Tweet-Based Sentiment Analyzer

Gresha Bhatia, Chinmay Patil, Pranit Naik and Aman Pingle

Abstract People, these days, express their opinions regarding any particular topic or issue widely on social media. One such popular social media platform among masses is twitter with over 320 million monthly users. Users also express their thoughts on any political announcements or decisions taken by a particular party. Analyzing these tweets on a specific topic can help in determining what people think about measures undertaken by the government. It will give an idea on how many percent of people are in favor of any announcement, and how many of them stand against it. This will in turn provide areas of improvement for the ruling or opposition party. This paper thus aims on finding sentiments of tweets on a political leader, some party or announcements like a union budget. This can further be generalized to any particular measure undertaken by any organization.

Keywords Sentiment · Twitter API · Scraping · Live graphing · Multinomial Naive Bayes

1 Introduction

The number of users on Internet is increasing day by day with over 3.9 billion people in the world accessing Internet for their day to day tasks. This also leads to people utilizing various social media sites and connecting to their acquaintances. Thus, individuals tend to posit their views on various issues or advancements on social media networks. Such arguments or thoughts of people are found to be useful for

G. Bhatia · C. Patil (✉) · P. Naik · A. Pingle
Vivekanand Education Society's Institute of Technology (V.E.S.I.T), Chembur, Mumbai, India
e-mail: 2016.chinmay.patil@ves.ac.in

G. Bhatia
e-mail: gresha.bhatia@ves.ac.in

P. Naik
e-mail: 2016.pranit.naik@ves.ac.in

A. Pingle
e-mail: 2016.aman.pingle@ves.ac.in

S. Fong et al. (eds.), *ICT Analysis and Applications*, Lecture Notes in Networks and Systems 93, https://doi.org/10.1007/978-981-15-0630-7_36

analysis of propensity of people toward a particular side through their contention on social networks. Advantage of such analysis can be used by various political parties to recognize the mindset of the population. This forms the base of a Sentiment Analyzer.

Twitter is a popular microblogging site which lets people put forth their thoughts on to a worldwide platform [1]. The popularity of twitter is bolstered by the fact that 500 million tweets are posted everyday. This in turn provides large amount of data available for any topic with people casting their views on it. Twitter allows users to put forth their thoughts in a short text with a limit of 280 characters per tweet. Also, being in the category of blogging, twitter has more users using textual format to contend their views on a pool of issues. Twitter allows us to view various trends and lets us determine the inclination of users on it, i.e., either positive or negative. The above factors make analysis on the data faster, easier, and more accurate. As a result, we narrowed down our approach to a Tweet-Based Sentiment Analyzer from a Sentiment Analyzer [2].

Despite having an enormous amount of data, the text tweeted by users often contains discrepancies. The tweets are often multilingual, have acronyms used, contain various emojis, misspelt words, incorrect grammar, etc.

Such inconsistencies act as noise in data [3]. However, we can still classify the data into positive and negative tweets with the help of significant words or terms stated in the text. The paper thus depicts the approach of the project to classify such data by tackling these difficulties [2, 4].

2 Related Work

Twitter being a humongous platform for people to express their opinions, its data was analyzed by different researchers for varying purposes. Azam et al. [2] built a system to cluster tweets based on similarity using Markov Clustering technique with each cluster depicting an event. Various scenarios considered were for Israel—Gaza conflicts, Delhi assembly election and union Budget 2015. This was carried out by considering tweets as nodes in a social graph and weighted edge between them representing the similarity between the tweets. Norman et al. [5] performed sentiment analysis by gathering English tweets on demonetization and Indian Budget 2017. They used a Naive Bayes Classifier to predict sentiment of tweets fetched in real time to classify them into either positive, negative or neutral. The results helped to determine the feeling and estimation of the general population about the government's call to demonetization and its outcome on the proposed Budget in 2017. Naiknaware [6] worked upon estimating the inclination of people by scrutinizing the tweets of Union Budget of India from 2016 to 2018 in order to classify them into three classes, viz., positive, negative, or neutral. Kaur [7] worked upon improving the accuracy of a sentiment analyzer by proposing a system design that combined Lexicon-based and machine learning approaches. This also brought into light an

