LABELLING OF SHORT MESSAGES IN TWITTER FOR CHARACTER ANALYSIS-A STUDY

Anjana P[1], Syam Sankar[2]
Department of Computer Science & Engineering,
NSS College of Engineering Palakkad, Kerala

Abstract

Microblogging has become an essential tool for internet users. Most of them are members of any one of the social media networking sites. So the peoples are connected with the world and the world is on the finger tip. Twitter, Facebook etc. are the very popular social networking sites. This paper proposes a method to process Twitter data (tweets) for the purpose of categorizing the account holder's character and behavior from their tweets. It is done by assigning predefined labels into tweets that matches the meaning of the corresponding label and thus analyzing interested area of a particular person or the user. K-means clustering algorithm is used in the formation of clusters of datasets (tweets) in the training phase. The system uses four labels such as "normal", "political", "sports" and "technological". The departments like cyber forensic can make use of this system for analyzing the comments or opinion in social networking sites by setting appropriate datasets.

Index Terms: Classification, Clustering, Information Retrieval, tweet extraction.

I. INTRODUCTION

Twitter is one of the most popular microblogging sites. Through social media, a user can exchange the information like news, trending topics and can post about anything. The user's tweet (Twitter short message) is 140 characters in length and the tweet of a particular person can be viewed only by the followers of that person [1]. The web-based technology is the working principle of social media networking sites. The impact of social media sites is growing day by day. As per a recent survey, about 500 millions of tweets are being generated per day. The social media sites give substantial changes to communities and individuals and these media change the way peoples and organizations communicate.

In this paper, a system model to categorize tweets into different labels or assign labels into tweets is discussed. There are so many works currently going on related to twitter sentiment analysis. Twitter data is used for the purpose of sentiment analysis by many researchers. Millions of peoples are sharing their views on various aspects of life every day. Therefore social media websites are the large big source for opinion mining [2], [3].Clustering methods are used in the system's training phase. A large set of tweets (dataset) are taken and are clustered based on the labels.

II. RELATED WORKS

Tweet classification approaches are under the area of machine learning [4]. Machine learning approaches can be divided into two categories: supervised machine learning and unsupervised machine learning techniques. In the machine learning method, natural language processing techniques play the most relevant position. Support vector machines (SVM), Naive Bayes, Maximum Entropy are some of the most common techniques used in supervised machine learning approach.

The semantic orientation technique performs classification based on the meaning of words contained in each text [5]. Semantic orientation does not need any crucial training method to mine the data. The sentiment analysis is mainly

classified into two categories such as corpus-based and lexicon based methods [6].
Network-based text classification checks for the topic Classification by decision tree method [7].There are so many methods proposed for Twitter data classification. Many researchers use movie review dataset [8] for sentiment classification. Text mining or data mining is the process of extracting high quality and useful information from text [9].

III. SYSTEM ARCHITECTURE

This paper proposes a model for labelling the twitter data or twitter short messages. This system will reduce the data sparseness and make the accurate prediction on tweet dataset. The proposed system contains mainly two phases: training phase and testing phase.

A. Training Phase

The first phase is the training phase. It uses the training dataset from Twitter and it will preprocess the training data. The preprocessing includes stemming and stop word removal. In stemming, it will process a word in its base form For example; the word "play" is the base form of words like “playing”, “played”. Next is the stop word removal in which it removes the stop words like “is”, “was”, etc., on the training dataset. In the word-to-vector model, it trains N-dimensional word vectors of documents based on a given corpus. The neural network is just used to aggregate the dimensions of the document vector and also capturing some relationship between words. But what should be mentioned is that this is not really semantically related, it just reflects the structural relationship in our training body. Then we will use K-means clustering algorithm to cluster the data with respect to word clusters which include Normal, Technological, Sports and Political category based tweets [10]. These clusters are used for labelling the twitter short messages. Mainly four labels are used: Normal, technological, Political and Sports. These labels simply mean the category name.
The K-means algorithm works based on the operation of calculating the Euclidean distance between the sample vectors.

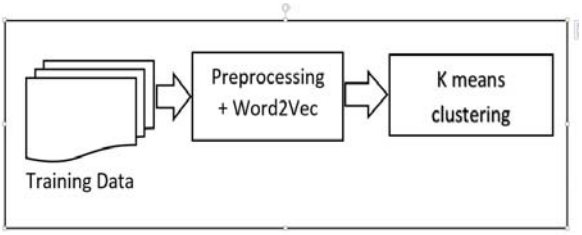

Fig.1. Training Phase

The Fig.1 represents the training phase of proposed approach. It uses the K-means algorithm for forming clusters of tweets and each cluster corresponds to a label. The word to vector and K-means scheme improved the accuracy.

