Real Time Analysis of Top Trending Event on Twitter: Lexicon Based Approach

A.Haripriya
Computer Science and Engineering Department
M S Ramaiah University of Applied Sciences
Bengaluru, India
avadhanampriya@gmail.com

Santoshi Kumari
Computer Science and Engineering Department
M S Ramaiah University of Applied Sciences
Bengaluru, India
Santoshik29@gmail.com

***Abstract*— Social media has become an integral part of everyone's daily routine in today's digital era. It has evolved in an immense way with its growing number of channels, never imagined of a generation ago. Nowadays, it has become rare to see anyone without their mobiles in their hand, browsing and logging in through their social media network such as Facebook, Twitter, WhatsApp and so on. Enormous amounts of real-time data being generated every second across the world mostly in unstructured text messages. It brings in a huge challenge and opportunity to analyze in order to discover answers to the various questions and solve many real time problems. This paper mainly focusses on two aspects. One is to identify the real-time top trending event on twitter for a particular location and next is to consider the first trending event for doing sentiment analysis. Sentiment analysis is performed using lexicon based approach classified and visualized in ten different emotion parameters.**

***Index Terms*— Social Media, Text Mining, Sentiment Analysis, Lexicon, Real Time.**

I. INTRODUCTION

In this information age, social media plays a pivotal role across the world changing the way of communication and building in relationships tremendously. The rise in social media sites and its usage has also opened up the possibilities of discovering, sharing and learning new information faster than ever with real time data. More than 1.4 billion online users are spending 22 percent of their time on social platforms. Most of this data are being captured in unstructured format like Facebook likes, comments, tweets from twitter in text format, videos, voice recording, blogs, and emails. Approximately 85 percent of the data is in unstructured textual form [1] making it more challenging to analyze it to find, understand the sentiments and emotions of public.

As Twitter is one of the most used social media site with huge volumes of data since its start in 2006 with its strength in real-time data, it is considered for sentiment analysis. In Twitter, every single second, on an average, around 6,000 tweets are tweeted, corresponding to over 350,000 tweets sent per minute, 500 million tweets per day and almost around 200 billion tweets per year [2]. Capturing and analyzing the sentiments of these rich tweets with real time data provides a huge opportunity for various businesses and organizations by providing a platform to interact with customers and can yield high benefits in all the fields of research.

Real-time analytics has become essential in this digital evolution and it is vital for taking actions or decisions in almost all the sectors. Faster it is available, faster a decision can be made and in some cases, it has the potential to save the lives and prevent the loss of lives. For example, Twitter data is extensively used in Japan on research of earthquakes, Paris attacks, played a major role in US presidential elections etc. and many more.

Sentiment analysis, also known as opinion mining aims to determine the attitude of a writer or a topic of interest or an emotional reaction to an event or interaction or to a document. The main objective is to classify the polarity of a given text to positive, negative or neutral and to understand the emotional state. They deal with natural language text stored in semi structured or unstructured format [3]. These are being applied in various domains such as Finance, Healthcare, various business organizations for brand building and also to increase customer reach and sales, Fraud detection etc.

The main aim of this paper is to find the sentiment analysis on top trending topics using real time twitter data for any given location. Sentiment analysis is performed using lexicon based approach and the details are captured and the results are discussed.

This paper is organized in following sections. Related work is discussed in Section II. The proposed system details are provided in Section III followed by the results and discussion in Section IV with detailed information. The conclusion of this paper is provided in Section V with future work direction.

II. RELATED WORK

There have been many research works that are being carried out on sentiment analysis over a decade and many studies are being conducted till now. Initial studies have proven that machine learning techniques are efficient than human generated baselines [4]. According to many researchers, a strong positive correlation between the presences of adjectives exists in a sentence and in the presence of opinions [5]. Bag of Words technique is used for sentiment analysis in which relationships between the words are not considered and the

document itself is represented as a collection of words [6]. With initial document level classification task, sentence level [7], [8], phrase level [9] classifications are being explored. Using Lexical based approach, database WordNet, emotional content of a word with different dimensions are determined [10]. Automated dictionaries are generally larger such as SentiWordnet [11] that includes 38,182 non neutral words. Maryland dictionary [12] includes more words having 76,775 phrases tagged for polarity and words.

