International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-6, April 2020

Visualization of Real-time Twitter Data based on Sentiment Classification

Susmitha Peeka, Sabiya Tabassum, Dhanushya Pallem, Sowmya Aravapalli, M. Sobhana

Abstract*: Analyzing information from social media sites could bring great challenges and opportunities to solve many real time problems. It gives the public opinion about almost every product, personality or any service. The data from social networking sites is more accurate and useful to analyze the public sentiment about the trending topics. The activity of analyzing opinions, sentiments and also the subjectivity of data that is provided, is called sentiment analysis. Tweepy is an easy-to-use python library which is used to extract source data from twitter. From these tweets, features are extracted and then classified using Naïve Bayes algorithm to identify sentiment. This aims to provide an interactive automatic system which predicts the sentiment of the tweets posted in social media using python in real-time. These applications of sentiment analysis are broad and they tend to be very useful in today's lifestyle. It will evaluate people's sentiment about the trends, entertainment, political issues and products which helps to improve marketing strategies with the help of hashtags, keywords etc.

Keywords: Trending Topics, Real-time, Sentiment Analysis, Twitter, Python, Tweepy.

I. INTRODUCTION

The process of analyzing, identifying and categorizing sentiments of a piece of text or data based on its polarity is sentiment analysis. A sentiment analysis system for any text analysis is the combination of different algorithms in Natural Language Processing and any of the Machine Learning procedures to assign the polarity scores based on the brands, trending topics and other categories.

The activity of analyzing opinions, sentiments and also the subjectivity of data that is provided, is called sentiment analysis. We can obtain sentiment information mostly from internet. People can publicize their views or information through blogging and various networking sites like Twitter, Facebook etc., from users' point of view. Many of these sites are releasing their APIs, encouraging to collect and analyze the data which is useful for many developers and researchers. Hence, this sentiment analysis system appears to have a powerful fundamental support with the enormous amount of online data.

Revised Manuscript Received on April 30, 2020.
***** Correspondence Author
Susmitha Peeka*, CSE, V R Siddhartha Engineering College, Vijayawada, India. Email**:** peeka.susmitha@gmail.com
Sabiya Tabassum, CSE, V R Siddhartha Engineering College, Vijayawada, India. Email**:** sabiyatabassum99@gmail.com
Dhanushya Pallem, CSE, V R Siddhartha Engineering College, Vijayawada, India. Email**:** dhanushyapallemdhanu@gmail.com
Sowmya Aravapalli, CSE, V R Siddhartha Engineering College, Vijayawada, India. Email**:** sowmyamastan114@gmail.com
Dr. M. Sobhana, Sr. Assistant Professor, CSE, V R Siddhartha Engineering College, Vijayawada, India. Email: msobhana80@gmail.com

This helps large corporations with data analysts to observe brand and product reputation, and understand customer experiences, measure public opinion and to manage fine distinct statistical surveying. In sentiment analysis, there may be several defects in the data we consider, which blocks the whole process. One of the defects is, since anyone can post anything freely their opinions cannot be undertaken as it affects the analysis. For example, online spammers who post irrelevant data and some people who post fake opinions. The other is, the data that is used for analysis is not always available.

We analyze the reviews given by the people on any topic considered. In order to do this, we consider user reviews from Twitter. A lot of research has been done on user data which indicates how the information can be used to fore show various outcomes. Our present research deals with prediction and explores all possible outcomes.

II. RELATED WORK

Feddah Alhumaidi AlOtaibi et al.,[1] presented a methodology to perform sentiment analysis by using unsupervised machine learning algorithm. The data here considered is about KFC and McDonald and find which is famous among them. The data that is collected was fed with various models and the output that is obtained is tested using different metrics like cross validation and f-score.

Rasika et al.,[2] suggested different types and techniques to perform sentiment analysis. They defined that there is an increase in accuracy when different opinion mining techniques are used.

Prakruthi et al.,[3] proposed that real-time sentiment analysis was done by fetching tweets using Twitter API. Then the pre-processed tweets are compared with the Visualization of Real-time Twitter Data based on Sentiment Classification and then the tweets are labelled as positive, negative and neutral and the outputs are shown by a pie chart.

Vishal et al.,[4] presented a model which can do analysis on any data set as the data collected is unstructured and difficult to obtain sentiment. There is an increase in accuracy when machine learning techniques are used. Different methods based on semantics, which can be used to yield efficient results to perform opinion mining.

