IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 4, April 2018

Real Time Twitter API Based Sentiment Analysis using Data Mining Algorithm

R. Jenifer[1], C. Bhuvaneshwaran[2]

ME Scholar, Mother Teresa College of Engineering and Technology, Mettusalai, Pudukkottai, India[1]

Assistant Professor, Mother Teresa College of Engineering and Technology, Mettusalai, Pudukkottai, India[2]

Abstract: Social media sites are places where citizens voice their opinions without fear. There is growing sense of urgency to understand public opinions because of the viral nature of social media. Making sense of these mass conversations for interacting meaningfully is in demand. Sentiment analysis is the study where sentiments are computed for a conclusion. It is extremely useful in monitoring and can help gain an overview of wider public opinions behind certain topics. The applications of sentiment analysis are broad and powerful. Shifts in sentiment on social media have been shown to correlate with shifts in the stock market or quickly understand consumer attitudes. From a managerial point of view, sentiment analysis can provide means to optimize marketing strategies. In marketing tactics, sentiment analysis can help fit marketing campaigns for target audiences. Success of a campaign also lies in positive discussions amongst customers, where sentiment analysis plays a major role. Sentiment analysis can also be the base for market research and quality improvement. Moreover, the volume of digital information on the Internet has been responsible in increasing access times on items of interest for users. This voluminous information has to be filtered, prioritized and delivered to users to satisfy their search requirements for recommendations. This paper underlines the need for sentimental analysis and recommender systems based on sentimental analysis for users. Further, it proposes a Sentiment Analysis Method based on Naïve Bayes data mining technique on unigram and bigram tweets. The proposed work aims to fulfill sentimental analysis with speed for recommender systems useful to end users.

Keywords: API, Data mining, Social media, Sentiment analysis.

I. INTRODUCTION

The media are exploring the most important way for the broad masses of viewers, and interacting with them in the event that they create events in virtual communities and networks that share and exchange information and ideas. Social media technology takes a variety of different forms, including magazines, web forums, blogs, social blogs, microblogs, wikis, social networks, podcasts, photos or pictures, videos, ratings and social bookmarks. Microblogging sites have evolved into a variety of sources of information. This is due to the ten-thousandth of the blog, people post their opinions on various topics related to the nature of real-time information, discuss current, complain, express positive feelings about the products they use in the problems of everyday life. In fact, companies in the production of these products have begun to investigate the general meaning of these micro-blogs to make your products feel. Often, these companies are studying the user's response and the user in response to micro blogging. The status and importance of social media in society is increasing. Public and private opinions on a variety of topics are constantly being expressed and disseminated through many social media, where twitter is the most fashionable. Social media has become one of the biggest forums for expressing opinions. Emotional analysis is intended to determine the attitude of a speaker or writer in terms of the overall background polarity of some subjects or documents. This attitude may be his or her judgment or assessment, the emotional state (the author's emotional state at the time of writing), or the emotional exchange of intentions (the author wishes to have an emotional effect on the reader). The basic task of emotional analysis is to determine the polarity of the text in a given document, sentence, or trait / level - if it is assumed that the entity in the document, phrase, or feature / aspect is positive, negative, or neutral. The emergence of advanced emotional categories, "transcendence", for example, in the emotional state of "anger", "sad" and "happy".

II. LITERATURE REVIEW

Sunny Kumar et el. R input language is powerful and suitable for implementing data extraction and data analysis tools. However, when the data size is in other words, the size of the data exceeds the size of the physical memory environment R, and R gives a poor result, sometimes ending the R-Talk.

Gayathiri.R at el. Based on the assessment of the views of people who only consider the high accuracy of the positive and negative evaluation rules, emotional analysis. Finally, the accuracy of the system and the accuracy of the recall measures are used to calculate the results.

DOI 10.17148/IJARCCE.2018.7465

Tian-Shyug Lee et al. The results show that in the TTE increase the number of visits, but the market optimism, the annual decline in the tourists. PEO should be able to quickly make quality decisions, accurately meet the needs of visitors, show the social media platform, constantly monitor how they are in the social media investment in who broke into ROI.

Amol Patwardhan et al. Detection group and the environment in the crowd spontaneous emotion. Edge detection uses a grid overlay with lines to extract features. The movement of the feature from the aspect of the movement of the reference point is used to filter through the color channel image sequence. In addition, video data collection was carried out by viewing spontaneous emotions in the participants of the sporting event. The method is independent of vision and obstruction, the results are not subject to multiple people expressing various emotions chaotically exist.

