic relations. In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1994, pp. 61—69.

[9] Joachims, T.: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. European Conference on Machine Learning (ECML) (1998) Text Categorization with Class-Based and Corpus-Based Keyword Selection.

[10] Ozgur, L., Gungor, T., Gurgen, F.: Adaptive Anti-Spam Filtering for Agglutinative Languages. A Special Case for Turkish, Pattern Recognition Letters, 25 no.16 (2004) 1819–1831.

[11] McCallum, A., Nigam, K.: A Comparison of Event Models for Naive Bayes Text Classification. Sahami, M. (Ed.), Proc. of AAAI Workshop on Learning for Text Categorization (1998), Madison, WI, 41–48.

[12] Yang, Y., Liu, X.: A Re-examination of Text Categorization Methods. In Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval, Berkeley, US (1996).

[13] Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Computing Surveys 34 no. 5 (2002) 1–47.

[14] Forman, G.: An Extensive Empirical Study of Feature Selection Metrics for Text Classi.cation. Journal of Machine Learning Research 3 (2003) 1289–1305.

[15] Ozgur, A.: Supervised and Unsupervised Machine Learning Techniques for Text Document Categorization. Master's Thesis (2004), Bogazici University, Turkey.

[16] Burges, C. J. C.: A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery Vol. 2 No. 2 (1998) 121–167.

[17] Joachims, T.: Advances in Kernel Methods-Support Vector Learning. chapter Making Large-Scale SVM Learning Practical MIT-Press (1999).

[18] Lin, S-H., Shih C-S., Chen, M. C., Ho, J-M.: Extracting Classi.cation Knowledge of Internet Documents with Mining Term Associations: A Semantic Approach. In Proc. of ACM/SIGIR (1998), Melbourne, Australia 241–249.

[19] Azcarraga, A. P., Yap, T., Chua, T. S.: Comparing Keyword Extraction Techniques for Websom Text Archives. International Journal of Artificial Intelligence Tools 11 no. 2 (2002).

[20] Aizawa, A.: Linguistic Techniques to Improve the Performance of Automatic Text Categorization. In Proceedings of 6th Natural Language Processing Pacific Rim Symposium (2001), Tokyo, JP 307–314.

[21] Yang, Y., Pedersen J. O.: A Comparative Study on Feature Selection in Text Categorization. In Proceedings of the 14th International Conference on Machine Learning (1997) 412–420.

[22] Mladenic, D., Grobelnic, M.: Feature Selection for Unbalanced Class Distribution and Naive Bayes. In Proceedings of the 16th International Conference on Machine Learning (1999) 258–267.

[23] Salton, G., Yang, C., Wong, A.: A Vector-Space Model for Automatic Indexing. Communications of the ACM 18 no.11 (1975) 613–620.

[24] ftp://ftp.cs.cornell.edu/pub/smart/ (2004).

[25] Porter, M. F.: An Algorithm for Suffix Stripping. Program 14 (1980) 130–137.

[26] Salton, G., Buckley, C.: Term Weighting Approaches in Automatic Text Retrieval. Information Processing and Management 24 no. 5 (1988) 513–523

[27] Lewis, D. D.: Reuters-21578 Document Corpus V1.0, http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html.

[28] Kroha, P.; Baeza-Yates, R. A Case Study: News Classification Based on Term Frequency , Sixteenth International Workshop on Database and Expert Systems Applications, 2005. Proceedings.

[29] Lin Lv; Yu-Shu Liu; Research of English text classification methods based on semantic meaning: ITI 3rd International Conference on information and Communications Technology, 2005. Enabling Technologies for the New Knowledge Society.

[30] Islam, Md. Rafiqul; Islam, Md. Rakibul; An effective term weighting method using random walk model for text classification, 11th International Conference on Computer and Information Technology, 2008. ICCIT 2008. 24-27 Dec. 2008.

[31] Sang-Bum Kim; Kyoung-Soo Han; Hae-Chang Rim; Sung Hyon Myaeng; Some Effective Techniques for Naive Bayes Text Classification, IEEE Transactions on Knowledge and Data Engineering, Volume 18, Issue 11, Nov,2006.

[32] Miao Zhang; De-xian Zhang; Trained SVMs based rules extraction method for text classification, IEEE International Symposium on IT in Medicine and Education, 2008. ITME 2008, 12-14 Dec. 2008.

[33] Agarwal, S.; Godbole, S.; Punjani, D.; Shourya Roy; How Much Noise Is Too Much: A Study in Automatic Text Classification, Seventh IEEE International Conference on Data Mining, 2007. ICDM 2007.

[34] Makrehchi, M.; Kamel, M.S.; Combining feature ranking for text classification, IEEE International Conference on Systems, Man and Cybernetics, 2007. ISIC. 7-10 Oct. 2007.

[35] Guifa Teng; Yihong Liu; Jianbin Ma; Fang Wang; Huiting Yao; Improved Algorithm for Text Classification Based on TSVM, First International Conference on Innovative Computing, Information and Control, 2006. ICICIC '06. Volume 2, Aug. 30 2006-Sept. 1 2006.

[36] Hui He; Bo Chen; Jun Guo; Semi-supervised Chinese compound word extraction based on HMM, 7th World Congress on Intelligent Control and Automation, 2008. WCICA 2008. 25-27 June 2008.

[37] Wei Hu; Dong-Mo Zhang; Huan-Ye Sheng; Vague events-based Chinese Web news classification, Proceedings of 2004 International Conference on Machine Learning and Cybernetics, 2004.

[38] Kroha, P.; Baeza-Yates, R.; A Case Study: News Classification Based on Term Frequency, Proceedings. Sixteenth International Workshop on Database and Expert Systems Applications, 2005. 26-26 Aug. 2005.

[39] Lisbon, Proceedings of the ACM first Ph.D. workshop in Information and Knowledge Management, Portugal, 2007.

[40] Jun-Peng Bao Jun-Yi Shen Xiao-Dong Liu Qin-Bao Song , A new text feature extraction model and its application in document copy detection, I International Conference on Machine Learning and Cybernetics, 2003.

[41] Harriman, Feature selection and feature extraction for text categorization, Proceedings of the workshop on Speech and Natural Language, Human Language Technology Conference archive.

[42] www.hinduonnet.com

[43] www.expressbuzz.com

[44] www.thehindubusinessline.com

[45] http://economictimes.indiatimes.com

[46] http://timesofindia.indiatimes.com