Prabhsimran et al.,[5] examined different opinions of people regarding the demonetization policy of government by performing sentiment analysis. The data that is used for this approach is twitter data. Tweets are collected based on location of the user by using hashtags or keywords.

Xing et al.,[6] resolved the issue of polarity score which is considered as one of the main problems in this sentiment analysis. Investigation has been achieved for both sentence-level categorization and review-level categorization.

Classification techniques such as Naïve Bayesian, Random Forest, SVM are used for categorization.

Geetika et al.,[7] developed a sentiment classification procedure for reviews given by customers for any products. The tweets are retrieved as a dataset and then performed preprocess on them. Feature Vector is used, that is the keywords are considered from the dataset. Naïve Bayes and SVM is used for classification and also WordNet is used for feature extraction process.

Rajasree et al.,[8] developed a new feature vector for reviewing the posts of electronic products from twitter. The Feature Vector had 8 distinct features. Naïve Bayes and SVM classifiers were used for classification

Niketan Jivane et al.,[9] suggested a model for the reviews on Hollywood and Bollywood films. It used SVM and Naïve Bayes classifiers and it was observed that SVM was more accurate but has less precision and recall compared to Naïve Bayes.

Liza et al.,[10] performed primary pre-processing, processing and validation. In processing, it used Naïve Bayes for classification of the textual data and then Cross Validation is performed as the final step in this process.

Chin-Sheng Yang et al.,[11] suggested different methods of feature extraction procedures like bigram and unigram. K-means clustering is used as the main technique for clustering and then is classified using Naïve Bayes algorithm.

Fei Liu et al.,[12] considered processing techniques for opposite opinion and non-opinion contents. Thus, K-means clustering-based approach was applied to find the results on reviews, comments, blogs.

III.METHODOLOGY

The methodology proposed consists of several steps and is represented as,

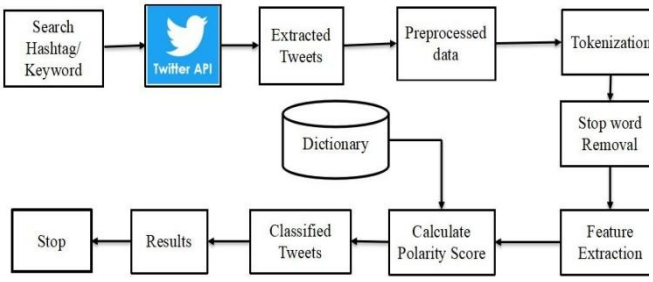

Fig. 1. Overview of Proposed Methodology

A. Data Preprocessing

The data that is extracted from twitter is not organized and structured. Therefore, data preprocessing mainly consists of two steps. They are Data cleaning and Stop word removal.

1. Data Cleaning

It comprises Removal of unnecessary data such as HTML Tags, white Spaces and Special characters obtained from the Tweets. Process of cleaning the data is as follows:

a. The First step in this is to remove the URL.
b. URL is not considered as an essential element in the tweets and just to reduce complexity, URLs are removed.
c. Twitter handlers such as '@abs' are also removed as they do not provide any importance in sentiment classification.
d. Punctuation removal takes place here and from that point forward, the evacuation of unique character happens.
e. Non-textual contents and contents that are not relevant for the analysis are identified and eliminated.
f. Single white space is replaced by Extra white spaces.
g. White spaces are removed from the beginning and also from the end.

2. Stop Words Removal

Stop word Removal is done by eliminating the unnecessary words from the Tweets that are taken from Twitter, so that, the resultant data set contains only the information required for the analysis. The process of stop word removal is shown below:

a. First of all, tokenization takes place. "Tokens" are usually individual words and "tokenization" is the process of considering text and breaking the text into its individual words.
b. After that, all the unnecessary words are removed such as 'a', 'an', 'the', and so on. These unnecessary words are nothing but stop words which have no meaning.

B. Feature Extraction

One of the crucial steps in opinion mining is feature extraction. After the preprocessing phase, only necessary and useful words are left in tweets which are used for analysis. The motive of feature extraction is to extract opinion sentences which contain one or more features, aspects, and opinions. Machine learning techniques are required in representing the major features, which are considered as feature vectors.