Anne Veenendaal et al. Check the use of color and depth data (RGB-D) motion analysis and frame recognition of human behavior. Specifically, the identification of attacks such as throwing, kicking, punching, using 3D depth sensors threatens aggression. The RGB-D data obtained from the operation and the infrared detection of the device is recorded. Was asked to take 23 students to take positive measures. The SVM classification uses the training from the series of frames and the functional life of the combat scene based on the simulation of the combat scene. The results show that the performance of single individual stocks is better, but the system performance of the poor performance of the detection group activities.

Mustofa Kamal et al. System to analyze the Indonesian people's views on the ASEAN Free Trade Area and extract useful information about the sensory analysis of his opinion. Our system consists of five calculation steps: (1) data acquisition, (2) text mining, (3) sentiment analysis engine, (4) time projection engine and (5) processing unit context. We use a new approach, based on the analysis of time and space to explore the views of public opinion on the views of the ASEAN Free Trade Area. This method automatically processes and extracts textual data to obtain information on the mood phrases. The results show that the simple method of effectiveness to obtain information on the views of people and some stakeholders and who are interested in the ASEAN Free Trade Area students have already felt that this solution has been evaluated.

Shaohua Wan et al. The proposed framework is expected to maximize the performance criteria for the repeated establishment of a specific gold standard for the given gold standard, and then refine the gold standard based on performance indicators. At the same time, to relieve the potential overlap of the geometrical features and the appearance of the different facial expressions, repeat the distance between our update and the performance score of the new estimated scorer. Head movement is mainly limited to the plane of the image. The use of high-back chairs accompanied by spontaneous smiles and smiling conditions may only have limited movement of the aircraft, and a few occur from the analysis.

III. SYSTEM ANALYSIS

A. Existing System

The existing system works only on the dataset which is constrained to a particular topic. The existing systems also do not determine the measure of impact the results determined can have on the particular field taken into consideration and it does not allow retrieval of data based on the query entered by the user i.e. it has constrained scope. In simple words, it works on static data rather than dynamic data. Unsupervised algorithms like Vector Quantization are used for data compression, pattern recognition, facial and speech recognition, etc and therefore cannot be used in determining sentiment in twitter data. Apriori algorithm fails to handle large datasets and as a result can generate faulty results.

B. Proposed System

The proposed system architecture consists of five main steps. The first step for sentiment analysis of twitter data is to acquire the data that has to undergo sentiment analysis. Raw tweets were obtained at the end of this stage and stored in a database which was the input for preprocessing step. Preprocessing consists of several steps which are the removal of URLs from the tweet, hashtag removal so that the tweet becomes more cleaned, removal of slangs, emoticon conversion to text, stop word removal. Filtered tweet is obtained at the end of this step which becomes the input of the Sentiment Influencers Identification step. Filtered text is passed to a tokenizer which separates each word in the tweet. This set of words (W) is passed to a POS tagger which tags the words of the tweet as nouns (N), verbs (V), adverbs (Adv) and adjectives (Adj). The output of this step is tagged text (T.T) which is a set containing noun, verbs, adjectives and adverbs. This set of POS is named as sentiment influencers. The output tagged text of the previous phase becomes the input of next step. The score of the T.T is calculated using SentiWordNet dictionary which assigns a score to each word in the tweet. The score obtained from SentiWordNet dictionary is named as standard score (S.S). Tagged score (T.S) is calculated by ranking the sentiment influencers which will be discussed in detail in implementation section. After the calculation of T.S, negation is checked in the tweet if it’s present then the final score of the tweet is inversed

DOI 10.17148/IJARCCE.2018.7465

that is multiplied with -1. The final score is calculated by summing the tagged score obtained by tagging the text and dividing it with the total number of words (W) in a tweet. The final score is passed to sentiment classifier which classifies the tweet into positive, negative or neutral. Sentiment classifier will classify the tweet as follows. If the final score of the tweet is greater than zero, the tweet is declared as positive. If the final score of the tweet is less than zero, the tweet is declared as negative. If the final score of the tweet is equal to zero, the tweet is declared as neutral.

IV. IMPLEMENTATION

A. Data Streaming

Extracting real time tweets using Twitter Streaming API For classification and training the classifier we need Twitter data. For this purpose we make use of API's twitter provides. Twitter provides two API's; Stream API1 and REST API2. The difference between Streaming API and REST APIs are: Streaming API supports long-lived connection and provides data in almost real -time. The REST APIs support short-lived connections and are rate-limited (one can download a certain amount of data [*150 tweets per hour] but not more per day).

