Hybrid Semantic Analysis of Tweets: A Case Study of Tweets on Girl-Child in India

Mani Madhukar
IBM India Pvt Ltd
Bangalore, India
manimmadhukar@gmail.com

Seema Verma
Banasthali University
Rajasthan, India
seemaverma3@yahoo.com

Abstract—**Social networks have become one of the major and important parts of daily life. Besides sharing ones views the social networking sites can also be very efficiently used to judge the behavior and attitude of individuals towards the posts. Analysis of the mood of public on a particular social issue can be judged by several methods. Analysis of the society mood towards any particular news in form of tweets is investigated in this paper. The key objective behind this research is to increase the accuracy and effectiveness of the classification by the process of Natural Language Processing (NLP) Techniques while focusing on semantics and World Sense Disambiguation. The process of classification includes the combination of the effect of various independent classifiers on one particular classification problem. The data that is available in the form of tweets on twitter can easily frame the insight of the public attitude towards the particular tweet. The proposed work implements a hybrid method that includes Hybrid K, clustering and boosting. A comparison of this scheme versus a K-means/SVM approach is provided. Results are shown and discussed.**

Keywords-natural language processing (NLP); sentiment analysis; social networking analysis; social networking sites (SNS)

I. INTRODUCTION

With the continuous increase in the use of social networking sites (SNS), it has become quite possible to judge the sentiments and reactions of the public towards any news or tweet posted, a fact with interesting possibilities [1]. Every conversation that occurs in the SNS has several different reactions that depict the effectiveness of the post and help analyze the public sentiment towards it [2]. Sentiment analysis is an interdisciplinary field that crosses natural language processing, artificial intelligence, and text mining. Since most opinions are available in text format and its processing is easier than other formats, sentiment analysis has emerged as a subfield of text mining [3]. Sentiment analysis appeared in the literature in 90's for the first time and it became a major research topic in 00's. Classifying the polarity of a given text as positive or negative is the basic task of sentiment analysis. Due to its many aspects it is often referred to with different names such as opinion mining, sentiment classification, sentiment analysis, and sentiment extraction. It is widely believed that sentiment analysis is needed and useful while it is also widely accepted that extracting sentiment from text is a hard semantic problem even for human beings. Additionally, sentiment analysis is domain specific, therefore the polarity of some terms depends on the context in which they are used. For example, while "small" for "size" as a feature in the electronic products is positive, in agricultural products such as fruit it has a negative polarity. Sentiment analysis is used in different domains such as shopping, entertainment, politics, education, marketing, and research and development. There are several sentiments of the public towards the SNS posts that include: positive and negative along with the n-point scale that includes very good, good, satisfactory, bad and very bad [3]. Text mining is the famous way to analyze and understand the sentiment of people integrated with the content posted and its methods are: Machine Learning, Statistical/Quantitative Techniques or Natural Language Processing [4]. The sentiment analysis is of two kinds, supervised or unsupervised. In [5] a method to apply an approach of MLT, based on Maximal Discrepancy concept, to the problem of SVM model selection has been detailed. Sentiment analysis approaches are reported in [6-8] and an overall review can be found in [9]. Social media insights and especially twitter are also essential [10-13].This paper, following [14], focuses on sentiment classification in social issues, thus on determining the public's reaction over a particular tweet in order to get the proper assistance to determine the public reaction towards the news and its effects.

II. IMPLEMENTIG A HYBRID APPROACH

Social Issues Sentiment Analysis automatically analyzes social issues. It identifies the positive, negative or neutral opinion. The conceptual framework comprises of four stages: Data collection and cleaning; preprocessing; sentiment analysis and finally experiments and results. The architecture of the proposed framework for sentiment analysis is presented in Figure 1.

A. Text Collection and Cleaning Stage

For the analysis of social issues tweets, data is collected. The input data will be raw text from tweets on social causes in India, in particular on 'JNU agitation', "Intolerance in India". The motivation for the topics has been derived from the a web interview of Twitter India Director in 2016, as the topics had the power to polarize the entire country and to shape opinion of common countrymen leading to a sharp divide in Indian

society. For creating the corpus of tweets, the tweets are fetched from the Twitter database based on HashTags (#), using Twitter API for connecting and authenticating. The collected text is noisy and cleaning and parsing the data to form a corpus for further processing is required.

B. Preprocessing Stage

At this stage, the corpus is transformed into feature vectors. Strings are converted into words using some filtration techniques. We adapted a simple feature selection as a pre-processing method to transform or tokenize the text stream to words. These methods constitute a sequence of the following asks: removing delimiters, removing numbers and stop words. For stop words removal, a list of stop words is provided in the filtration process.