Some of them are,

- *Parts of Speech Tags*
  Grammatical forms like modifiers, qualifiers and a few gatherings of verbs and nouns are acceptable pointers of subjectivity and sentiment.
- *Opinion Words and Phrases*
  Beside some words, idioms and different phrases can be used as features which convey sentiments.
- *Position of Terms*
  The position of any term will make a difference in determining the complete sentiment of the text.
- *Negation*
  Negation is significant, but also a difficult feature to interpret. The polarity of the opinion normally gets affected by the negation

C. Sentiment Identification

After feature extraction phase, we identify the positive and negative orientation of words. Features that are obtained in feature extraction phase are searched into positive and negative word record of the dictionary. If any of the word is present in the positive word record, then the positive Sentiment will be assigned to the corresponding feature.

If it is present in the negative word record, then that sentiment is assigned to the corresponding feature.
If the corresponding word is not present in any of the records, then the sentiment is viewed as neutral. So, the overall polarity score for the tweet is determined by performing subtraction operation with variables negative score and positive score.

D. Sentiment Classification

We analyze the sentiment into various sentiments such as neutral, positive and negative which are classified further. If the final calculated polarity score of the tweet considered is between 0.6 and 1, then the tweet is labelled as strongly positive. If the polarity score of the tweet considered is between 0.3 and 0.6 then the tweet is labelled as positive. If the polarity score of the tweet considered is between 0 and 0.3 then the tweet is labelled as weakly positive. If the polarity score of the tweet considered is equal to 0 then the tweet is labelled as Neutral. If the polarity score of the tweet considered is between 0 and -0.3 but not equal to 0 then the tweet is labelled as weakly negative. If the polarity score of the tweet considered is between -0.3 and - 0.6 but not equal to -0.3 then the tweet is labelled as Negative. If the polarity score of the tweet considered is between -0.6 and -1 then the tweet is labelled as Strongly Negative.

E. Subjectivity Identification

A subjective sentence may not express any sentiment. In our system, we get the subjectivity score for the tweets using the Text Blob library function. Text Blob Library already has a dictionary that contains subjectivity score for the words. Modifiers increase the subjectivity of a word or sentence.

IV. RESULTS AND DISCUSSION

We consider the real time data/tweets that are streaming live in twitter. Using the Twitter API, we get those tweets and classify them accordingly.
. As we take the data in real time, the analysis of any specific keyword will be changing from time to time. That is, the latest tweets will be considered and then the polarity score will be calculated. We use different python libraries for analyzing the data and visualizing it.
We consider the data in the form of keywords or hashtags and then select the count of tweets that are to be analyzed according to the requirement. The result will be obtained in the form of pie chart where it is represented in various colors to visualize the sentiments.

Table- 1. Experimented results of the Data

	Nike	Lee Cooper	Reebok	Puma	Adidas	Skechers
Positive	5.30	0.70	15.20	11.10	10.20	10.70
Weakly Positive	6.10	1.20	27.30	13.60	13.40	16.80
Strongly Positive	1.00	0.80	5.70	20.40	8.60	1.70
Negative	0.90	0.20	2.10	1.30	5.00	1.80
Weakly Negative	5.20	0.30	6.50	11.20	6.80	17.50
Strongly Negative	0.20	0.10	0.90	0.90	5.00	2.10
Neutral	81.20	2.40	42.20	41.20	49.20	49.30

Based on the sentiment of 1000 tweets considered for every brand, polarity scores of different brands are divided as in the Table-1 and this score is then visualized with the help of MatPlot library.