B. Preprocessing

In this phase, the tweets are available as text data and each line contains a tweet. Initially we clean up or remove retweets as that will induce a bias in the classification process. We need to remove the punctuations and other symbols that doesn’t make any sense as it may result in inefficiencies and may affect the accuracy of the overall process

C. Sentiment Polarity Analysis

MapReduce is a new parallel programming model, hence the classical Naive Bayes based sentiment analysis algorithm is adjusted to fit into Map Reduce model. we choose to employ a Naive Bayes classifier and empower it with an English lexical dictionary SentiWordNet

D. Visualization

Tweets are presented using several different visualization techniques. Each technique is designed to highlight different aspects of the tweets and their sentiment.

E. Heat Map

The Heat Map visualizes the number of tweets within different sentiment regions. It highlights "hot" red regions with many tweets, and "cold" blue regions with only a few tweets.

F. Tag Cloud

The tag cloud visualizes the most frequently occurring terms in four emotional regions: upset in the upper-left, happy in the upper-right, relaxed in the lower-right, and unhappy in the lower-left. A term's size shows how often it occurs over all the tweets in the given emotional region. Larger terms occur more frequently.

G. Timeline

The timeline visualizes when tweets were posted. Pleasant tweets are shown in green above the horizontal axis, and unpleasant tweets in blue below the axis.

H. Map

The map shows where tweets were posted. Twitter uses an "opt-in" system for reporting location: users must explicitly choose to allow their location to be posted before their tweets are geotagged.

I. Affinity

The affinity graph visualizes frequent tweets, people, hashtags, and URLs, together with relationships or affinities between these elements.

J. Evaluation Metrics

It will evaluate our experiment results by using following Information Retrieval matrices.

Precision = TP/(TP + FP)
Recall = TP/(TP + FN)
F-measure = 2*Precision*recall/(Precision + recall)
Accuracy = TP + TN /(TP + TN + FP + FN)

DOI 10.17148/IJARCCE.2018.7465

TABLE I
PERFORMANCE MEASURES OF TWEETER SENTIMENTAL ANALYSIS

Keyword	Positive Rate	Negative Rate	Accuracy
“Modi”	30	9	97.88
“Donald Trump”	32	12	93.45
“Cauveri”	23	7	90.11
“duststrom”	19	19	87.41
“election”	37	11	95.21
“tamilnadu”	12	6	82.91

