INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

(Volume2, Issue4)

Available online at: www.Ijariit.com

Lexicon Analysis Based Automatic News Classification Approach – A Review

Kamaldeep Kaur, Maninder Kaur

deepsandhu9190@gmail.com , *manindererecediet@gmail.com*

Doaba Institute of Engineering & Technology, Kharar

ABSTRACT—The news classification approach is the primary approach for the online news portals with the news data sourced from the various portals. The various types of data is received and accepted over the news classification portals. The lexicon analysis plays the key role in the categorization of the news automatically using the automatic news category recognition by analyzing the keyword data extracted from the input image data. The N-gram news analysis approach will be utilized for the purpose of the keyword extraction, which will further undergo the support vector classification. The support vector machine based classification engine analyzes the extracted keywords against the training keyword data and then returns the final decision upon the detected category. The proposed model is aimed at improving the overall performance of the existing models , which will be measured on the basis of precision, recall, etc.

KEWORDS - News Classification, Regression, Probabilistic Classifier, Automatic Categorization, Multi-domain news analysis.

I. INTRODUCTION

Data mining is process of discovering interesting knowledge such as patterns, associations, changes, anomalies and significant structures, from large amount 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 consists of an iterative sequence of following step:

- Data cleaning, which handles 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 from 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 pattern 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.

Major tasks of data mining

In general, data mining tasks can be classified into two categories: descriptive data mining and Predictive data mining. The former describes the data set in concise summary manner and presents interesting general properties of data. A data mining system may accomplish one or more of the following data mining tasks.

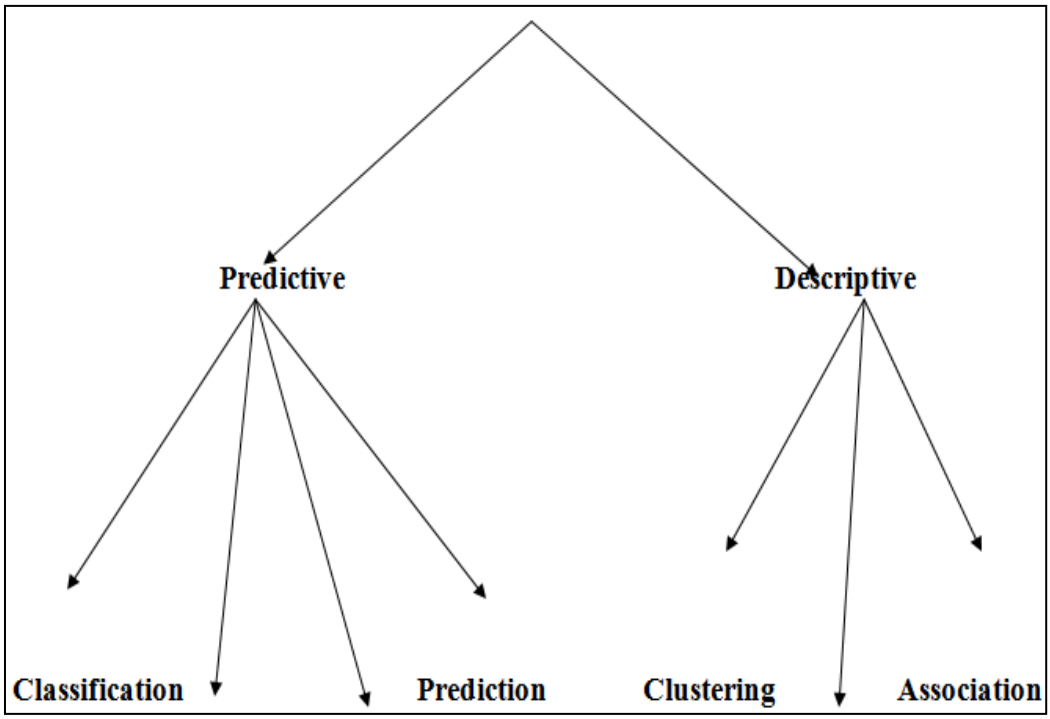

Figure 1: The tree-structure for the inter-relationship between the data mining techniques

- **Class description:** Class description provides a concise and summarization of collection of data and distinguish it from others. The summarization of collection of data is called class characterization the comparison between two or more collections of data is called comparison or discrimination.
- **Association:** Association is discovery of association relationships or correlations among a set of items. There are various association analysis algorithms like Apriori search, mining multiple level, multi dimensional association, mining association for numerical.
- **Classification**: Classification analyzes a set of training data (a set of object whose class label is known) and constructs a model for each class based on the features in the data. A decision tree or set of classification rule is generated by such a classification process. There have been many classification method developed in the field of machine learning, static, database, neural network.
- **Prediction:** This mining function predicts the possible value of some missing data and the value distribution of certain attributes in a set of objects. It involve the finding of set of attributes relevant of the attribute of interest and predicting the value distribution based on set of data similar to select object.
- **Clustering:** Clustering analysis is tool identify clusters embedded in data where a cluster is a collection of data object that is similar to one another. Similarity can be specified by user of experts.

