ISSN 2277-3061

Design and Implementation of Hybrid Algorithm for e-news Classification

Harneet Kaur, Dr. Kiran Jyoti
Research fellow, M.Tech, Department of Information Technology,
Guru Nanak Dev Engineering College , Ludhiana
harneetkaur87@yahoo.co.in
Assistant Professor, Department of Information Technology,
Guru Nanak Dev Engineering College , Ludhiana

ABSTRACT

Data mining involves the use of data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. As the use of internet is increasing day by day and with the advancement of internet news also publish online. So to handle this bulk amount of news various data mining techniques for classification had been used. In this paper we are using an intelligent system based on Hybrid algorithm (HMM, SVM and CART) for e-news classification. An intelligent system is designed which will extract the online news and then will find out category and subcategory wise news. System involves four main stages: a) Keyword Extraction b) Implementation of Hybrid Algorithm (HMM, SVM and CART). Data have been collected for experimentation from online newspapers like The Hindu, Hindustan Times and Times of India. The experimental results are based on the news categories and sub categories such as Entertainment: Bollywood 100% and Hollywood 90%, Sports: Cricket 90%, Football 90% and Hockey 78%, Matrimonial :Hindu 100% and Muslim 80%. In this paper we also compare the result of Hybrid algorithm (HMM, SVM and CART) with individual HMM and SVM Algorithm and conclude that Hybrid algorithm (HMM, SVM and CART) gave better result than that of what HMM and SVM individually gave.

Keywords

Knowledge base, SVM, HMM, Hybrid, CART

Council for Innovative Research
Peer Review Research Publishing System

Journal: INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY
Vol 12, No.1
editor@cirworld.com
www.cirworld.com, member.cirworld.com

INTRODUCTION

Data mining also known as knowledge discovery, which is computer-aided process of identifying hidden patterns by digging and analyzing enormous sets of data and then extracting the meaning of the data. By using pattern recognition technologies and statistical and mathematical techniques to go through warehoused information, data mining helps analysts recognize significant facts, relationships, trends, patterns, exceptions and anomalies that might otherwise go unnoticed. In this paper we are considering online news for classification with hybrid algorithm (HMM, SVM and CART). Online news classification is very essential to obtain relevant information quickly from a bulk amount of news articles published.

For the purpose of developing text classification system many researchers devoted their time for developing automated text classification .In early days the work of classification and indexing of online news was totally manual where rule base was generated by human expertise. So it was very time consuming process with less accuracy and more effort. On the other hand , statistical approach based on keyword extraction from training texts is a popular method of generating a knowledge base[13] . We need to follow some guidelines on how to select training data. We use Hybrid algorithm (HMM, SVM and CART) which is automated intelligent system that will conclude the result without taking much time, with less effort and with high accuracy rate.

RELATED WORK

We had collected detailed information on related work from where we got an idea to do our work. Before going to our new intelligent system , it is essential to have an overview of various existing methodologies related to our work:

A) Manual and Fuzzy Text classification

B) Automated Text classification

In earlier days of online news classification work was totally manual and it was very time consuming , labor intensive and expensive work. It was very time consuming and difficult process to classify various categories. It was very expensive and laborious work. Also there were problems of accuracy. So it was necessary to build an automated system that can resolve these issues and then Carnegie Group took an initialized step to build an automated news categorization system that was based on fuzzy rule based text categorization. In the last 15 years and so, substantial research has been conducted on text classification through supervised machine learning techniques.Techniques like KNN (K Nearest Neighbour) for classification of text. But their accuracy that is the major issue , was not good for online news and journals. Instead of manually classifying documents or hand-crafting automatic classification rules, statistical text categorization uses machine learning methods to learn automatic classification rules based on human-labeled training documents. So in this research we have done classification by implementing Hybrid Algorithm (HMM, SVM and CART) that gave better results of classification. Our intelligent system gave better accuracy results for news classification as well as sub-classification.