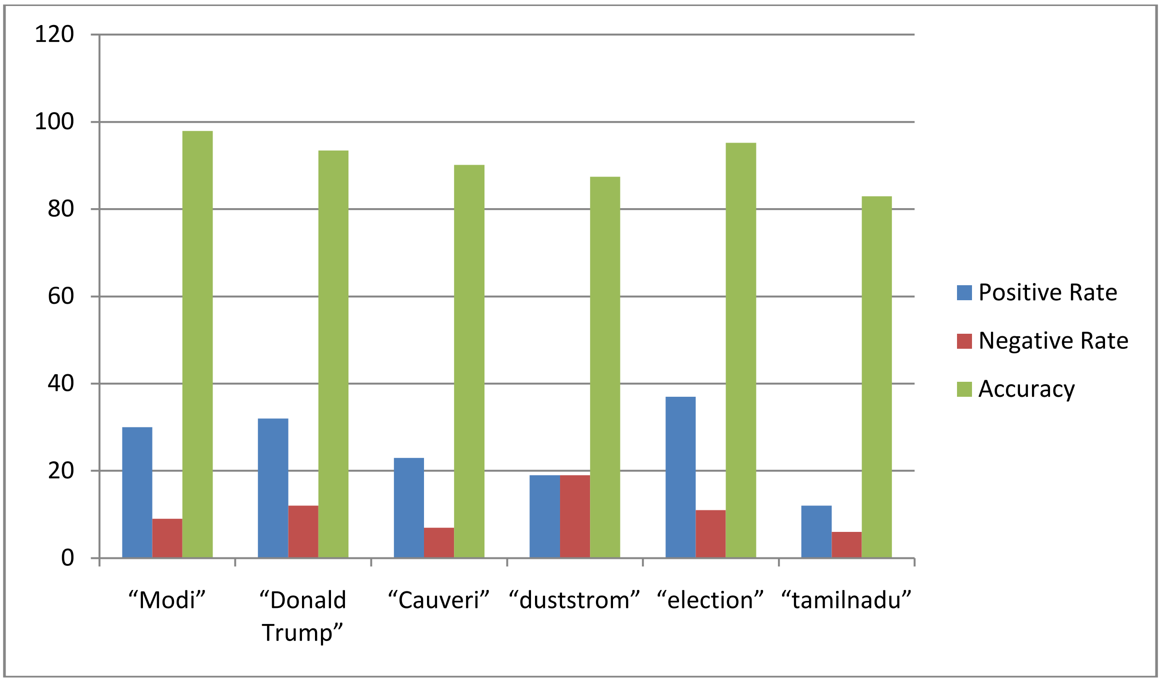
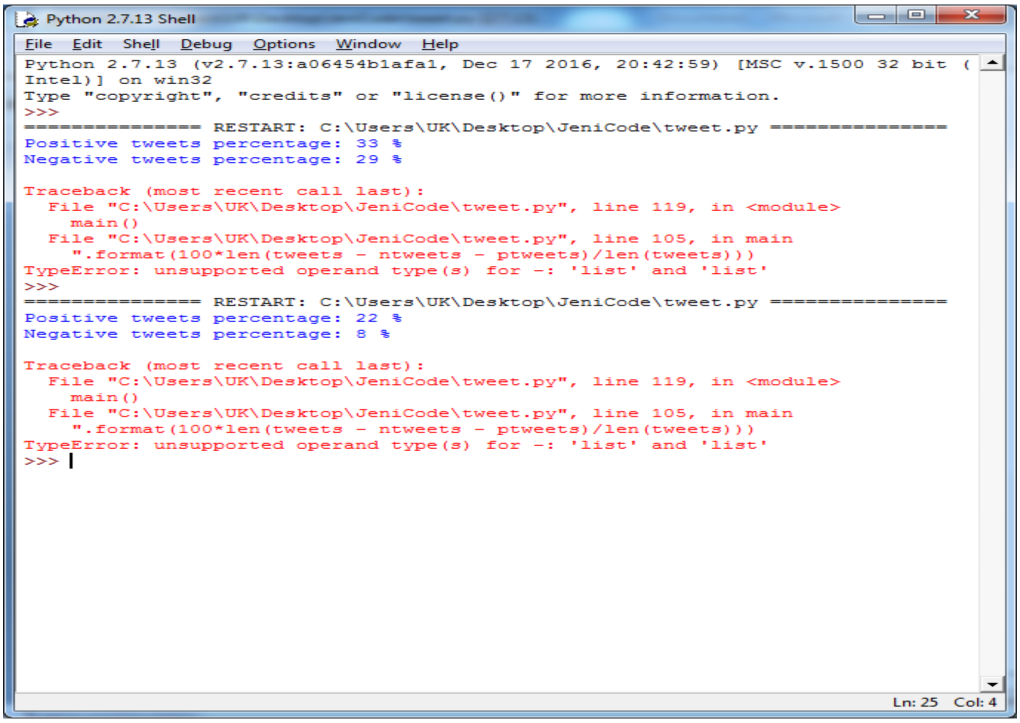


Fig. 1 Performance Chart



```
Python 2.7.13 (v2.7.13:a06454b1afa1, Dec 17 2016, 20:42:59) [MSC v.1500 32 bit (
Intel)] on win32
Type "copyright", "credits" or "license()" for more information.
>>> 
=============== RESTART: C:\Users\UK\Desktop\JeniCode\tweet.py ===============
Positive tweets percentage: 33 %
Negative tweets percentage: 29 %

Traceback (most recent call last):
  File "C:\Users\UK\Desktop\JeniCode\tweet.py", line 119, in <module>
    main()
  File "C:\Users\UK\Desktop\JeniCode\tweet.py", line 105, in main
    ".format(100*len(tweets - ntweets - ptweets)/len(tweets)))
TypeError: unsupported operand type(s) for -: 'list' and 'list'
>>> 
=============== RESTART: C:\Users\UK\Desktop\JeniCode\tweet.py ===============
Positive tweets percentage: 22 %
Negative tweets percentage: 8 %

Traceback (most recent call last):
  File "C:\Users\UK\Desktop\JeniCode\tweet.py", line 119, in <module>
    main()
  File "C:\Users\UK\Desktop\JeniCode\tweet.py", line 105, in main
    ".format(100*len(tweets - ntweets - ptweets)/len(tweets)))
TypeError: unsupported operand type(s) for -: 'list' and 'list'
>>> 
```

Fig. 2 Screen Shot of Keyword Analysis in Tweeter Data.