- **Time Series Analysis:** Time series analysis is to analyze large set of time series data to find certain regularities and interesting characteristics, including search for similar sequences and sub sequences, mining sequential patterns, periodicity, trends and deviation.

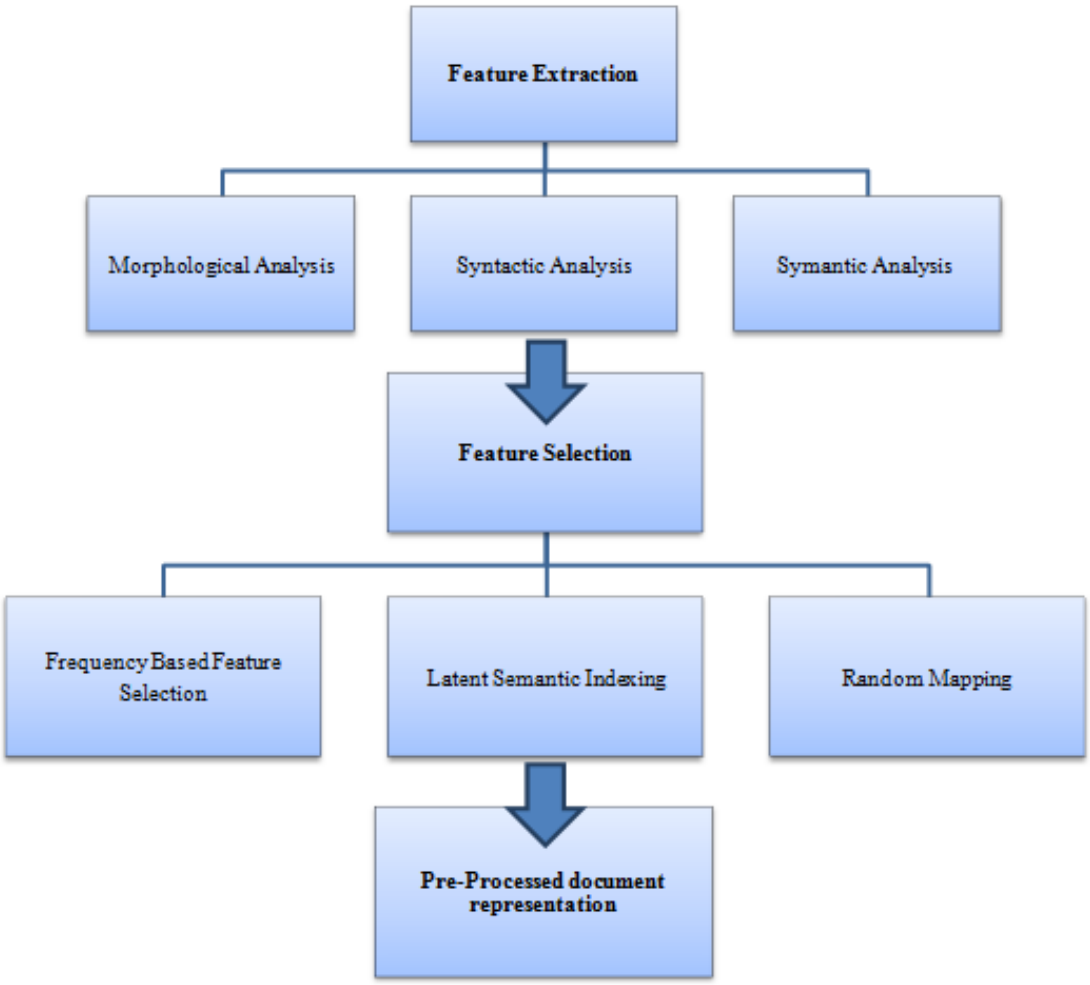

Figure 2: Basic structure of the lexicon analytical engine for the keyword analysis and extraction