TEXT CLASSIFICATION PROCEDURE

Text classification (a.k.a. text categorization) is the task of assigning predefined categories to free-text documents. It can provide conceptual views of document collections and has important applications in the real world. For example, news stories are typically organized by subject categories (topics) or geographical codes; academic papers are often classified by technical domains and sub-domains; patient reports in health-care organizations are often indexed from multiple aspects, using taxonomies of disease categories, types of surgical procedures, insurance reimbursement codes and so on. Another widespread application of text categorization is spam filtering, where email messages are classified into the two categories of spam and non-spam, respectively. The procedure for online news classification is as follow:

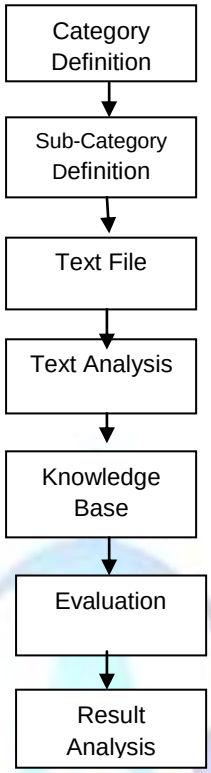
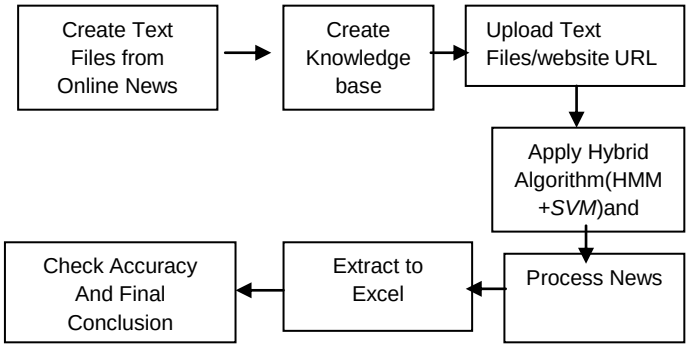


Fig 1: Classification Steps

- Category Definition: First of all the categories to be considered should be defined eg. Entertainment and Sports.
- Sub-Category Definition: Subcategories that are to be considered should be defined eg. Bollywood and Hollywood under category Entertainment and Cricket, football, Hockey under category Sports .
- Text File: Text file should be created by using text from online newspapers that will be used as training set.
- Text Analysis: Length of text, No. of Text under each category and Keyword Distribution is very important.
- Knowledge Base Generation: Training set is used to generate Knowledge Base. Weighted keywords should be used to generate knowledge base. We should carefully select the keywords in order to maintain good accuracy percentage.
- Evaluation: Classification correctness is to be calculated.
- Result Analysis: Knowledge base will need to be refined if result will not be sufficient.

PROPOSED WORK

The proposed intelligent system is designed and developed by implementing Hybrid algorithm(HMM,SVM and CART). The phases of our research work is as given in fig.2



\

Fig.2 Proposed Work Flow

A. Create Text File

We will feed news in our system in the form of text files and these text files will be created from online news papers that we consider in our work. These text files will contain news of Entertainment – Bollywood and Hollywood and Sports – Cricket, Football and Hockey .

B. Creation of Knowledge base

Creation of Knowledge base for classification plays an important role. It is generated using training set .Keywords are used which defines the character of each category and subcategory. The algorithms that we used will consider these keywords in their work to reach to the final conclusion that is why it played an important role. Before choosing the keywords we consider some important parameters that plays an important part for accuracy:

- Stop words : Stop words are known as “noise” words. Basically there are two types of stop words: (a) the words that appears in every category (these words depends on text domain, number of categories and volume of training texts) ,and (b)the words which do not characterize any of the category eg. is, am ,are, for, doing, where etc that appear frequently in all the documents , but these words are not useful for our knowledge base. So we avoid using these type of “noise” words.
- Keywords scope: Scoping for keywords means the area where from where we consider the keywords in our knowledge base. We know mostly keywords are included in headline or in the first paragraph of any particular news so we took keywords from headline or first paragraph. Selection of training articles i.e. keywords plays a very important role in the over all performance of the classifier so we must be very careful while choosing the keywords.

  In our paper we considered two types of parameters for choosing the keywords:

  i. Quantity: If there will be less keywords then our result will be very poor so we should consider more keywords for better result but it will increase our computing time.

  ii. Publication date-time lapse between training and evaluation articles: As we know , news are frequently changing so new words are also frequently appearing day to day.so we assume that publication date of training articles and evaluation articles should be closer in order to obtain the classification correctness.