DOI 10.17148/IJARCCE.2018.7465

V. CONCLUSION & FUTURE WORK

Twitter is an excellent initial point for social media analysis. People directly share their opinions through Twitter to the general public. One of the very common analyses which can perform on a large number of tweets is sentiment analysis. In the proposed work, tweets are collected using Twitter streaming API from twitter. The collected tweets are preprocessed using Natural Language Toolkit techniques. The features of the tweets are selected based on Chi-Square test and Naïve Bayes classifier is used to classify the tweets as positive and negative. This proposed work is implemented using Python. proposed system would be easy for user to obtain the Scalable Sentiment Classification also used to support them in decision making process in for their daily life activities. It is proposed to stream real time live tweets from twitter using Twitter API, and the large volume of data makes the application suitable for Big Data Analytics. A method to predict or deduct the keyword of a tweet based on the tweet's information and the user's information should be found in the future.

REFERENCES

[1] Nikita Jain , Vishal Srivastava, "Data Mining Techniques: A Survey Paper", IJRET, eISSN: 2319-1163 | pISSN: 2321-7308.
[2] Yihao Li, "Data Mining: Concepts, Background and Methods of Integratign Uncertainty in Data Mining"
[3] Vishal A. Kharde, S.S. Sonawane," Sentiment Analysis of Twitter Data: A Survey of Techniques", 2016, International Journal of Computer Applications, Volume 139 – No.11
[4] Deepali Arora, Kin Fun Li and Stephen W. Neville," Consumers' sentiment analysis of popular phone brands and operating system preference using Twitter data: A feasibility study", 2015, IEEE, 1550-445X
[5]Haseena Rahmath P , Tanvir Ahmad, "Sentiment Analysis Techniques - A Comparative Study", IJCEM International Journal of Computational Engineering & Management, Vol. 17 Issue 4, July 2014
[6] P. Grandin and J. M. Adán," Piegas: A System for Sentiment Analysis of Tweets in Portuguese", 2016, IEEE LATIN AMERICA TRANSACTIONS, VOL. 14, NO. 7
[7] Alexander Porshnev, Ilya Redkin, Alexey Shevchenko," Machine learning in prediction of stock market indicators based on historical data and data from Twitter sentiment analysis," 2013, IEEE, 879234-645-345
[8] LI Bing, Keith C.C. Chan, Carol OU," Public Sentiment Analysis in Twitter Data for Prediction of A Company's Stock Price Movements", 2014, IEEE, 978-1-4799-6563-2 [9] Ryan M. Eshleman and Hui Yang," A Spatio-temporal Sentiment Analysis of Twitter Data and 311 Civil Complaints", 2014, IEEE, 978-1-4799-6719-3
[10] Rincy Jose, Varghese S Chooralil," Prediction of Election Result by Enhanced Sentiment Analysis on Twitter Data using Word Sense Disambiguation", 2015, IEEE, 978-1-4673- 7349-4
[11] Nehal Mamgain, Ekta Mehta, Ankush Mittal, Gaurav Bhatt," Sentiment Analysis of Top Colleges in India Using Twitter Data", 2016, IEEE, 978-1-5090-0082-1
[12] Aldo Hernández, Victor Sanchez, Gabriel Sánchez, Héctor Pérez, Jesús Olivares, Karina Toscano, Mariko Nakano and Victor Martinez," Security Attack Prediction Based on User Sentiment Analysis of Twitter Data", 2016, IEEE, vol. 56, pp.45
[13] Anurag P. Jain, Mr. Vijay D. Katkar," Sentiments Analysis Of Twitter Data Using Data Mining", 2015 International Conference on Information Processing (ICIP), 978-1-4673- 7758-4
[14] Ming Hao, Christian Rohrdantz, Halldór Janetzko, Umeshwar Dayal, Daniel A. Keim, Lars-Erik Haug, Mei-Chun Hsu," Visual Sentiment Analysis on Twitter Data Streams", 2011, IEEE, 3927504-365-4-54
[15]. Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4):541–551, 1989.
[16]. Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In IEEE CVPR, 2014.
[17]. D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J.Z. Wang, J. Li. and J. Luo, "Aesthetics and emotions in images," Signal Processing Magazine, IEEE, vol. 28, no. 5, 2011, pp. 94-115.
[18]. S. Siersdarfer. E. Minack, F. Deng, and J. Hare. "Analyzing and predicting sentiment of images on the social web," in Proceedings of the 18th ACM international conference on Multimedia, 2010, pp. 715-718.
[19]. D. Borth. R. Ji, T. Chen, T. Breue!, and S.-F. Chang, "Large-scale visual sentiment ontology and detectars using adjective noun pairs," in Proceedings of the 21st ACM international conference on Multimedia, 2013, pp. 223-232.
[20]. L. P. Marency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in Proceedings ofthe 13th international conference on multimodal interfaces, 2011, pp. 169-176

DOI 10.17148/IJARCCE.2018.7465