approach of hybrid model which may consist of multiple machine learning methods like SVM, Naive Bayes, etc. to determine the polarity of a tweet. Sarlan et al. [8] in their research, developed a model to obtain opinion of customers on an organization or company which will turnout to be beneficial for the company by measuring the perceptions of their customers. The model gave output in the form of a pie chart on an HTML page after classifying the tweets into two classes, i.e., positive and negative. The accuracy was improved by incorporating Natural Language Processing before actually classifying the data. Verma et al. [9] worked upon opinion mining for movies to be released in India in real time. The tweets were streamed in real time with help of a Twitter Streaming API. This helps in determining the mood of viewers and how the movie will perform in box office upon its release. Rahman et al. [10] proposed an approach to sentiment analysis on various topics by categorizing tweets into sentiments by a trained model on Machine Learning algorithm, i.e., Naive Bayes Method. Guha et al. [4] proposed a system for analyzing twitter data of SemEval 2015 by training a linear SVM.

2.1 Lacuna of Existing Systems

The existing systems classify data for English-based tweets only, i.e., there is no multilingual support. Also, these systems do not fetch data in a dynamic manner or produce tweet sentiments in real time along with no functionality to get and classify historical tweets for analysis purpose.

3 Proposed System

- The user first logs into the system.
- Then, he can choose between two options:
  - Get the sentiments of old tweets.
  - Get the sentiments of live tweets.
- The old tweets are fetched using web scraping and the live tweets are fetched using Twitter API.
- These tweets are processed and then fed to the trained model.
- Training the model:
  - First, the dataset is preprocessed.
  - Then it is fed to the algorithm which outputs the model.
- The results obtained from the model are plotted as graphs and presented to the user (Fig. 1).

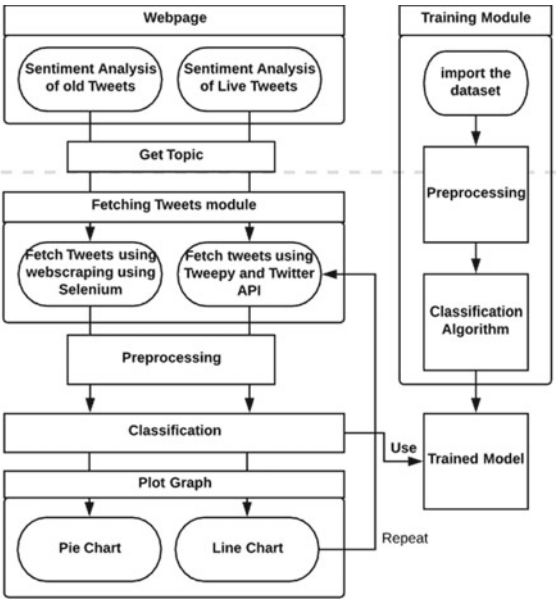


Fig. 1 Block diagram

4 Methodology

The project was developed in various phases. First, dataset was collected for training the classifier after which an appropriate algorithm was selected to categorize tweets into different classes. Tweets were captured based on users' input text using two main approaches, i.e., through Twitter API and Web Scraping. The input tweets were fed to the trained model to predict the sentiment. The output of the algorithm was then displayed to the end user.

4.1 Dataset Used

A predefined dataset of tweets by Indians on a Union Budget or government decision is not readily available. As a result, a movie review dataset for short text is used for training purpose. The dataset comprises of two different types of tweets in two files—positive and negative—with each having over 5000 movie reviews.

These files are then combined and shuffled randomly to obtain a mixed dataset with both positive as well as negative tweets. From a line of text, according to the study, an adjective plays the most vital role in determining the polarity of the sentence [11]. A ratio of positive to negative occurrence of an adjective is calculated to find in which of the two cases positive or negative the word is more associated. If the word "excellent"

occurs 30 times in positive classified data and only 3 times in negative data, then it is more closer to the positive side. Thus, adjectives are extracted as features of a sentence and mapped to the respective polarity. All such words form feature sets which are classified according to frequency of occurrence in the dataset. Among them, top 5000 or most frequent 5000 feature sets are picked and the corresponding model is trained based on these selected frequent feature sets along with picking 3000 sentences from the shuffled dataset. The trained model is then tested on next 1000 sentences to test the accuracy of the classifier.