  So these are the two parameters that we considered for better output of our work.

C. Upload Text file/website URL and process news

First of all we will give url of web page of e news that we are considering at a time then HMM algorithm will extract the text by removing unnecessary tags and by removing the space . After that we will select the news which we want to know to which category that news belongs and then when we process that news . SVM algorithm will distinguish the keywords and CART will compare the keywords with the training set of every category then SVM will be used to classify the news to particular category to which that news actually belongs. Our system then show us that this news belongs to this category. The result will be stored in Microsoft excel sheet.

D. Hidden Markov Model

A hidden Markov model (HMM) includes a method of feature extraction and is an effective technique of classifying them. While classifying text, words included in them are used as classification features[5]. HMM is regarded as one of the most significant state-of-the-art approaches for sequence learning.Hidden Markov models are especially known for their application in temporal pattern recognition such as speech, handwriting, gesture recognition,part-of-speech tagging, musical score following, partial discharges ,bioinformatics and extraction.

A hidden Markov model can be considered a generalization of a mixture model where the hidden variables (or latent variables), which control the mixture component to be selected for each observation, are related through a Markov process rather than independent of each other.

It is a finite set of states, each of which is associated with a (generally multidimensional) probability distribution . Transitions among the states are governed by a set of probabilities called transition probabilities. In a particular state an outcome or observation can be generated, according to the associated probability distribution. It is only the outcome, not the state visible to an external observer and therefore states are ``hidden'' to the outside; hence the name Hidden Markov Model.

A hidden Markov model (HMM) is a triple (π,A,B).

Π= π(i) the vector of the initial state probabilities;

A=(a(i ,j)) the state transition matrix;

B=(b(i , j)) the confusion matrix;

Each probability in the state transition matrix and in the confusion matrix is time independent - that is, the matrices do not change in time as the system evolves. In practice, this is one of the most unrealistic assumptions of Markov models about real processes.

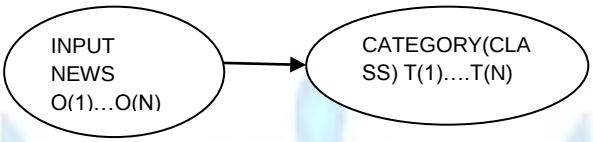
HMM is used for two main purposes:

a) Feature extraction of input news

b) Primary classification of input news

The news that we input are in the form of text which is basically a sequence of observations O=(O(1),……,O(n)). We will attach a semantic tag T(i) to some of the tokens O(n).

The extraction algorithm's work will be to map an observation sequence O(1).....O(n) to a single sequence of tags (T(1).....T(n)).



Tags are the categories in which we are dividing our text file .

Input text i.e. T= (W(1),.....W(n)) equivalent to O(i).....O(n).

HMM λ

Set of tags T(1).....T(n) equivalent to target HMM states S(1).....S(n).