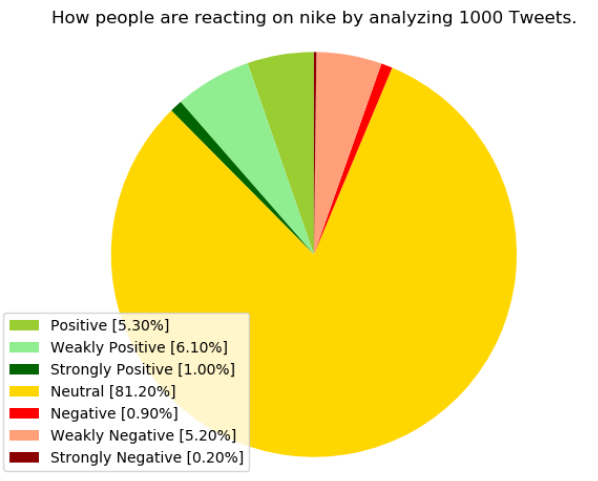

Fig. 1. Visualization of Output on NIKE

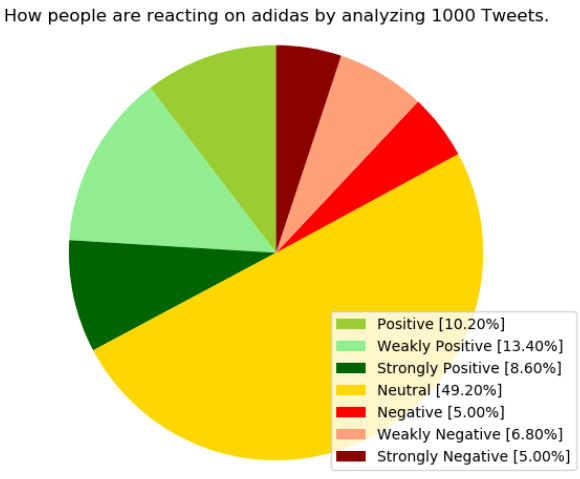

Fig. 2. Visualization of Output on Adidas

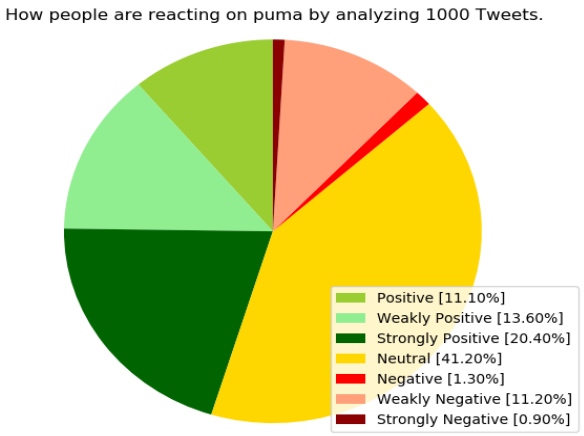

Fig. 3. Visualization of Output on PUMA

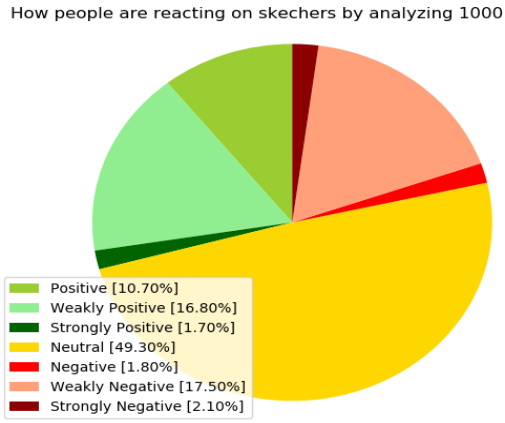

Fig. 4. Visualization of Output on Skechers

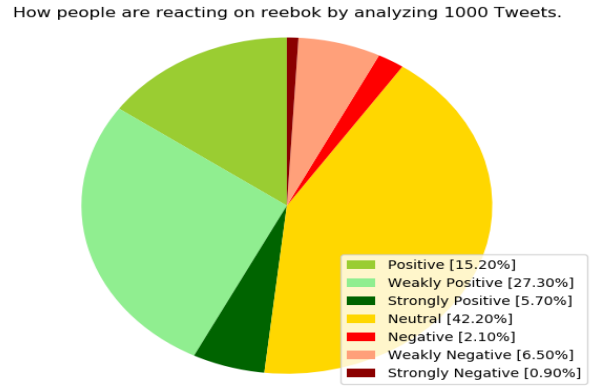

Fig. 5. Visualization of Output on Reebok

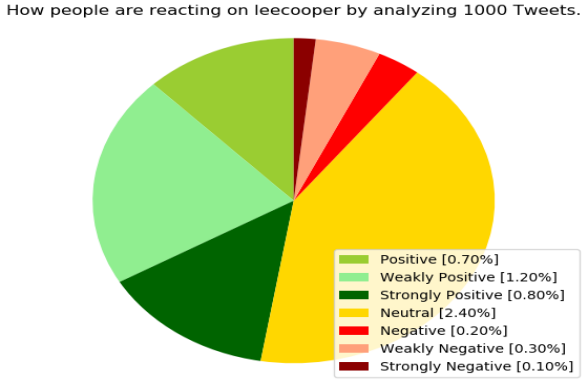

Fig. 6. Visualization of Output on Lee Cooper

V. CONCLUSION

Finally, we were able to analyze the opinion of each tweet by determining whether the it is positive or negative, which is beneficial in several fields like politics, finance, business and sociology. Here analysis is done in four steps such as Data Preprocessing, Feature Extraction, Classification. This classification algorithm is used to predict to which class the data belongs to. For example, we can predict the public opinion of the leaders participating in any election by considering the tweets using hashtags and keywords. Our method will help companies and organizations to know the real opinion about their products and can act accordingly.
Currently, we are striving to interpret sarcasm in our future work. Sarcasm is the use of irony to praise or convey, contempt transforms the polarity of favorable or bad utterance into its reverse. It will be useful to predict the actual feelings of the public and improves the result. We are also looking forward to implement it for stronger outcomes.