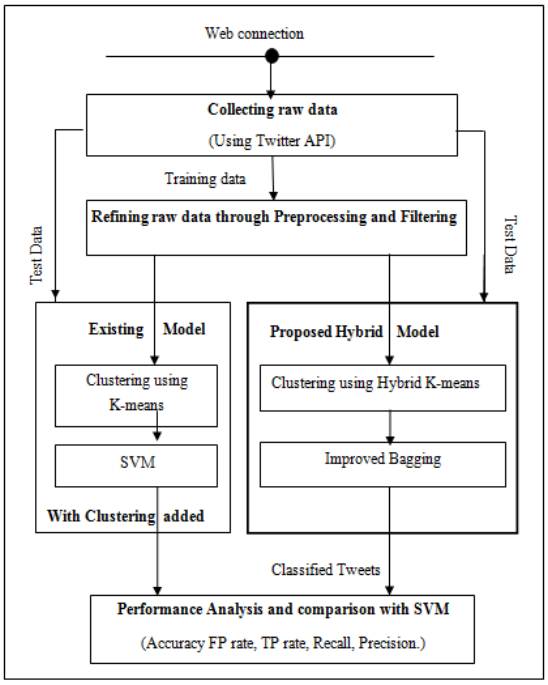

Fig. 1. Architecture of the Proposed Methodology

C. Sentiment Analysis stage

This stage of framework handles the polarity measurement and sentiments. We approach these tasks by employing the following hybrid machine learning methods. In this approach, we have combined clustering with classification. In clustering, hybrid clustering approach is used which combines the feature of k-modes and k-medoid algorithms. The functionality of the Hybrid Clustering is as follows:

- The size of cluster is fixed and the output of the first phase forms initial clusters. Here, the input array of elements is scanned and split up into sub-arrays, which represent the initial clusters.
- The cluster sizes vary and the output of this phase are the finalized clusters. Initial clusters are inputs for this phase. The centroids of these initial clusters are computed first, on the basis of which distance from other data elements are calculated. Furthermore, the data elements having less or equal distance remains in the same cluster otherwise they are moved to appropriate clusters. The entire process continues until no changes in the clusters are detected.
- For classification, Improved Bagging is used. Classifying the clustered data is performed by using improved bagging technique which decreases the variance of the prediction using dataset using combinations with repetitions to produce multi-sets of same size of the dataset. For each multi set the Boosting learning algorithm is applied to classify the instances and a model is created and a vote related to that model is generated. The average of all the predicted votes is considered to be the result of the classifier.

III. EXPERIMENTAL RESULTS AND DISCUSIONS

A. Performance Parameters

- Accuracy: Accuracy is the percentage of testing set examples correctly classified by the classifier. It is the proportion of total number of predictions that are correctly classified in class.

$$\text{Accuracy} = (TP + TN) / (TP + FP + TN + FN) \quad (1)$$

where, TP is True Positive, TN is True Negative, FP is False Positive, FN is False Negative

- Precision: Percentage of selected instances that are relevant and are correctly classified in class out of all documents in class.

$$\text{Precision} = (TP) / (TP + FP) \quad (2)$$

- Recall: Percentage of correct documents that are selected in class from the entire document actually belonging to class.

$$\text{Recall} = (TP) / (TP + FN) \quad (3)$$

- Confusion Matrix: Also, known as contingency table or error matrix in supervised learning and in unsupervised learning is called matching matrix. In confusion matrix, ROC is used to plot graph. Various measures could be defined basing on values in a confusion matrix.
- True Positive (TP) and False Positive (FP) Rate: For multiple comparisons TP and FP is used and it is a type of error. TP is also called Sensitivity as if a person has a disease how often will the test be positive is referred to as true positive rate. FP is an error in a test result indicates presence of a condition.

$$TPR = (TP) / (TP + FN) \quad (4)$$

$$FPR = (FP) / (FP + TN) \quad (5)$$

- F-measure: A measure that combines precision and recall.

$$\text{F-measure} = 2*(\text{Precision} * \text{recall}) / \text{Precision} + \text{Recall} \quad (6)$$

So, considering every 6th frame leads to good enough results and speeds up the task. The output images of these are

then fused together for the final text detection. Thus, fusion results obtained are more informative when every 6th frame is fused together. Figures 2-4 shows the comparison graphs of TP rate and FP rate and Precision Recall and F Measure of proposed technique with previous techniques.

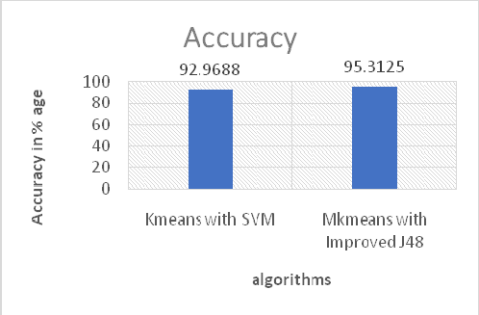

Fig. 2. Classification cccuracy comparison

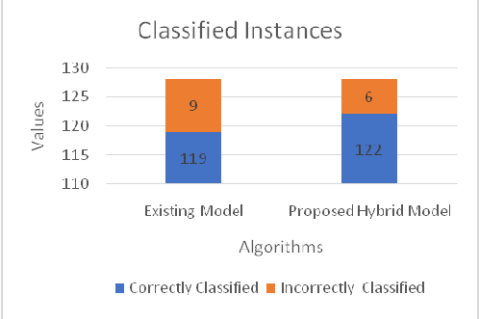

Fig. 3. Number of classified instances comparison.