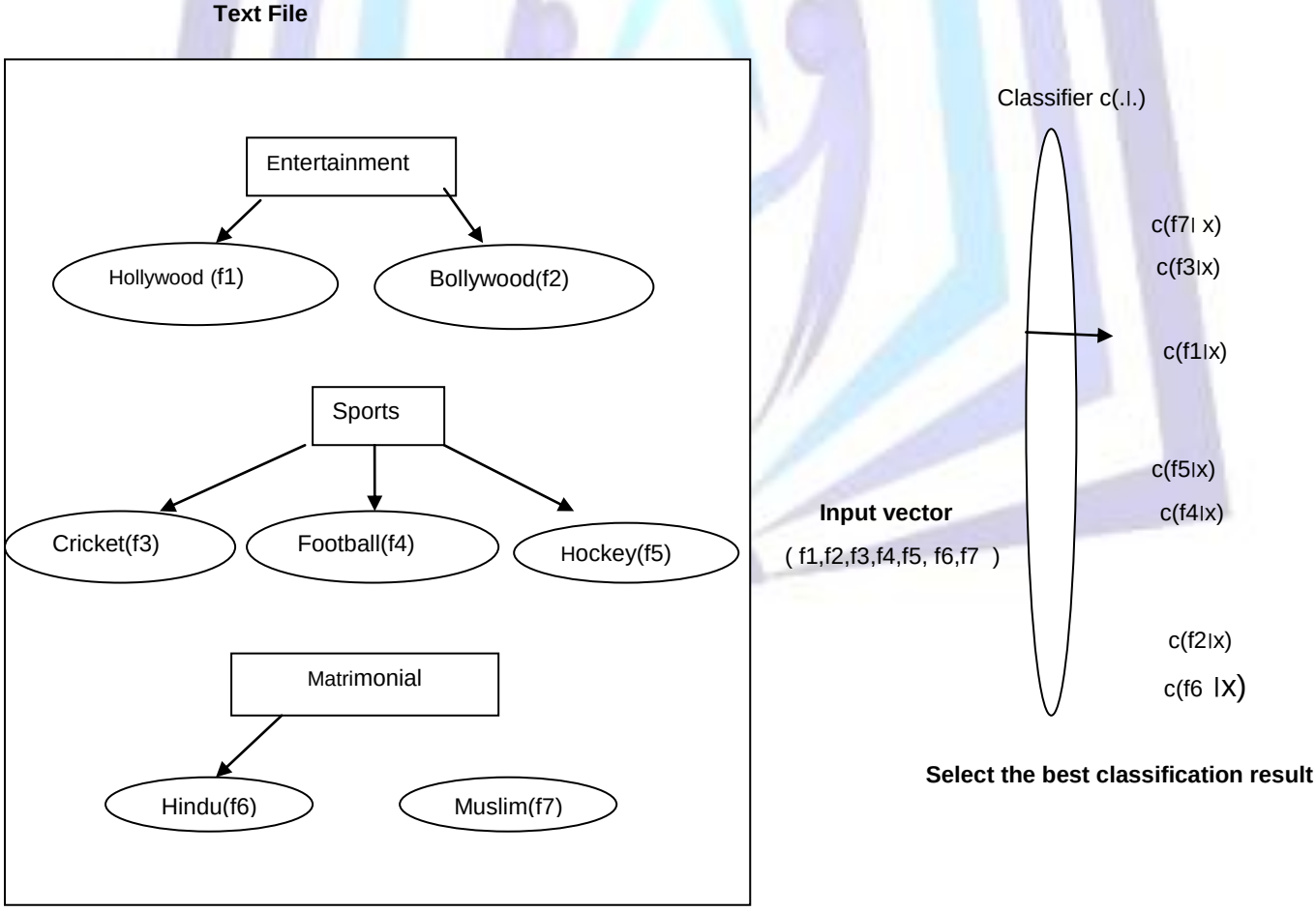


Fig.3 HMM based Feature Extraction

E. Support Vector Machine

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on [8].

In previous work for classification, SVM used all words in document instead of choosing meaningful keywords .On the other hand there are various studies on keywords[9] selection for text categorization, Main focus of these studies is on keyword selection metrics and employ either corpus based or class based keyword selection approach, do not use standard data sets. Most studies do not use SVM as classification algorithm e.g. Yang[10] and Pederson[5] used KNN , and Mladenic and Grobelnic [11] use Naïve Bayes in their studies on keyword selection metrics. but later studies reveal that SVM is best for classification.

Writing the classification rule in its unconstrained dual form reveals that the maximum-margin hyperplane and therefore the classification task is only a function of the support vectors, the subset of the training data that lie on the margin.

Using the fact that $\|\mathbf{w}\|^2 = w \cdot w$ and substituting $\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x_i}$, one can show that the dual of the SVM reduces to the following optimization problem:

Maximize (in α_i)

$$\tilde{L}(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \quad (1)$$

subject to (for any i= 1,….,n)

$$\alpha_i \geq 0,$$

and to the constraint from the minimization in b

$$\sum_{i=1}^{n} \alpha_i y_i = 0. \quad (2)$$

Here the kernel is defined by . $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$ (3)

W can be computed thanks to the α terms :

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i . \quad (4)$$

F. CART

CART (Classification and Regression Tree) is a classification methodology that uses the present data to form decision trees and these decision trees will then be used to classify new data. In our work we are using this CART algorithm which will create a decision tree before going to the final conclusion.

