

DETECTING EMERGING TOPICS IN TWITTER USING CONCEPT LEVEL SENTIMENT ANALYSIS

¹J.Prathipa, ²Avish Dhawan, ³Sajja Balasankar, ⁴Sindhu C

Department of Computer Science and Engineering

SRM Institute of Science and Technology

¹prathipavadivelan@gmail.com

Abstract: The prominence of social media and especially Twitter is well known to all of us as it comprises the public opinion. In this, we concentrate on extracting the tweets, applying them to pre-processors and quantifying the tweets by assigning certain values through which emerging topics will be retrieved. For the following process, different data mining techniques like third party authentication, clustering and topic based modelling is used. The model shows how the particular tweet have trended and how much the particular tweet has been preferred by the public. Using the aggregation we determine the in and out time of the tweet in the TREND column.

Keywords: Clustering, Classification, Topic based Modelling, Data Mining, Prediction.

1. Introduction

In this era, Social media has played a vital role in decision making in every aspect. For example-predicting election results and the resulting outcomes to provide a user's point of view. A prerequisite is required to provide an accurate estimation of the popularity of the particular leader. If not, the result will be invalid. A variety of methods has been proposed to understand and remove biasing. Some methods involves connecting media links to certain reference points. In context of twitter, we faced a problem i.e. is it possible to assign a particular value to a particular tweet or not. But the problem is solved using some aspects of the twitter:

TOPIC SKETCH: It basically displays the tweet topic extracted from twitter and the URL of that particular tweet. **AGGREGATION:** This is used to show the in-time and the out-time of the tweet from the trending column.

BURST ANALYSIS: It differentiates what the difference was between clustering based detection and topic modelling based detection in terms of time

i.e. how long a particular tweet was trending in both the detection modes.

PERFORM TRAINING: Choosing a particular tweet and thereafter shows the users who have tweeted something regarding the tweet (in clustering) and shows an additional count of retweet in the case of Topic based Modelling detection.

CHANGE POINT ANALYSIS: It gives the total count of people who have tweeted something regarding the selected tweet and the time for which the particular topic is trending.

Certain good and bad keywords have been assigned a particular value and based on the content of the tweet, that particular tweet is considered as a positive/negative/neutral tweet. Our requirement is only a number of tweets and after pre-processing a simple score is interpreted. Score is taken as average of positive, negative and neutral tweets. The real point of the field of work can be outlined as taking after: To pick up the data from twitter; Pre-processing of information i.e. assigning particular values to certain keywords of the tweets and displaying the burst analysis about that particular tweet; Analyzing and Prediction of understudy execution in early stage through the information mining systems which gives the strategy to basic leadership; To bring ongoing execution forecast arrangements at whatever time, anyplace. **CLUSTERING:** It extracts data from a pile of unordered data, resulting it into groups of objects.

2. Literature Review

Many researchers have proposed various models for predicting student's performance.

[6] M. D. Conover and group developed a model in which prediction of the political of twitter users was done. It came up with 91% accuracy because of opposing political signals, differences in opinion and it used a data set of

1000 individuals and supports SVM. The structure isn't provided but is identified. Practical application is done on web based political advertisement.

[7] S. Golder and group proposed a framework in which the practice of re-tweeting is considered in such a way like it is a conversation because retweeting is considered like a comment to the tweet. It is fed with case studies and empirical data. As it has a limit up to which infers 140 characters, so it is convenient also.

[4] M. Pennacchiotti and group proposed a machine learning approach infers attribute of the user like user's orientation judged by the user behavior. Basically 3 tasks i.e. Affiliation detection, affinity detection, and ethnicity identification is done. It relying on set of derived user information.

[10] d. boyd and group Built a framework in which the sentiment detected in the dataset are compared and estimation of political popularity is done. If out of the box tools is preferred, it result in poor performance. So instead multi- stage approach for general purpose classification is used.

3. Proposed Methodology

In this, we propose a framework to screen tweets and to display what are the views of the public regarding the person. Facilitate we extend the current framework to foresee election polls using different set of dictionaries and keywords that are provided by the user.

4. Future Term Grade Forecast

The above figure shows how the whole model i.e. from data retrieving from twitter till providing the prediction survey and showing what's trending. There is a different set of files (dictionary) which is compared with the scrutinized set of keywords from tweets. Particular values are assigned to match and a tweet is declared as a positive/negative/neutral tweet. Choosing a particular tweet in the topic based modelling detection, the number of retweets can be retrieved.