Liu et al. [16] has compared products based on the sentiment. The features are extracted and the words associated with them are compared in their work. Emotional value of the sentence apart from polarity sentiment is identified in the work done by saif and peter. They have combined three sentiment dictionaries: 'afinn', developed by Finn Arup Nielsen and second one, 'bing' is developed by Minqing Hu and Bing Liu and third one, 'nrc' is developed by Mohammad, Saif M. and Turney, Peter.D[14] ,[15], Hu and Liu's approach extracts specific sentences and identifies as subjective sentences for opinion analysis and then performs the analysis [16]. Maning et al describes about the 'coreNLP' tool created at Stanford [17] for natural language processing.

Machine learning approaches are also most prominently in place with Support Vector Machine, Naïve Bayes classifier, Maximum entropy and many more techniques. Many research works are being carried out in microblogging websites such as twitter with initial works of Alec et al. involving machine learning approaches [18]. As twitter provides meta data such as user id, latitude and longitude, Barbosa and Feng [19] has suggested to use them as traditional approaches are not well suited for microblogging sites. Most of the research work done till now are based on a specific topic or an event. This paper attempts to provide a generic approach to find the sentiment analysis of real-time events happening on twitter based on a given location quickly for taking prompt actions.

III. Proposed System

In this paper, Sentiment analysis of top trending event is carried out in following two main steps. One is to identify the top trending events on twitter for a particular location and next is to carryout sentiment analysis of top first trending topic using lexicon based approach. This is performed as shown in the Figure 1. Sentiment Analysis Model for top trending event.

Steps involved in this analysis are given here:

a. Establish connection: Here, using twitter API, OAuth handshake is done for each and every request for getting data from twitter to R. An app is created using the twitter account. Once an app is created, user's own twitter authentication credentials such as api_key,api_secret etc. are generated. In the tool necessary packages have to be installed and an authorization object is created using the twitter credentials for establishing the connection.

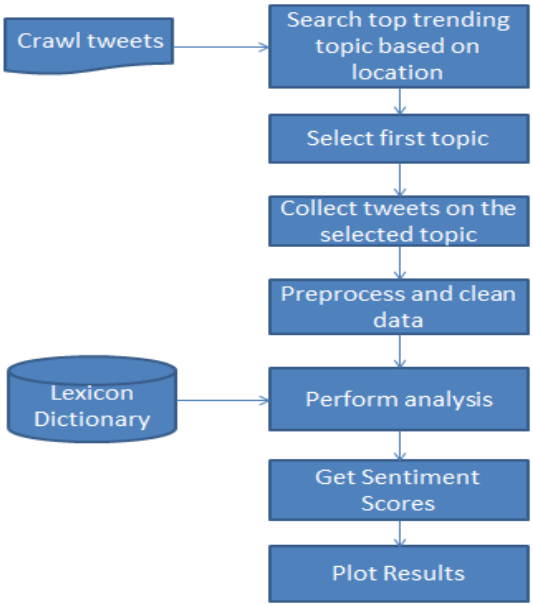

Figure 1. Sentiment Analysis Model for top trending event

b. Data Collection: Once the twitter connection is established, tool can access the real time tweets from twitter. Here, based on the given location, top trending topics are retrieved from twitter. On getting the trending topic, first topic or event is considered for analysis. Tweets related to this trending topic are extracted and stored in a file for doing analysis.

c. Preprocess and clean data: Data stored is read to perform cleaning and preprocessing. For preprocessing data, various methods such as removing special characters, removing stopwords, numbers, punctuation are proposed along with stemming methods.

d. Perform Analysis: Using Term Frequency weighting, the frequency of words is proposed in the text mining. For performing sentiment analysis, 'syuzhet' package is used. This package extracts the sentiment and also sentiment derived plot arcs from a given text. Here in this paper, 'get_nrc_sentiment' is used from NRC sentiment dictionary for calculating the presence of various emotions. The function, 'get_nrc_sentiment', is designed based on the work of saif and Peter. It has a list of words and also has an association of the words with different emotions: anticipation, anger, fear, trust, surprise, sadness, disgust and joy.

e. Plot the Results: Using the document term matrix, a word cloud is generated from the maximum count words obtained for better visual representation of text data. The sentiment scores are collected and bar graph is plotted for understanding various emotions of the real time top trending topic based on a given location.

The proposed model with main steps used for performing sentiment analysis is provided in this section.

IV. Results and Discussion

The results of the proposed model are provided here. Using twitter API and relevant R packages, connections are established between twitter and R. Once the connections are established, real time top trends are extracted from twitter based on India location in this paper.