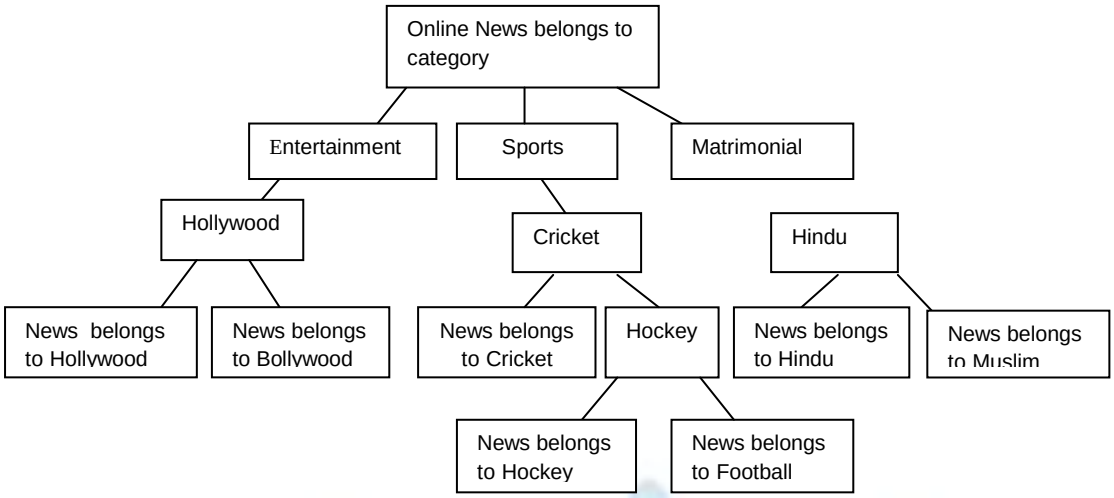


Fig.4 CART Decision

Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. It is one of the predictive modeling approaches used in statistics, data mining and machine learning. More descriptive names for such tree models are classification trees or regression trees. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels .These are represented by a set of questions which will splits the learning sample into smaller parts then CART will asks question in the form of Yes or No e.g. in our work Cart will ask the question like If news belongs to Entertainment then News belongs to Hollywood? If yes then final conclusion will be “News belongs to Hollywood” .If No then it will check with the keywords of Bollywood and if keywords will match then CART will conclude the result as “News belongs to Bollywood.

G. Evaluation

Evaluation process will give output of our work, accuracy of our work.We check the accuracy by entering the no. of news actually entered in the form of text file for particular category and then output will show us that how many news our Hybrid algorithm(HMM,SVM and CART) had accurately classified. In this way we can check percentage accuracy of each category.

Experimental Results

To evaluate the effectiveness of news classification method proposed in this paper, we choose text set of 124 news as testing set as shown in table 1. These news are taken from The Hindu [16] , Hindustan Times[14] and The Times of India[15] and distributed among categories Entertainment and Sports news which which sub- categorized as Bollywood, Hollywood, Cricket, Football and Hockey news.

Table 1. NEWS DISTRIBUTION AMONG CATEGORIES AND SUB- CATEGORIES

Source / Classes	THE HINDU	Hindustan times	The times of India	Total
ENTERTAINMENT (BOLLYWOOD)	0	20	20	40
ENTERTAINMENT (HOLLYWOOD)	0	20	15	35
SPORTS(CRICKET)	30	20	15	65
SPORTS (FOOTBALL)	18	18	15	51
SPORTS (HOCKEY)	15	0	15	30
MATRIMONIAL (HINDU+MUSLIM)	25	0	0	25
TOTAL	88	78	80	246

HMM Algorithm has been trained for feature extraction which extracts features of every class and SVM has been used to maximize the margin of classified categories in order to find the accurate result. We use 75 news articles for entertainment which is divided into sub categories news Bollywood (40) and Hollywood (35). Out of 40 Bollywood ,news all 40 news were classified correctly, out of 35 Hollywood news 32 were correctly classified, out of 65 cricket news 59 were correctly classified, out of 51 football news 47 were correctly classified ,out of 30 hockey news 23 were correctly classified and out of 25 matrimonial news 22 were correctly classified.. We analyzed that the reason behind misclassification is ambiguity in text features. However the classification accuracy of the proposed system for category Entertainment's sub- categories Bollywood and Hollywood are 100 % and 90 % respectively while for Sports's sub-categories Cricket, Football and Hockey are 90%, 90% and 78% respectively .For Matrimonial Hindu news result is 100 % and for Muslim matrimonial news is 80%.We compared classification accuracy of proposed method with that of HMM and SVM individually and the testing results are shown below in Table 2.