As of now the existing system is considering on displaying only the positive, negative, neutral tweets. The future advancement maybe comparison on basis of different domain. We are not having a framework which would help us to incorporate the execution and undesirable into thought. The current system doesn't have a static database to display the comparison between keywords. It doesn't understand the positive negation tweets for the decision making.

Proposed System

Our plans are to build an accurate model describing views utilizing data mining strategies which mines data.

It utilizes the algorithm to extract the tweets and allows pre- processing and displaying the according data. The model may help the leader to have a chance of either increasing his/her likelihood (while positive and negative response).

Firstly the tweets are scrutinized using hashtags. Then the user mines the keywords from the tweets and compares with the dictionary.

Polarity re-tweet

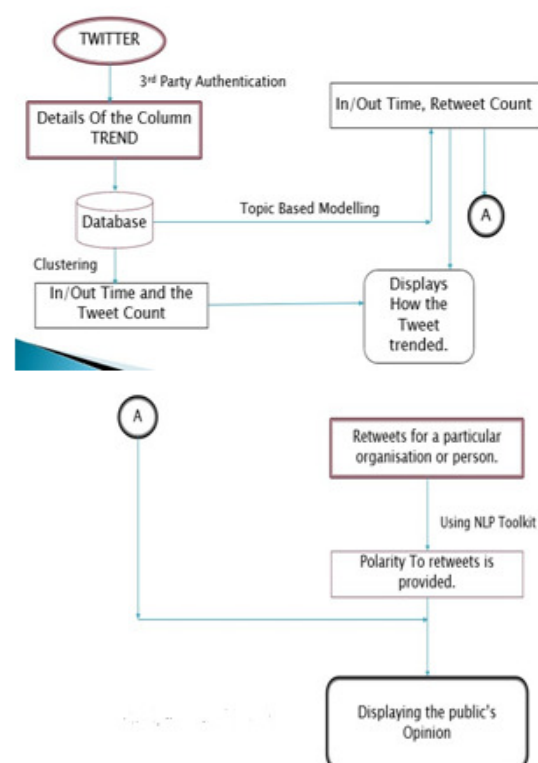


Figure 1. System Architecture

These techniques are meant to mine the data and display the evaluation of the tweets on the basis positive, negative and neutral responses and suggest what actions have to be taken by the leader to improve his likelihood. Showing the results may reflect the increase/decrease of demand of the person by considering a month before result and a current result, permitting to maintain or improve their reputation. The number of retweet on a particular tweet additionally offer data with positivity or negativity of the output, which helps us to display an accurate result and handle their time. Besides, the regular liberal tweeters can likely profit by knowing that the amount of retweets is considered as well and the

importance of their tweets and thus empowering them to make according modification etcetera.

A. Data Preparation

This procedure signifies the ways or the steps how to extract data from the twitter. The follows steps were conducted to extract the data and perform pre-processing: Download the file trends.txt and search.txt from which the tweets are confined on the basis of hashtags. Categorize the number of good words and number of bad word in a particular tweet and assign them values accordingly. Classify the tweets on the basis of total number of good words and number of bad words as a positive/negative or neutral tweet.

B. Data Pre-Processing

The method of data mining consists of three stages:

Stage 1: This stage involves the information arrangement which includes extracting and comparing keywords and clearing the leftover i.e. choosing a subset of it is considered and the other nouns are not. When the model is compiled, the fresh data is considered with the help of trends.txt and data extraction and evaluation is considered on the latest tweets done on twitter. A certain time limit has been set so that there is not a huge pile of data to pre- process and hence decreasing the time of compilation.

Stage 2: This module comprises of a pre-existing dictionary of good and bad words which when compared with the words of the tweets are programmed to evaluate and categorize a particular tweet. There are scope of systems created to apply distinctive models to the indistinguishable informational index and after that contrasting their accomplishment with select the best. However, comparing words with a positive negation may result in opposite manner and may need a horrendously muddled strategy.

Stage 3: The progression includes choosing the scrutinized tweets and extracting data and applying it to the opinion poll to achieve a forecast or an accurate prediction.

C. Clustering and Topic Based Modeling Detection

To categorize appositve, negative or a neutral tweet, natural language processing is being used. It is programmed in such a way that the documents are fed into the algorithm and stores the result in a manner such that it is easy to understand for the user. There are two techniques used for pre-processing: Clustering and topic based modelling detection. Below are provided the list of

formulas used by the clustering technique for data extraction and pre-processing.

