# Online News Classification: A Review

Harmandeep Kaur
*Computer Science and Engineering Department, Shri Guru Granth Sahib World University Fatehgarh Sahib*

Sheenam Malhotra
*Assistant Professor, Computer Science and Engineering Department, Shri Guru Granth Sahib World University Fatehgarh Sahib*

**ABSTRACT - Data mining is a field of researches and modifications. Online news classification has been challenge always in terms of manual transaction. In presented work , trying to create a novel algorithm which can classify the inner structures of the simple classified news. For the current scenario , the work has just been done to identify the outer clusters of the system but no work till now has been done for inner cluster of the datasets. This proposed work will be creating inners clusters for each and every field of the proposed system like for SPORTS , ENTERTAINMENT and MATRIMONIALS .**

**Keywords: Hidden Markov Model(HMM),Support Vector Machine(SVM),K Mean, CART**

## I. INTRODUCTION

Data mining is process of discovering interesting knowledge such as patterns, associations, changes, anomalies and significant structures, from large amounts of data stored in database, data warehouse, or other information repositories. Data to the wide availability of huge amount of data in electronic form, and imminent need for turning such data into useful information and knowledge for broad application including market analysis, business management and decision support, data mining has attracted a great deal of attention in information industry in recent year. Data mining has popularly treated as synonym of knowledge discovery in database, although some researchers view data mining as an essential step of knowledge discovery. A knowledge discovery process consist of an iterative sequence of following steps[9]:

- Data cleaning, which handle noisy, erroneous, missing, or irrelevant data.
- Data integration, where multiple, heterogeneous data source may be integrated into one.
- Data selection, where data relevant to analysis task are retrieved from database.
- Data transformation, where data are transformed or consolidated into form appropriate for mining by performing aggregate operations.
- Data mining, which is essential process where intelligent methods are applied in order to extract data patterns.
- Pattern evaluation, which is to identify the truly interesting patterns represent knowledge based on some interestingness measure.
- Knowledge presentation, where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

## II. TEXT CLASSIFICATION

Mostly the information store in the form of text like e mails, web pages, newspaper article, market research reports, complaint letter from customer and internally generated reports. As for as online news papers provide news under various categories like national, international, politics, finance, sports, entertainment etc[4]. Text classification is also an important part of text mining[5] . Text classification based on expert knowledge, how to classify the document under the given set of categories. Data mining classification start with training set of document that are already label with class. Text classification has two flavours as single label and multi label[5]. A single label document is belong to only one class and multi label document may be belong to more than one class. Data stored in most text databases are semi structure data in that they are neither completely unstructured not completely structured. For example a document may contain a few structured field such as title, authors, publication date, category But also contain some largely unstructured text components such as abstract, contents.

## III.TEXT CLASSIFICATION PROCESS

The stages of text classification are discussing as following points[5].

a. Document Collection

In this step collect the different types of document like html, .pdf, .doc etc.

b. Pre-Processing

In this present the text document into clear word format. for example Perform Tokenization, stemming word, removing stop words. In tokenization a document is treated as a string and then partition into list of tokens. In removing stop word remove the stop words for example ”the” ”a”, ”and”. In stemming words converts different words from into similar canonical form.

c. Indexing

In this provide the index to every document with this easily identify each document.

d. feature selection

After preprocessing and indexing the important step of text classification is feature selection . The main idea of feature selection is to select subset of features from the original document . It is performed by keeping the words with highest score according to predetermined measure of the importance of the word.

e. Classification

In this Classify documents into predefined categories. The documents can be classified by supervised, unsupervised methods. When the class label of each document is known that is called supervised classification when the class label of documents are not known that is called unsupervised classification.

f. Performance Evaluation

This is the last stage of text classification. This is experimentally, rather than analytically. In this measure the performance. Many measures have been used like precision and recall.