Table 2. NEWS CLASSIFICATION ACCURACY OF THREE METHODS

Method / Category (sub-category)	HMM (%)	SVM (%)	HYBRID (HMM+SVM+CART) (%)
ENTERTAINMENT (BOLLYWOOD)	64	76	100
ENTERTAINMENT (HOLLYWOOD)	62	70	90
SPORTS(CRICKET)	80	89	90
SPORTS (FOOTBALL)	70	80	90
SPORTS (HOCKEY)	60	64	78
MATRIMONIAL (HINDU)	42	48	100
MATRIMONIAL (MUSLIM)	48	52	80

Graphical representation of Three methods HMM, SVM and HYBRID (HMM, SVM and CART) is shown below in fig. 5

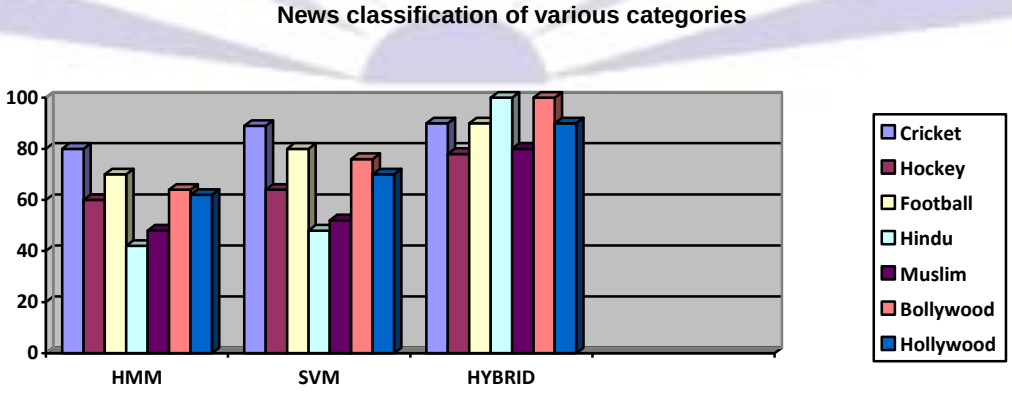


Fig. 5 Graphical Representation of classification Accuracy of Hybrid Algorithm (HMM, SVM and CART)

Conclusion

The proposed news classifier applied on online news for categories Entertainment's sub-categories Bollywood and Hollywood and Sport's sub-categories Cricket, Football and Hockey. We introduced several pre-processing techniques also before applying Hybrid (HMM, SVM and CART) Algorithm .The Hybrid algorithm (HMM, SVM and CART) that we applied provides extremely good results. In existing work SVM and HMM were applied to only categories Sports, Finance and Politics while we extend the work and applied Hybrid (HMM, SVM and CART) for sub-categorization of news also. The experimental results of Hybrid (HMM, SVM and CART) also compared with individual HMM and SVM.

REFERENCES

[1] 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.

[2] 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.

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

[4] Krishnalal G, Babu S Rengarajan and K G Srinivasagan. Article: A New Text Mining Approach Based on HMM-SVM for Web News Classification. International Journal of Computer Applications (0975-8887) Volume1- No.19.

[5] 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.

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

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

[8] 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.

[9]JOACHIMS, T.: Advances in Kernal Methods-Support Vector Learning. Chapter Making Large-Scale SVM Learning Practical MIT- Press(1999).

[10]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).

[11] Mladenic, D., Grobelnic , M.: Feature Selection for Unbalanced Class Distribution and Naïve Bayes . In proceedings of the 16th International conference on Machine Learning(1999) 258-267.

[12] McCallum, A., Nigam, K .: A comparison of Event Models for Naïve Bayes Text Classification . Sahami , M. (Ed.) , Proc. Of AAAI Workshop on Learning for Text Categorization (1998), Madison , WI , 41-48.

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

[14]www.hinsustantimes.com

[15]www.indiatimes.com

[16]www.thehindu.com