II. LITERATURE REVIEW

Denial I.Morariu, Lucian N. Vintan, and Volker Tresp [1] investigated three approaches to build an efficient meta-classifier. In this select 8 different SVM Classifiers. For each of the classifier modified the kernel, the degree of the kernel and input data representation based on the selected classifier calculate the upper limit of our meta- classifier that is 94.21 %. Compare one simple static model based on majority vote with two adaptive methods. With majority vote the classification accuracy was 86.38%. as we expected, the documents that are correctly classified by only one classifier can't correctly classified by this method. The SBED method obtains best results, growing upto 92.04% after 14 learning steps with 2.17% smaller than the upper limit. Also, this method is the fastest one because it selects the first acceptable classifier and because the computation cost is lowers. The last method (SBCOS) is the most rigorous one because it finds the best component classifier. As a consequence, the training time for SBCOS is longer at an average of 21 minutes comparatively with SBED. The goal of ongoing work is to classify larger text data sets. Also want to develop a pre classification of all documents, obtaining fewer samples. After that use the obtained samples as entry vectors for the already developed features selection and classification for web mining applications, in order to extract and categorized online news.

Hyeran Byun 1 and Seong-Whan Lee2 [2] present brief introduction is presented on SVMs and several applications of SVMs in pattern recognition problems. SVMs have been successfully applied to a number of applications ranging from face detection and recognition, object detection and recognition, handwritten character and digit recognition, speaker and speech recognition, information and image retrieval, prediction and etc because they have yielded excellent generalization performance on many statistical problems without any prior knowledge and when the dimension of input space is very high but did not compare the performance results for same application.

D. Morariu, R. Cre,Tulescu and L.,Vin,tan [3] says that building up on the meta- classifier presented based on 8 SVM components, we add to these a new bayes type classifier which leads to a significant improvement of the upper that the meta classifier can reach. Thus, the meta- classifier upper limit has increased from 94.21% **when using** 8 SVM classifier to 98.63% when using the 8 SVM classifiers plus the Bayes classifier. Moreover, in the case of the 9- classifier SBED meta- classifier we obtain even lower results, on average dropping from 92.04% to 90.38%. in the case of 9-classifiers, SBCOS, the classification accuracy of the meta- classifier has increased from 89.74% to 93.10%.

Lie Lu, Stan Z. Li and Hong –Jiang Zhang [4] presented in detail our approach that uses SVM for classification and segmentation of an audio clip. The proposed approach classifies audio clips into one of five classes: Pure speech, Music, Environment sounds and silence. We have also proposed a set of new features to represent a one second sub clip, including band periodicity**, LSP divergence shape and spectrum flux.** The experimental evaluation have shown that the SVM method yields high accuracy and with high processing speed. We are extending this work to incorporate visual information to help video content analysis; the result is also very satisfying.

III. PROBLEM DEFINITION

Researchers have done a lot for the text classification in online news classification but they have least about the sports category there is no classification of e-Sports, e-bollywood and e-matrimonial news. So it is time consuming task to select the most interesting one as there is proper classification of news articles. This categorization is essential to obtain the relevant information quickly. For this use texts from data source are testing set than compute the classification results and will compare those values with real values to check the accuracy. We hope to get best results with more accuracy and less time consumption.

IV. CONCLUSION

The purpose of this work is to modify the current evaluation techniques of the classification of the online news and to make the inner cluster so that the better efficient algorithm can be generated to reduce the burden of the manual system of data entry of online news classification. The purpose also involves checking out for better accuracy of the implemented technique so that the future researches get a change to enhance the modified results. In the work till now the successful implementation has been done to extract the news from the online portal for the further processing. In addition to this, the clusters of different categories hs been also created so that the further combination of HMM and SVM could be applied to it to regain the better efficiency. In future the work can be made to the inner cluster of the main cluster till now.

REFERENCES

[1] Denial I.Morariu, Lucian N. Vintan, and Volker Tresp, "Meta- Clalssification using SVM classifiers for text documents ", "World Academy of science engineering and technology" 21,2006.

[2] Hyeran Byun 1 and Seong-Whan Lee2, " Application 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 the SVM Meta Clssifier for text document by using Naïve bayes,"Int. J. of Computers, communication and control, ISSN 1841-9844.

[4] Lie Lu, Stan Z. Li and Hong –Jiang Zhang, " Content based Audion Segmentation using Support vector machine."

[5] Krishnlal 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) Volumn 1- No.19,2010.

[6] Vandana Korde, C namrata Mahender, "Text classification and classifier a survey," International Journal of Artificial Intelligence and Application (IJAIA), vol.3, No.2, March2012.

[7] Mita K. Dalal, Mukesh A.Zaveri," Automatic text Classification," International Journal of Computer Applications (0975-8887) Volumn 28- No.2, August 2011.