Table 1.Toptrends in twitter in India on 3rd May

Name	Woeid
WorldPressFreedomDay	2295420
TubelightTeaserTomorrow	2295420
QuintMemories	2295420
MereDilMein	2295420
LoanKaHaq	2295420

Table 1. Top trends in twitter in India on 3rd May 2017, shows the current real time top 5 trending topics taken on third May 2017 at 2: 30 pm. Where On Earth Identifier: woeid; 2295420 represents Indian location. Similarly, based on the geolocation or given latitude or longitude, real time top trending topic or an event or personal etc. can be considered for analysis using this generic approach.

Here, from the top trending topics, first topic, 'World Press Freedom Day' is considered for performing sentiment analysis in this paper. For analysis, 500 tweets are extracted using twitter authorization credentials. The United Nations General Assembly had declared 3rd May as World Press Freedom Day. The United Nations Educational, Scientific and Cultural Organization (UNESCO) marks this day each year. They bring together media professionals, press organizations and the United Nation agencies in order to assess the state of worldwide freedom of press by conducting conferences centered on a theme related to freedom of press and the role of media.

Table 2.Sample Tweets on World Press Freedom Day

1	RT @PalObserver: #WorldPressFreedomDay No such thing as Freedom of the Press by Israel. Exposing their Crimes to the world is considered a…
2	RT @OfficeOfRG: No better words than his to guide us to courage & freedom. Best wishes on #WorldPressFreedomDay https://t.co/tpUPcdWmQg

Table 2.Sample tweets on World Press Freedom Day represents the text messages on world Press Freedom day. Similarly, based on the current top trending topic, recent tweets are searched and stored in a csv format for further analysis. The data is then read and a corpus of words is created with the collected tweets. It is preprocessed and cleaned using various methods such as removing punctuation, numbers and stop words. Stemming of words is also performed. Document term matrix is then generated with sorting. It is a table containing the frequency of the words using term frequency weighing method. 887 unique terms are obtained with 99% sparsity. This data is used to create a word cloud for visual representation.

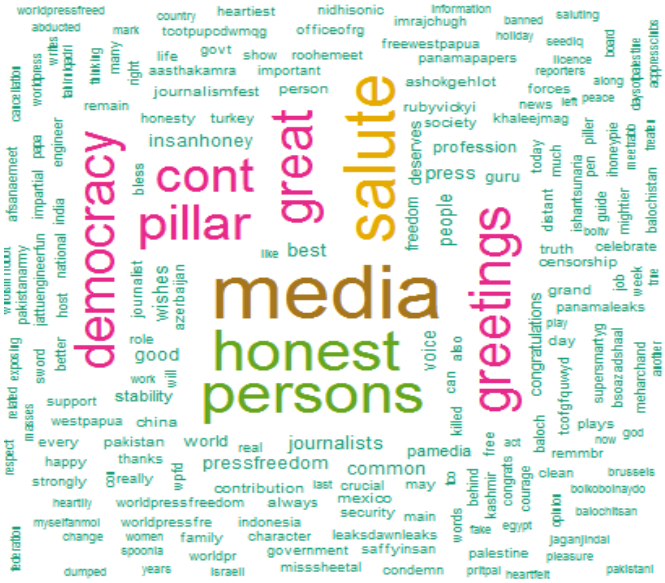

Figure 2.Word Cloud on World Press Freedom Day

Figure 2. Word Cloud on World Press Freedom Day represents the word cloud formed with the data obtained in twitter.

Using 'get_nrc_sentiment' function, sentiment analysis on the twitter text data is performed. Sentiment scores are obtained and a bar graph is plotted for understanding the emotions of the top trending location based real time data with ten emotions: anger, anticipation, disgust, fear, joy, negative, positive, sadness, surprise and trust.

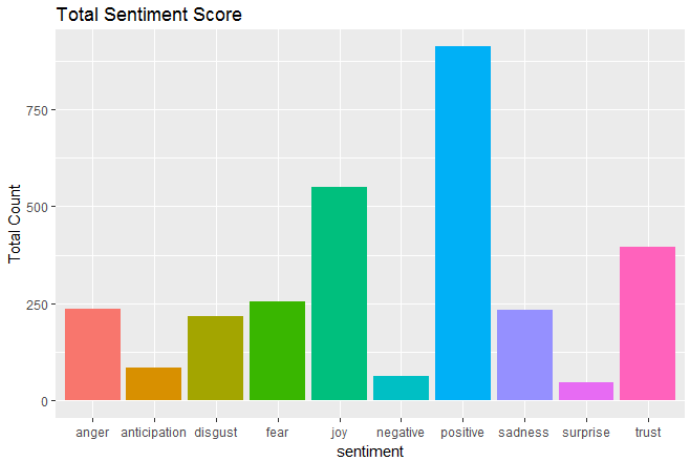

Figure 3.Sentiment analysis on World Press Freedom Day

Figure 3. Sentiment analysis on World Press Freedom Day displays the obtained results of the analysis. The sentiment is

more positive with trust and joy emotions. There are a very few emotions with anger, fear, anticipation and negative emotions related to world press freedom day topic. Similar to this, sentiment of netizens can be analyzed with the real time location based data on any datasets.

