Twitter Sentiment Analysis with Textblob

Nevil Gajera[1]
Department of Computer Science and Engineering,
Devang Patel Institute of Advance Technology and Research (DEPSTAR), Faculty of Technology and Engineering (FTE), Charotar University of Science and Technology (CHARUSAT)

Kishan Chanchad
Department of Computer Science and Engineering,
Devang Patel Institute of Advance Technology and Research (DEPSTAR), Faculty of Technology and Engineering (FTE), Charotar University of Science and Technology (CHARUSAT)

Sneh Vora
Department of Computer Science and Engineering,
Devang Patel Institute of Advance Technology and Research (DEPSTAR),
Faculty of Technology and Engineering (FTE), Charotar University of Science and Technology (CHARUSAT)

Abstract:- In the World, many social media sites exist like Twitter, Instagram, Facebook, Snapchat, etc. and data posted by people on these social media sites are increasing quickly that containing audio, video, text, and images. People use this site to share their thoughts and opinion and sometimes share their opinion and thoughts towards any company. For this, we have chosen twitter and applied sentiment analysis. In this paper, we discuss a method of data extraction through API, data cleaning, and use text blob python library for sentiment analysis.

Keywords:- Twitter, sentiment, opinion mining, social media, natural language processing, sentiment analysis.

I. INTRODUCTION

Social Media Websites are carrying a sea of data. Social Media sites have given a right to speak to every person who can access or use them. Twitter is used by a large number of people to write their emotions, opinions about their daily life, and a company or any organization reviews and opinions. Twitter is challenging because its users have to express their views in one or two key sentences and it can be seen as a good reaction to what is happening around the world. Sentiment analysis automates the extraction or classification of sentiment and views using text analysis, natural language processing, and computational approaches. This sentiment analysis benefits many fields like Customer information, Marketing field, books, mobile application, Social media, and websites. Many companies hire analysts who have a job to extract the emotions of people behind these posts or tweets. This helps businesses to get a good review about a product or service which helps them know public opinion and in addition, they make a better product in the future. In this project, we use the python text blob library for text classification. There are two ways to extract tweets using Twitter's official API and data scraper. For this project, we preferred API for collecting datasets.

II. LITERATURE REVIEW

In October 2017 Kirti Huda, Mrunmayee Deshpande, and Neshat Karim gave information on Classification Techniques for Sentiment Analysis of Twitter tweets data. In classification, there are mainly three techniques Naive Bayes, M E, and SVM. The author uses a pattern-based technique for feature extraction. In this, for feature extraction, The n-gram algorithm is used, which assigns a priority to each word that needs to be classified. In the last step of classification, they use a support vector machine (SVM). After the classification, they conclude that the accuracy of the E- Pattern-based algorithm is given a more accurate result than pattern based algorithm, and in terms of time also E-pattern based algorithm takes less execution time.

In June 2017 Shivam Singh, Sonal Agarwal, And Sakshi Agarwal proposed a Real-Time Twitter Sentiment Analysis. In this research, they use Hadoop with natural language processing. 1st is the Ingestion of tweets into HDFS in this Tweets are ingested from Twitter streaming using Twitter 4j API. 2nd is Post Processing, Construction of n- grams, and Spelling correction. 3rd is Query processing using HIVE Once the tweets are ingested into HDFS. Excel uses the ODBC driver to get the processed data in the form of graphs, geographical location, and charts-based data because the culture and diversity of a location matter very much.

In February 2018 Sahar A. El_Rahman, Feddah Alhumaidi AlOtaibi, And Wejdan Abdullah AlShehri proposed a Sentiment Analysis of Twitter Data. The author used sentiment analysis to classify English tweets about two famous restaurants that are McDonald's and KFC. In this method they use some packages and libraries, some packages are Twitter, R0Auth, and word cloud, after preparing tweets using an unsupervised learning algorithm they used a lexicon- based model used to classify Twitter tweets. To train the model they use different supervised algorithms: Naive Bayes, SVM, random forest, decision tree, and maximum entropy. For Accuracy They use Recall, Precision, F-score, And Cross- validation.

In September 2019 Brahmananda Reddy, D.N.Vasundhara, and P. Subhash proposed Sentiment research on Twitter data. This system was completed in seven stages. In this system, they overcome the drawbacks for better understanding the emotions they classified emotions into 7 categories ex. Strongly Positive, Positive, Weakly positive. Instead of static data, they use real-time data using Twitter API by giving a username or hashtag and They can look at a specific person's tweets or hashtags. In this research, the author uses a Naïve bye classifier and they use video games review data sets for training and testing. The Bayes theorem is used in the Naive Bayes technique, which uses a probabilistic learning function.

In November 2019 Senthil Jayavel, Jayakumar Sadhasivam, and Ramesh Kalivaradhan proposed a Survey of varied Algorithms utilized in Twitter For Sentiment Analysis. During this development for Classifying the text, they use Naive Bayes, Maximum Entropy, and SVM machine learning techniques. In the analysis of different sentiment analysis algorithms, they saw that the performance of supervised learning is better than unsupervised learning, and failure can occur in the case of supervised learning because of insufficient training data.