4.2 Classification Algorithm

The classification algorithm used for predicting sentiment of tweets is Multinomial Naive Bayes Classifier as it is suitable for classification with discrete features such as text based classification [12]. Under Naive Bayes assumption we have:

$$p(f_1, f_2, \ldots, f_i|c) = \prod_{i=1}^{n} p(f_i|c)$$

$$p(f_i|c) = p(c|f_i) * p(f_i)/p(c)$$

The term Multinomial Naive Bayes lets us know that each $p(fi|c)$ is a multinomial distribution, rather than some other distribution. $p(fi|c)$ denotes probability that fi lies in class c [13].

This works well for data which can easily be turned into counts, such as word counts in text. Consider following training dataset (Table 1).

Let us determine whether the statement "overall budget is good" results in a positive statement or negative statement [14] (Fig. 2).

Since positive probability is greater as compared to negative, the text "Overall budget is good" is classified as "Positive".

Table 1 Training dataset

Text	Sentiment
"Education has become easier"	Positive
"It is affordable. Overall, a good move"	Positive
"A truly horrible decision by XYZ"	Negative
"It is a very good decision"	Positive
"They will suffer"	Negative

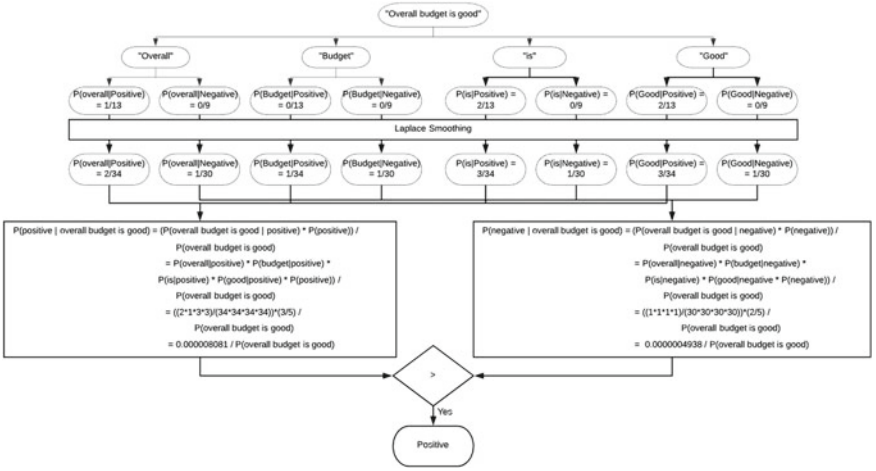


Fig. 2 Calculation of probability

4.3 Gathering Tweets

The project includes analysis of live tweets as well as old tweets or historical tweets. Thus, two different approaches have been used. The former uses Tweepy while the latter applies concept of web scraping using Selenium. Scraping is used, as twitter's privacy policy provides truncated old tweets using its API [15].

Live Tweets Using Tweepy. In order to fetch live tweets, Tweepy and Twitter API are used [16]. Steps to get Twitter keys:

- Apply for twitter developer account.
- Click on "Create an Application".
- Fill the details of the Application.
- The access tokens will then be available.

Tweepy handles the authentication, connection, creation, and destruction of the session.

Web Scraping using Selenium Webdriver. For fetching the old tweets, scraping of Twitter pages is done using Selenium Webdriver. Basically, Selenium is an automation testing tool. It can be used to perform various browser actions by writing a program [17]. Now, twitter is a dynamic website which loads more content upon scrolling with changing HTML as compared to other static web pages which have a fixed HTML code. As a result, we require a dynamic web scraping tool. Selenium is thus apt for the requirement. It simulates a human browsing the twitter pages loading more tweets by pressing page the down button. More the webdriver scrapes, more tweets are acquired from the HTML for sentiment prediction.

Based on the topic to be analyzed, a URL of search query is generated. Then, the corresponding page is visited on a browser. Followed by, all the content having "body" as the tag name being fetched. From the body, the "div" tag which has tweet text is reached and the tweet is captured.

4.4 Processing Tweets

A captured tweet contains variety of languages, emoticons, and noise. All of this has to be processed first to obtain a generalized format before predicting its polarity. So, it is passed through three different phases after which the sentiment is determined.

Convert Emojis to Literal Meanings. Tweets comprise of emoticons which play a vital role in expressing the sentiment of the user [18]. As a result, the system converts them into their meaning in textual format. This functionality is achieved using python module "Emoji" and the function being called as "demojize" in its documentation. E.g., the emoji ":)" is converted to the text "smile".