The top trending event changes with time and for a given location. As a future work, this model can be scheduled for any given time to get the up to date real-time analysis of top trending event for a given location and sentiment analysis can be derived. However, the accuracy of the model depends on whether the words entered in the tweets are there in lexicon based dictionary. This work can be extended by training algorithm and validating the model using confusion matrix. Further it can be improved by combining lexicon and machine learning approach.

V. CONCLUSION

The data explosion with social media sites and various blogs has thrown challenges in processing, identifying and extracting knowledge from unstructured data. Currently, it has become necessary to quickly analyze and understand the sentiments of public on top trending events on twitter. This helps to predict and take further actions. A generic sentiment analysis model is proposed for analyzing twitter data using lexicon based approach. In which analysis is carried out by identifying and collecting tweets on the top trending topic or an event at real time for any location. Most frequent words are identified and the sentiment scores are collected and plotted in a graph for better depiction of emotions.

Other lexicon based approaches can further be explored as part of future work. Lexicon with machine learning techniques such as Naïve Bayes classifier, Support Vector Machine along with feature extraction and selection methods can be used to improve the sentiment analysis and get better result accuracy.

REFERENCES

[1] Gerber D, Hellmann S, Bühmann L, Soru T, Usbeck R, Ngomo AC. “Real-time RDF extraction from unstructured data streams.” , In International Semantic Web Conference, Springer Berlin Heidelberg, pp. 135-150, October 2013.

[2] Hussain J, Islam MA., “Evaluation of graph centrality measures for tweet classification”. In Computing, Electronic and Electrical Engineering (ICE Cube), International Conference on IEEE, pp. 126-131, April 2016.

[3] Feldman R, Sanger J., “The text mining handbook: advanced approaches in analyzing unstructured data”. Cambridge university press; 2007.

[4] Pang, B., Lee, L., & Vaithyanathan, S , Thumbs up? Sentiment Classification using Machine Learning Techniques, 2002.

[5] Janyce Wiebe, Rebecca F. Bruce, and Thomas o’hara. "Development and use of a gold standard data set for subjectivity classifications". In Proc. 37th Annual Meeting of the Assoc. for Computational Linguistics (ACL-99), 1999.

[6] P. D. Turney, “Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews,” in Proceedings of the 40th annual meeting on association for computational linguistics, pp. 417–424, Association for Computational Linguistics, 2002.

[7] M Hu and B Liu., Mining and summarizing customer reviews. KDD, 2004.

[8] S M Kim and E Hovy., Determining the sentiment of opinions. Coling, 2004.

[9] Agarwal A, Biadsy F, Mckeown KR. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. InProceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, pp. 24-32, Association for Computational Linguistics, March 2009.

[10] J. Kamps, M. Marx, R. J. Mokken, and M. De Rijke, “Using wordnet to measure semantic orientations of adjectives,” 2004.

[11] Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani., SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC’10), pages 2200–2204, 2010.

[12] Mohammad, Saif, Bonnie Dorr, and Cody Dunn, Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2009), pages 599–608, 2009.

[13] Feinerer I. Introduction to the tm Package Text Mining in R. 2013-12-01], pdf. March, 2017.

[14] Jockers M. Package ‘syuzhet’, 2016.

[15] Saif Mohammad and Peter Turney. "Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon." In Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, LA, California,June 2010.

[16] Minging Hu and Bing Liu, “Mining Opinion Features in Customer Reviews”, Proceedings of 19th National Conference on Artificial Intelligence, pp. 755-760, 2004.

[17] Alec Go, Lei Huang and Richa Bhayani, “Twitter Sentiment Analysis”, Final Project Report, Stanford University, pp. 1- 16, 2009.

[18] Efthymios Kouloumpis, Theresa Wilson and Johanna Moore, “Twitter Sentiment Analysis: The Good the Bad and the OMG!”, Proceedings of 5th International Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social Media, pp. 538-541, 2011.

[19] Luciano Barbosa and Julan Freng, “Robust Sentiment Detection on Twitter from Biased and Noisy Data”, Proceedings of the 23rd International Conference on Computational Linguistics, pp. 36-44, 2010.