Keyword Spotting is the simplest among all for concept based analysis. Algorithm run in such a way that it indicates whether the tweet is a positive sentiment or a negative sentiment. Dictionaries are used here to compare the keywords with the data mined. Theses dictionary are also called "bag of words". For example- Good words- happy, good;

Bad words-terrible, sad.

In the formulae,

N-number of clusters,

Cx- centroid of cluster,

sie x- average distance,

d- Distance between centroids.

$$DB = \frac{1}{n} \sum_{i=1}^n \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right)$$

a. Decision Table

It is a way or method to decide that include extracting tweets and mining data and declaring a particular tweet as positive, negative or neutral tweet. Every tweet is assigned a value based on the number of good and bad words in it. When multiple number of tweets are received, a fair result is being displayed showing the public opinion or view regarding the particular person. Accordingly the person can take steps.

b. Dataset

The initial phase of the model is to gather tweets from twitter with the help of the file 'trends.txt' through third party authentication. On the basis of hashtag politics, various number of tweets are scrutinized and they are taken into consideration. We are taking nearly 100 numbers of tweets of a particular country and for a particular person.

For every tweet different emotions, emoticons, slangs are considered and are compared with the dictionary of good and bad words. The two type of user: liberal (who tweets on a regular basis) and the one who are dormant in nature, both are taken into consideration. On the number of positive and negative count, the tweet is generalized as positive, negative or neutral tweet. The count of all the categories of tweets are displayed using a pie chart.

Evaluation of tweets depends on the nature of tweets i.e. good or bad viewer's point of view for that particular person. The result presented in the model if used can be

used to influence the particular person and can provoke good deeds from him in the future, both from the re-tweeters and by the number of tweets. The result

helps the public as well, regarding how the particular person is working. This result may be very influential for anyone from public before election.

c. Process of Classification

The processed data is been shown showing the positive negative and the neutral tweets in green, red and blue color respectively.

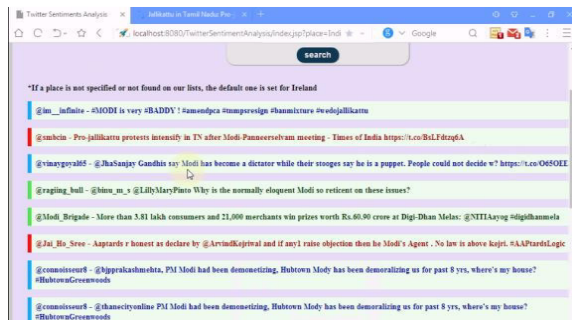


Figure 3. Positive, negative and neutral tweets colored

d. Interface

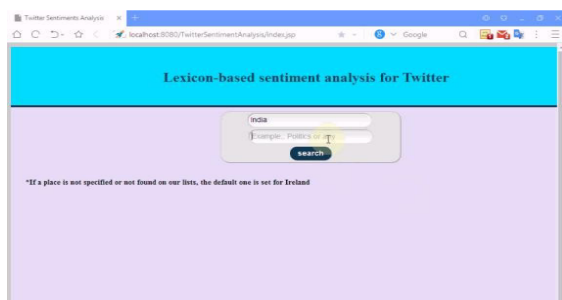


Figure 4: The web page for providing the input.

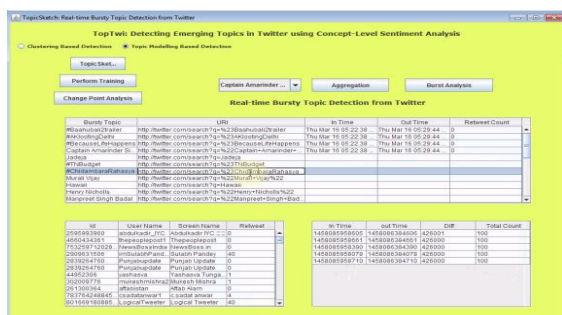


Figure 5: Interface displaying URL of tweets and buttons with different functionality.