Cleaning of Tweets. Tweets contain URLs if it has an image associated to it. In this case, the URL is also fetched with the text and it acts as noise in data and should be eliminated [19]. This is removed using python module called "preprocessor" [20]. It also removes hashes from the text in case there is any hashtag present. Input statement: "The decision is good. https://xyz.com/image.png" is processed to give "The decision is good."

Translation. Tweets captured can be in various languages and not just English. However, the model recognizes and is capable of processing only English language. Hence, "GoogleTrans" Python library is used to detect the language of tweets and translate them into English wherever required. The API is also capable to translate a tweet in some other language typed in English to English language.

4.5 Output Representation

Both the methods incorporated for sentiment analysis depict the results in different fashion. They are as follows:

Line Graph. Live capturing of tweets generates a live graph which updates continuously based on time. The X-axis represents time and the Y-axis shows sentiment value. The graph initially starts with the value of Y-axis being 0. When a tweet is classified as positive, the Y value increments by 1 else it is decremented by 1. This shows current mood of people on a particular topic. An upward moving graph denotes that people are happier about a decision or there is a positive feedback and a downward moving graph suggests a negative feedback (Fig. 3).

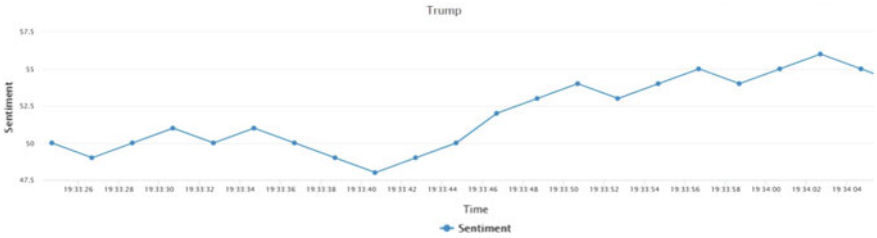
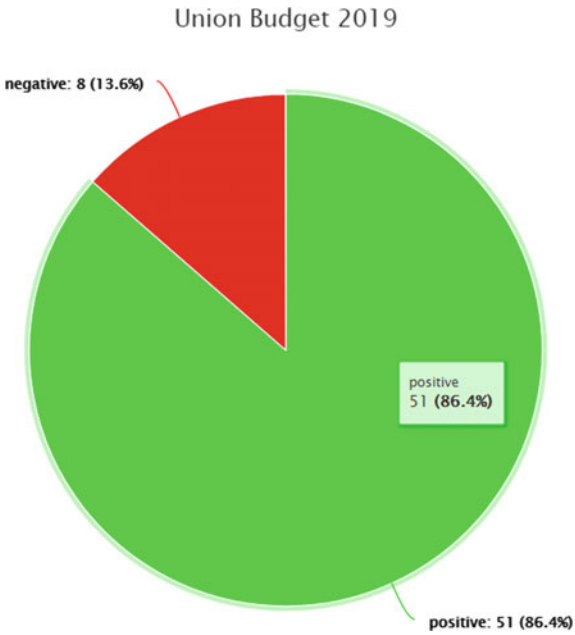


Fig. 3 Live sentiment analysis

Fig. 4 Scraped tweets analysis



Circular Statistics. Scraping of twitter gives its analysis in the form of a pie chart with a view of the number of tweets analyzed based on amount of scrolling done. The pie chart shows the percentage of tweets classified as positive and negative (Fig. 4).

4.6 Conclusion

The developed Tweet-Based Sentiment Analyzer can be used for analyzing various decisions or policies undertaken by the government to get an overall view of the public reaction. The sentiment analyzer provides an accuracy of about 76% upon taking multilingual input and an accuracy about 85% for input tweets with English language

only. The multilingual inputs' text is converted to English using GoogleTrans Python library each time for the classifier to recognize the input text. The accuracy of the classifier upon considering multiple languages decrease as the python library may at times incorrectly translate the tweet thus resulting in wrong predictions at a few instances. This problem, however, does not occur when considering English tweets, thus resulting in a better accuracy. This is achieved using Multinomial Naive Bayes algorithm to classify tweets and output is represented in the form of line graph when graphing the tweets live and pie chart when using web scraping to predict sentiment of old tweets.