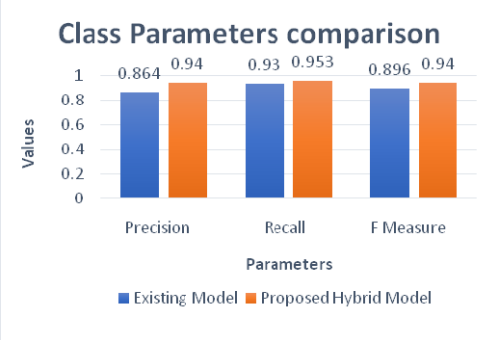

Fig. 4. Precision Recall and F Measure comparison.

IV. CONCLUSION AND FUTURE SCOPE

Sentiment analysis, as an interdisciplinary field that crosses natural language processing, artificial intelligence, and text mining, recognizes opinions of people regarding a product, service, object, or social issues expressed in a given text. Up until recently most, if not all, research in sentiment analysis has been done on the products and services. Determining public opinions regarding social issues is important for governance and decision making. This paper has focused on sentiment analysis of social issues. A new hybrid method for sentiment analysis of social issues is proposed. It extracts the opinions from each sentence, constructs correspondence opinion structures, and then determines their orientations regarding the social issue. This algorithm performs better in comparison with a previous k-means/SVM approach. In the future, the work can be extended by containing a larger dataset and considering more instances which may cooperate in higher accurate prediction analysis.

REFERENCES

[1] A. Shaikh, T. Pritam, P. Ankita, W. Shital, T. Pooja, "Stock Exchange Market Prediction", International Journal of Advances in Computer Science and Technology, Vol. 3, No. 5, pp. 349-351, 2014

[2] A. Joshi, A. R. Balamurali, P. Bhattacharyya, R. Mohanty, "C-Feel-It: A Sentiment Analyzer for Micro-blog", HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations, pp. 127-132, Portland, Oregon, June 21, 2011

[3] A. Abbasi, "Intelligent feature selection for opinion classification", IEEE Intelligent Systems, Vol. 25, No. 4, pp. 75-79, 2010

[4] S. Haykin, Neural Networks: A Comprehensive Foundation, Prentice Hall, 1998

[5] D. Anguita, A. Ghio, N. Greco, L. Oneto, S. Ridella, "Model Selection for Support Vector Machines: Advantages and Disadvantages of the Machine Learning Theory", Proc. of the International Joint Conference on Neural Networks (IJCNN), Barcelona, Spain, 2010

[6] D. Rajput, M. Madhukar, S. Verma, M. Sharma, "Sentiment Analysis on Big Data using Machine Learning for Holiday Destinations", 2015 IEEE European Modelling Symposium , Spain, October 6-8, 2015

[7] A. Kumar, T. Mary, "Sentiment Analysis: A Perspective on its Past, Present and Future", International Journal of Intelligent Systems and Applications, Vol. 10, pp.1-14, 2012

[8] A. Somla, S. V. N. Vishwanathan, Introduction to Machine Learning, Cambridge University Press, 2009

[9] A. M. Kaplan, M. Haenlein, "Users of the World, Unite! The Challenges and Opportunities of Social Media", Business Horizons, Vol. 5, No. 1, pp. 59-68, 2010

[10] N. Anitha, B. Anitha, S. Pradeepa, "Sentiment Classification Approaches", International Journal of Innovation Engineering and Technology, Vol. 3, No. 1, pp. 22-31, 2013

[11] B. J. Jansen, M. Zhang, K. Sobel, A. Chowdury, "Twitter Power: Tweets as Electronic Word of Mouth", Journal of The American Society for Information Science and Technology, Vol. 60, No. 11,pp. 2169–2188, 2009

[12] B. Pang, L. Lee, S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques", Proceedings of the Conference on Empirical Methods in Natural Language Processing, Philadelphia, pp. 77-86, 2002

[13] C. Vicient, A. Moreno, "Unsupervised Topic Discovery in micro-blogging networks", Expert Systems with Applications, Vol. 42, pp. 6472–6485, 2015

[14] S. Verma, M. Sharma, D. Rajput, M. Madhukar, V. Mittal, R. Singh, "Disclosing Tweet Polarity using feature representation factor", International Journal Of Latest Trends In Engineering and Technology, Vol. 5, No. 2, 2015