5. Discussion

The result shows out of 100 tweets, how many number of tweets are declared as positive or negative or neutral in nature. In the interface using the burst analysis, it is determined that for how the tweet has trended

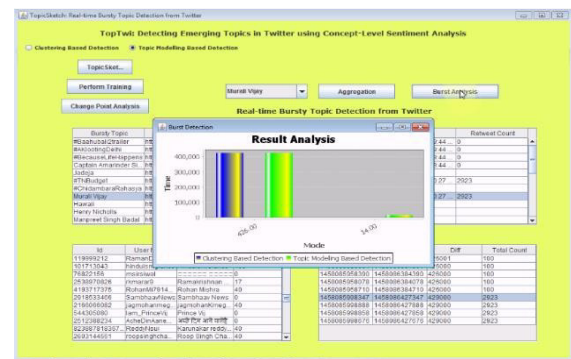


Figure 6: Displaying the burst analysis using different techniques.

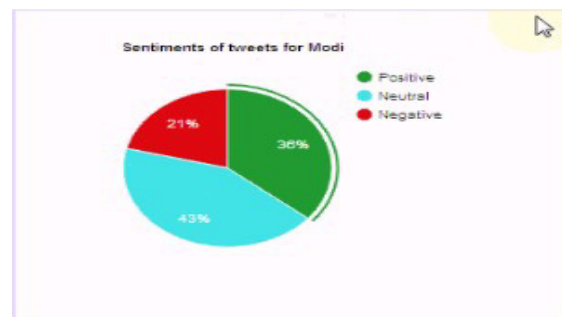


Figure 7. Displaying the genre of tweets received

In this, the tweets are taken into consideration to anticipate the opinion of the public. There are many pre-existing methodologies for information gathering but we have used third party authentication to extract tweets from twitter. Data resembles the tweets, re-tweeters, re-tweets which were gathered using the file trends.txt and search.txt. This model would help the public to have a view on how their leader is getting the response and furthermore it gives the leader a chance to diminish the poor responses he/she is receiving and making suitable move at correct time.

6. Conclusion

This field utilizes the recent tweets and further data processing is done using comparison. There are many Methodology to extract but we have used clustering

techniques and topic based modeling detection Techniques .Tweets were extracted to anticipate the trending ones and to know the public's point of view on it. Even the timing in which the particular tweet was in trends column, is displayed using the button Aggregation.

References

- [1] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 U.S. election: Divided they blog," in Proc. LinkKDD, 2005.
- [2] F. Al Zamal, W. Liu, and D. Ruths, "Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbors," in Proc. ICWSM, 2012.
- [3] J. An, M. Cha, K. P. Gummadi, J. Crowcroft, and D. Quercia, "Visualizing media bias through Twitter," in Proc. ICWSM SocMedNews Workshop, 2012.
- [4] M. Pennacchiotti and A.-M. Popescu, "A machine learning approach to Twitter user classification," in Proc. ICWSM, 2011.
- [5] S. Ansolabehere, R. Lessem, and J. M. Snyder, "The orientation of newspaper endorsements in U.S. elections," *Quarterly Journal of Political Science*, vol. 1, no. 4, pp. 393–404, 2006.
- [6] M. D. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer, "Predicting the political alignment of Twitter users," in Proc. IEEE SocialCom, 2011.
- [7] D. Boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter," in Proc. HICSS, 2010.
- [8] P. Barbera, "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data," *Political Analysis*, 2014.
- [9] A. Boutet, H. Kim, and E. Yoneki, "What's in your tweets? I know who you supported in the UK 2010 general election," in Proc. ICWSM, 2012.
- [10] D. Boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter," in Proc. HICSS, 2010.
- [11] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
- [12] M. Cha, H. Haddadi, F. Benevenuto, and K.P. Gummadi, "Measuring user influence in Twitter: The million follower fallacy," in Proc. ICWSM, 2010.
- [13] J. Clinton, S. Jackman, and D. Rivers, "The statistical analysis of roll call data," *American Political Science Review*, vol. 98, no. 2, pp. 355–370, 2004.
- [14] R. Cohen and D. Ruths, "Classifying political orientation on Twitter: It's not easy!" in Proc. ICWSM, 2013.
- [15] M. D. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer, "Predicting the political alignment of Twitter users," in Proc. IEEE SocialCom, 2011.
- [16] T. Padmapriya, V. Saminadan, "Performance Improvement in long term Evolution-advanced network using multiple input multiple output technique", *Journal of Advanced Research in Dynamical and Control Systems*, Vol. 9, Sp-6, pp: 990-1010, 2017.
- [17] T. Padmapriya, Ms. N. Dhivya, Ms. U. Udhayamathi, "Minimizing Communication Cost In Wireless Sensor Networks To Avoid Packet Retransmission", *International Innovative Research Journal of Engineering and Technology*, Vol. 2, Special Issue, pp. 38-42.