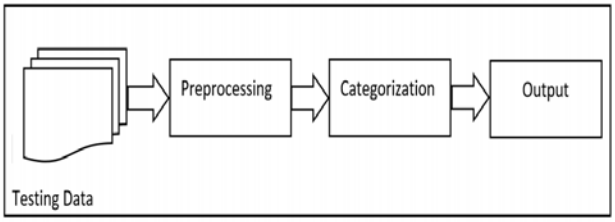

Fig.2. Testing Phase

B. Testing Phase

The second is the testing phase. In the testing phase, the user can give an input data into the model and it will preprocess it and categorize it into the appropriate category and produces the labels for the given input. The Fig.2 represents the testing phase in detail.

The Fig.3 shows the overall system model. In training phase it will make the cluster of training data, here uses the tweet dataset for training purpose. This cluster contains the processed data. The proposed system uses the sports, technological and political tweet dataset for the training purpose. So the output of training phase will be a model. In testing phase it extract 10 tweets from twitter account of a person and preprocessing it and then it will compare with the trained model. The output of testing phase is tweets with the appropriate category label.
Then the next step is to plot the graph. The proposed work coding is done in python. Graph plotting is done with the help of matplotlib library. Graph plotting is done by taking the tweet data as input and processing it and finds out the category it belonging.
For example, a test file contains three tweets two of them are in sports category and one is in technology category. The initial value of

category count will be set to zero and category count will incremented by checking the category of test tweet.

In this proposed system, user can type a tweet to find out the tweet category and can also give a text file for processing many numbers of tweets.

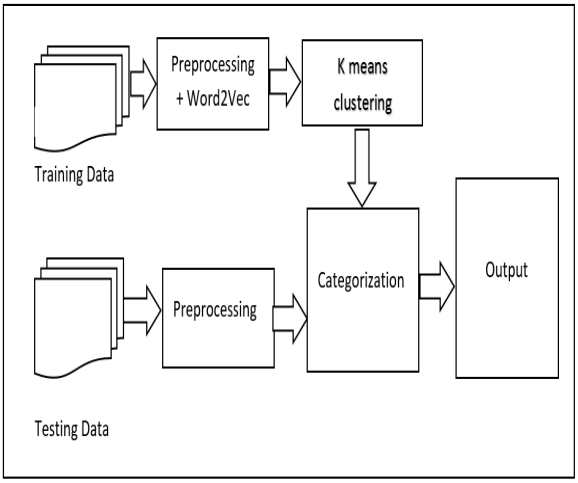

Fig.3. Overall System architecture

IV. RESULTS AND ANALYSIS

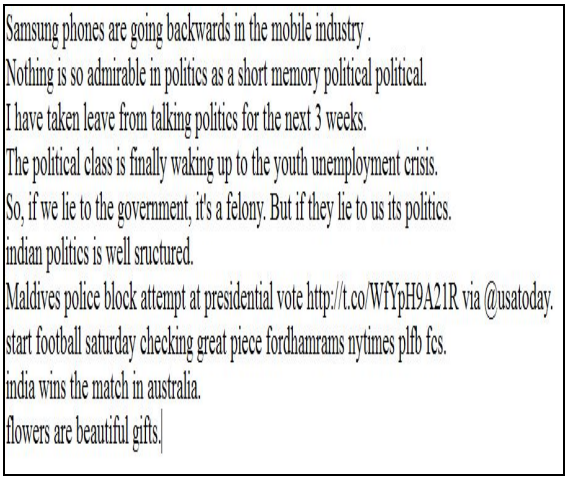

Fig.4. Input Data

The Fig.4 represents the input data i.e., it is given as an input to the system. This text file contains ten tweets that are extracted from a particular person's twitter account. Then that extracted tweet is saved into a text file i.e.Test_Data.txt. This text file is given to the system.

The Fig.5 represents the system output for the above mentioned input i.e. Test_Data.txt. The system uses the word to vector model to increase the accuracy of the result. The graph represents the output for the given input data. The given input data contains ten tweets and each of them is mapped into appropriate label clusters. Here we used four clusters: politics, normal, technological and sports. From the graph it is clear that the particular person's most of the tweet is related to political category. Also from the graph, from ten extracted tweets six of them are related to political category, two tweets on sports category and one for technology and normal category. So it can be concluded that his/her social media characteristics is more interested to tweet on politics. So the system, by taking his/her social media tweets predicts the type of tweets he used to make frequently and thus analyzing his/her behavior.