## IV. RELATED WORK

Text categorization is the problem of classifying text documents into a set of predefined classes[1]. In this investigated three approaches to build a meta classifier in order to increase the classification accuracy. The basic idea is to learn a meta classifier to optimally select the best component classifier for each data point. The result show that combining classifiers can significantly improve the accuracy of classification and meta classification strategy gives better result than each individual classifier.brief introduction on SVM and several application of SVM in pattern recognition problem are given in [2].SVM apply on many application ranging from face detection and recognition, object detection and recognition, handwritten character and digit recognition, information and image retrieval, prediction and etc. two approaches to develop a classifier for text document based on Naïve Bayes Theory and integrate this classifier into a meta classifier in order to increase the classification accuracy[3]. the intelligent news classifier is developed and experimented with online news from web for the category Sports, Finance and Politices[4]. The novel approach combining two powerful algorithms, Hidden Markov Model and Support Vector Machine in online news classification domain provides extremely good result. An intelligent system is design to extract the keywords from online news paper content and classify it according to predefined categories .Three different stages are designed to classify the content of online newspapers such as (1) text pre processing (2) HMM based feature extraction (3)classification using SVM. Brief introduction to various text representation schemes[5]. The existing classification methods are compared and contrasted based on various parameters namely used for classification, algorithm adopted, and classification time complexities. Different algorithms are perform differently depending on data collection. Mostly information stored as text , text mining is believed to have a high commercial potential value, knowledge may be discovered from many source of information , unstructured texts remain the largest readily available source of knowledge. This paper give the introduction of text classification, process of text classification as well as the overview of the classifiers and compare some existing classifier on the basis of few criteria like time complexity, principal, performance. Automatic text classification is semi supervised machine learning task that automatically assign a given document to a pre defined categories based on textual content and extracted features[6]. Automatic text classification has important application in content management, opinion mining, product review analysis and survey existing solution to major issue such as dealing with unstructured text, handling large number of attribute and selecting a machine learning technique appropriate to text classification application. Classify the financial news based on the content of relevant news articles that is accomplished by building a prediction model[7]. Which is able to classify the news as either rise or drop. In this paper forecasting system have been introduced. Apply SVM classification method to personalized classification[8] .In personalized classification user can define their personal categories using few keywords for search queries using these keywords. Categorizer obtain both positive or negative document required for construction of classifiers.[9] paper gives the introduction of data mining and various applications of data mining, life cycle of data mining, knowledge discovery process and various methods of data mining.

## V. PROPOSED WORK

Flow chart of research is given below .It shows the process of news classification.

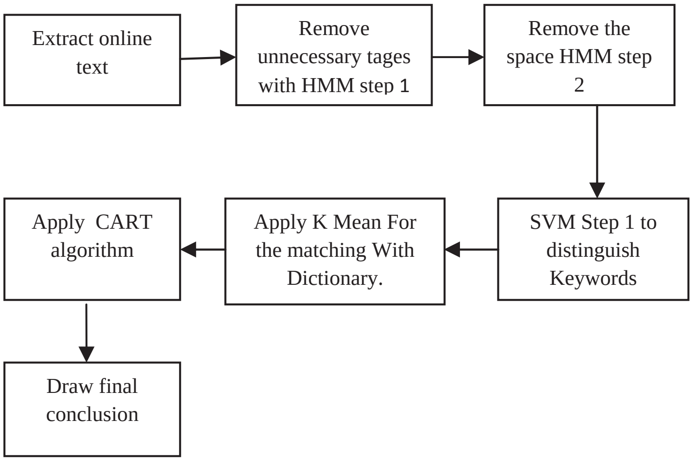
Extract online text
Remove unnecessary tages with HMM step 1
Remove the space HMM step 2
SVM Step 1 to distinguish Keywords
Apply K Mean For the matching With Dictionary.
Apply CART algorithm
Draw final conclusion

Figure no. 1 : News Classification Process

For this work we use two model HMM and SVM. Hidden Markov Model(HMM) use for text extraction and Support Vector Machine (SVM) use for classification. First of all give the url address of online news paper then text will display then various operations will apply for remove the html tages and unnecessary spaces. . K mean algorithm will be use for create the cluster and CART algorithm use for represent it into hierarichal form.

## VI. CONCLUSION

In the work till now , a successful implementation has been done to extract the news from the online portals for the further processing. In addition to it the clusters of different categories has been also created so that the further combination of HMM AND SVM could be applied to it to regain a better efficiency.

## REFERENCES

[1] Daniel I. Morariu, Lucian N. Vintan, and Volker Tresp,”Meta-Classification using SVM Classifiers for Text Documents,”World Academy of Science, Engineering and Technology 21 2006.

[2] Hyeran Byun1 and Seong-Whan Lee2,”Applications of Support Vector Machines for Pattern Recognition: A Survey,”SVM 2002, LNCS 2388, pp. 213-236, 2002.

[3] D. Morariu, R. Cre¸tulescu and L. Vin¸tan,”Improving a SVM Meta-classifier for Text Documents by using Naïve-Bayes,”Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844.

[4] Krishnalal G, S Babu Rengarajan, K G Srinivasagan ,”A new text mining approach based on HMM -SVM for web news classification,”International Journal of Computer Applications (0975 - 8887) Volume 1 – No. 19,2010.

[5] Vandana Korde, C Namrata Mahender,”Text classification and classifier a survey,” International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2, March 2012.

[6] Mita K. Dalal, Mukesh A.Zaveri,”Automatic text classification,”International Journal of Computer Applications (0975 – 8887) Volume 28– No.2, August 2011.

[7] Rama Bharath Kumar, Bangari Shravan Kumar, Chandragiri Shiva Sai Prasad,”Financial news classification using SVM”,International Journal of Scientific and Research Publications, Volume 2, Issue 3, March 2012 .

[8] Chee-Hong Chan Aixin Sun Ee-Peng Lim,”Automated Online News Classifcation with Personalization,”4th international conference on asian digital libraries, Dec 2001.

[9] Mr. S. P. Deshpande , Dr. V. M. Thakare,”data mining system and applications: a review,”International Journal of Distributed and Parallel systems (IJDPS) Vol.1, No.1, September 2010.