REFERENCES

1. El Rahman, Sahar A., Feddah Alhumaidi AlOtaibi, and Wejdan Abdullah AlShehri. "Sentiment analysis of Twitter data." In *2019 International Conference on Computer and Information Sciences (ICCIS)*, pp. 1-4. IEEE, 2019.
2. Rasika Wagh, and Payal Punde. "Survey on sentiment analysis using twitter dataset." In *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp. 208-211. IEEE, 2018.
3. Prakruthi, V, D. Sindhu, and S. Anupama Kumar. "Real Time Sentiment Analysis Of Twitter Posts." In *2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS)*, pp. 29-34. IEEE, 2018.
4. Kharde Vishal, and Prof Sonawane. "Sentiment analysis of twitter data: a survey of techniques." *arXiv preprint arXiv:1601.06971* (2016).
5. Prabhsimran Singh, Ravinder Singh Sawhney, and Karanjeet Singh Kahlon. "Sentiment analysis of demonetization of 500 & 1000 rupee banknotes by Indian government." *ICT Express* 4, no. 3 (2018): 124-129
6. Xing Fang, and Justin Zhan. "Sentiment analysis using product review data." *Journal of Big Data* 2, no. 1 (2015):5.
7. Geetika Gautam, and Divakar Yadav. "Sentiment analysis of twitter data using machine learning approaches and semantic analysis." In *2014 Seventh International Conference on Contemporary Computing (IC3)*, pp. 437-442. IEEE, 2014.
8. Neethu M. S., and R. Rajasree. "Sentiment analysis in twitter using machine learning techniques." In *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, pp. 1-5. IEEE, 2013.
9. Akshay, Niketan Jivane, Mahavir Bhandari, and M. Venkatesan. "Twitter sentiment analysis of movie reviews using machine learning techniques." *international Journal of Engineering and Technology* 7, no. 6 (2016): 1-7.
10. Wikarsa Liza, and Sherly Novianti Thahir. "A text mining application of emotion classifications of Twitter's users using Naive Bayes method." In *2015 1st International Conference on Wireless and Telematics (ICWT)*, pp. 1-6. IEEE, 2015.
11. He Yunchao, Chin-Sheng Yang, Liang-Chih Yu, K. Robert Lai, and Weiyi Liu. "Sentiment classification of short texts based on semantic clustering." In *2015 International Conference on Orange Technologies (ICOT)*, pp. 54-57. IEEE, 2015.
12. Li Gang, and Fei Liu. "Sentiment analysis based on clustering: a framework in improving accuracy and recognizing neutral opinions." *Applied intelligence* 40, no. 3 (2014):441-452.
13. Hima Suresh, "An unsupervised fuzzy clustering method for twitter sentiment analysis." In *2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, pp. 80-85. IEEE, 2016.

14. Nagamma P, H. R. Pruthvi, K. K. Nisha, and N. H. Shwetha. "An improved sentiment analysis of online movie reviews based on clustering for box-office prediction." In *International Conference on Computing, Communication & Automation*, pp. 933-937. IEEE, 2015.

AUTHORS PROFILE

Ms. Susmitha Peeka, studying Bachelor of Technology, Department of Computer Science and Engineering, VR Siddhartha Engineering College, Vijayawada.

Ms. Sabiya Tabassum, studying Bachelor of Technology, Department of Computer Science and Engineering, VR Siddhartha Engineering College, Vijayawada.

Ms. Dhanushya Pallem, studying Bachelor of Technology, Department of Computer Science and Engineering, VR Siddhartha Engineering College, Vijayawada.

Ms. Sowmya Aravapalli, studying Bachelor of Technology, Department of Computer Science and Engineering, VR Siddhartha Engineering College, Vijayawada.

Dr. M. Sobhana, currently working as Sr. Assistant Professor, Department of Computer Science and Engineering, VR Siddhartha Engineering College, Vijayawada. She received Ph.D. degree in Computer Science and Engineering in 2018 from Krishna University. She has 13 years of teaching experience. Her research interests lies in areas such as Artificial Intelligence, Machine Learning, Data Analytics, Cyber Security and Software Engineering. She published 14 papers in National and International journals and also published 3 patents.