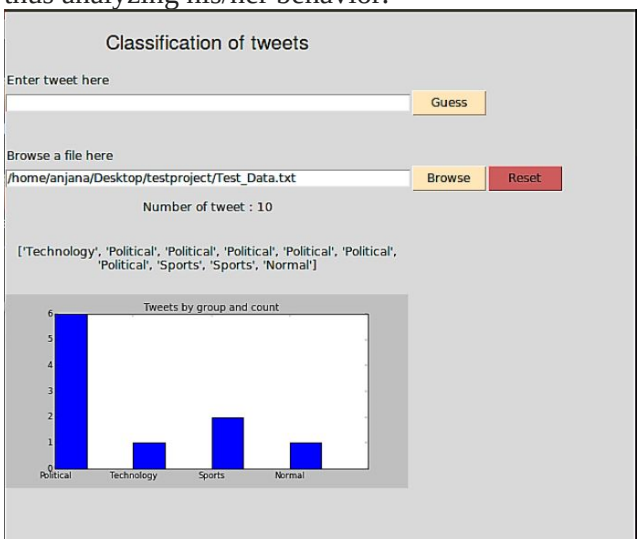

Fig.5. System Output

Consider another test input (Fig.6) and the corresponding output (Fig.7).It shows that the tweets are from a person who makes technology related tweets frequently

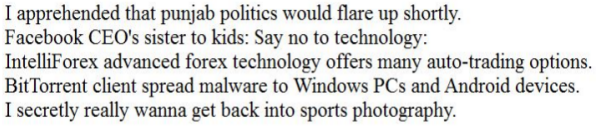

Fig.6. Input Data 1

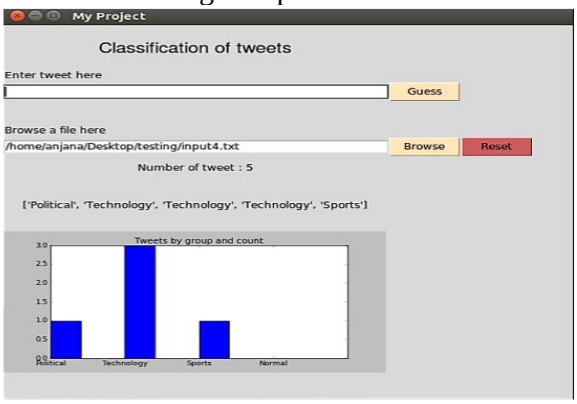

Fig.7. System Output 1

V. CONCLUSION

This paper introduced a labelling approach on Twitter message. The sparseness of twitter data is high so labelling of twitter data is more difficult. The proposed method consists of mainly training phase and testing phase. In training phase, it will train the tweet dataset and create a model for labelling and in the testing phase it will check the model with the test data to produce the output. This proposed method can be used for analysing the character of a particular person from their tweet i.e. mainly the area of interest of that particular person. Political, technological, sports and normal are the four labels used here. The system works well on the tweet data. We can improve the proposed system or method by using appropriate datasets.

.

REFERENCES

[1] B. Krishnamurthy, P. Gill, and M. Arlitt, "A few chirps about Twitter," in Proceedings of the First Workshop on Online Social Networks, 2008, pp. 19–24.

[2] DiMicco, D. Millen, W. Geyer, C. Dugan, B. Brownholtz, and M. Muller, "Motivations for social networking at work," in Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work (CSCW '08),2008, pp. 711–720.

[3] A. Java, X. Song, T. Finin, and B. Tseng, "Why we Twitter: Understanding microblogging usage and communities," in Proceedings of the Ninth WEBKDD and First SNA–KDD Workshop on Web Mining and Social Network Analysis,2007, pp. 56–65.

[4] S. Kinsella, A. Passant, and J. G. Breslin, "Topic classification in social media using metadata from hyperlinked objects," in Proceedings of the 33rd European conference on Advances in information retrieval , 2011, pp. 201–206.

[5] Y. S. Yegin Genc and J. V. Nickerson, "Discovering context: Classifying tweets through a semantic transform based on wikipedia," in Proceedings of HCI International, 2011.

[6] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," 2009.

[7] M. N. Hila Becker and L. Gravano, "Beyond trending topics: Real-world event identification on twitter," in Proceedings of AAAI , 2011.

[8] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?. sentiment classification using machine learning techniques," in Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, Association for Computational Linguistics, 2002, pp. 79–86.

[9] R. Narayanan, "Mining Text for Relationship Extraction and Sentiment Analysis," Ph.D. dissertation, 2010.

[10] C. D. Manning, P. Raghavan, and H. Schtze, "Introduction to Information Retrieval," New York, NY, USA: Cambridge University Press, 2008.