References

1. Phand SA, Phand JA (2017) Twitter sentiment classification using stanford NLP. In: 2017 1st International conference on intelligent systems and information management (ICISIM)
2. Azam N, Jahiruddin, Abulaish M, SMIEEE, Haldar NAH (2015) Twitter data mining for events classification and analysis. In: Proceedings of the 2nd international conference on soft computing and machine intelligence (ISCMI'15). IEEE CPS, Hong Kong, Nov 23–24
3. Wang H, Can D, Kazemzadeh A, Bar F, Narayanan S (2012) A system for real-time twitter sentiment analysis of 2012 U S presidential election cycle. In: Proceedings of the 50th annual meeting of the association for computational linguistics. pp 115–120, Jeju, Republic of Korea, July 8–14
4. Guha S, Joshi A, Varma V (2015) Sentibase: sentiment analysis in twitter on a budget. In: SEM 4th joint conference on lexical and computational semantics denver, Colorado, USA. Report No: IIIT/TR/2015/-1
5. Norman J, Mangayarkarasi R, Vanitha M, Praveen Kumar T, UmaMaheswari G (2017) A Naive-Bayes strategy sentiment for sentiment analysis on demonetization and Indian budget 2017-case-study. Int J Pure Appl Math 117(17):23–31. ISSN: 1311-8080
6. Naiknaware B (2018) Peoples opinion on Indian budget using sentiment analysis techniques. Int J Res Eng Appl Manag (IJREAM). ISSN: 2454-9150, Special Issue-NCCT (2018)
7. Kaur J (2016) A review paper on twitter sentiment analysis techniques. Int J Res Appl Sci Eng Tech (IJRASET) 4(X). Guru Nanak Dev Engineering College, Ludhiana, Oct 2016, IC Value: 13.98, ISSN: 2321-9653
8. Sarlan A, Nadam C, Basri S (2014) Twitter sentiment analysis. In: 2014 International conference on information technology and multimedia (ICIMU). Putrajaya, Malaysia, Nov 18–20
9. Verma A, Singh KPA, Kanjilal K (2015) Knowledge discovery and twitter sentiment analysis: mining public opinion and studying its correlation with popularity of Indian movies. Int J Manag (IJM) 6(1):697–705. ISSN 0976–6502
10. Rahman E-U, Sarma R, Sinha R, Sinha P, Pradhan P (2018) A survey on twitter sentiment analysis. Int J Comput Sci Eng 6(11). Open Access Survey Paper, India, e-ISSN: 2347-2693
11. Kouloumpis E, Wilson T, Moore J, Twitter sentiment analysis: the good the bad and the OMG! In: Proceedings of fifth international AAAI conference on weblogs and social media (ICWSM)
12. Xu S, Li Y, Wang Z, Bayesian multinomial naive bayes classifier to text classification. In: International conference on multimedia and ubiquitous engineering international conference on future information technology
13. Heba M, Ismail, Harous S, Belkhouche B (2016) A comparative analysis of machine learning classifiers for twitter sentiment analysis. In: 17th International Conference on Intelligent Text Processing and Computational Linguistics-CICLing
14. Shinde PD, Rathod S (2018) A comparative study of sentiment analysis techniques. Int J Innov Adv Comput Sci 7(3). ISSN 2347–8616

15. Tugores A, Colet P (2013) Mining online social networks with Python to study urban mobility. In: Proceedings of the 6th European conference on python in science
16. Dhanush M, Ijaz Nizami S, Patra A, Biswas P, Immadi G (2018) Sentiment analysis of a topic on twitter using tweepy. Int Res J Eng Tech 5(5):2881. e-ISSN: 2395-0056
17. Jagannatha S, Niranjanamurthy M, Manushree SP, Chaitra GS (2014) Comparative study on automation testing using selenium testing framework and QTP. IJCSMC 3(10):258–267. ISSN 2320–088X
18. Anand N, Kumar T (2017) Text and emotion analysis of twitter data. Int J Comput Sci Eng 5(6). Open Access Research Paper, e-ISSN: 2347-2693
19. Chirawichitchai N (2013) Sentiment classification by a hybrid method of greedy search and multinomial naïve bayes algorithm. In: 2013 Eleventh international conference on ICT and knowledge engineering
20. Gupta B, Negi M, Vishwakarma K, Rawat G, Badhani P (2017) Study of twitter sentiment analysis using machine learning algorithms on python. Int J Comput Appl 165(9):(0975–8887)