In 2019 Vishal A. Kharde and S.S. Sonawane proposed A Survey of Techniques for Sentiment Analysis of Twitter Data. Sentiment analysis includes sentiment classification, sentiment extraction, opinion summarization, and subjectivity categorization among other things. Author Mathematically represent an opinion as a quintuple (o, t, h, so, f). The preprocessed dataset has many distinctive properties. They also do the same survey with SVM using bigram and unigram and they get bigram with more accuracy after that they experiment with Maximum Entropy and the last results show that bi-gram with SVM is given a more accurate result.

In February 2019 Faizan proposed Twitter Sentiment Analysis using Python. Here the author discusses 5 stages of sentiment analysis data collection, data processing, Feature selection, Model Selection, and Model Evaluation. The author uses Twitter API and takes tweets of people airlines and they use text blob python library for tweets pre-data annotation of polarity. The second stage of preprocessing need more convenient raw data in author use the data that are derived from Twitter using Twitter API which is initially totally unstructured for this matter they use a python module named regular expression. Now, in the third stage of feature selection, many features are available like Unigram, N-gram, Bigram, Subjective, POS tagging, and objective features.

In March 2019 Hetu Bhavsar, and Richa Manglani, proposed a Sentiment Analysis of Twitter Data using Python. They used, SVM, decision tree, and Adaboosted decision tree- based hybrid sentiment classification model are used to increase the accuracy of Twitter sentiment analysis. The model of Twitter data analysis will be implemented using Anaconda python. The authors used a gaggle dataset that had been crawled and categorized as positive or negative. In Kaggle data they provided emotions, usernames, and hashtags, which are required to be processed and converted into a better form.

$$p(\frac{A_k}{X}) = \frac{p(A_k)p(X/A_k)}{p(X)}$$

In 2019 Vishal A. Kharde, S.S. Sonawane proposed A Survey of Techniques for Sentiment Analysis of Twitter Data. Sentiment analysis includes sentiment classification, sentiment extraction, opinion summarization, and subjectivity categorization among other things. Author Mathematically represent an opinion as a quintuple (o, t, h, so, f). The preprocessed dataset has many distinctive properties. They also do the same survey with SVM using bigram and uni-

gram and they get bigram more accuracy after that they experiment with Maximum Entropy and the last results show that bi-gram with SVM is give a more accurate result.

In Jan 2019 Udayan Ghose, and Ankita Sharma proposed a Sentimental Analysis of Twitter Data concerning India's General Elections. Text analytics uses NLP to accomplish this. The "twitteR" package in the R program is used to extract tweets from Twitter. They use R language for mining Twitter tweets. In this paper, the author uses 3 levels that are Sentence level, Document-level or Aspect level. The term "document level" refers to the classification of opinions over the full document. All sentences are checked at the sentence level for polarity, which might be positive or negative; mixed opinions may or may not be considered for asentence. Aspect level, also known as entry-level, Then they do textual content analysis using Rapid-Miner, Anyone can use them to get insight into emotional strength, sentiment subjectivity, and sentiment polarity, as well as assign a sentiment score.

In February 2020 Dr. KB Priya Iyer and Dr. Shakti Kumaresh did a sentiment analysis of people during corona pandemic. The researcher focuses on the sentiment analysis of COVID-19 using Twitter data. For sentiment analysis, they use a machine learning approach. For classification, the author uses the Naïve Bayes classifier. The author uses a proposed model dataset collected from Twitter API and this contains IDs and the sentiment score of the tweets during the COVID-19 pandemic and processes through a set of five phases. The author briefly described all five phases of the proposed model.

III. METHODOLOGY

- Start
- Read hashtag or username from the user
- Request twitter API through tweepy get tweets - tweepy.Cursor(api.search, q=hashtag).items(500):
- Make data frame of tweets - df = pd.DataFrame(msgs)
- Clean tweets using re - text = re.sub('@[A-Za-z0–9]+', '', text)
- Get polarity of tweets using textblob - twitter['Polarity']=twitter['Tweets'].apply(getPolarity)
- Get Analysis score positive, negative, neutral - positive=df.loc[df['Analysis'].str.contains('Positive')]
- Calculate the percentage of positive, negative, neutral tweets divides by the total number of tweets - positive_per= round((positive.shape[0]/df.shape[0])*100, 1)
- Give the Result
- Stop

- Text blob module work on below naïve byes theorem
NB is a probabilistic technique that computes sentiment distribution over data using Bayes theorem. NB, on the other hand, dissects any text into a bag of words, disregarding the word placements entirely. The Bayes equation to predict the sentiment probability is:

P(Ak|x): Probability (conditional probability) of occurrence of event Ak given the event x is true. P(Ak) and P(x): Probabilities of the occurrence of event Ak and x respectively. P(x|Ak): Probability of the occurrence of event x given the event Ak is true.
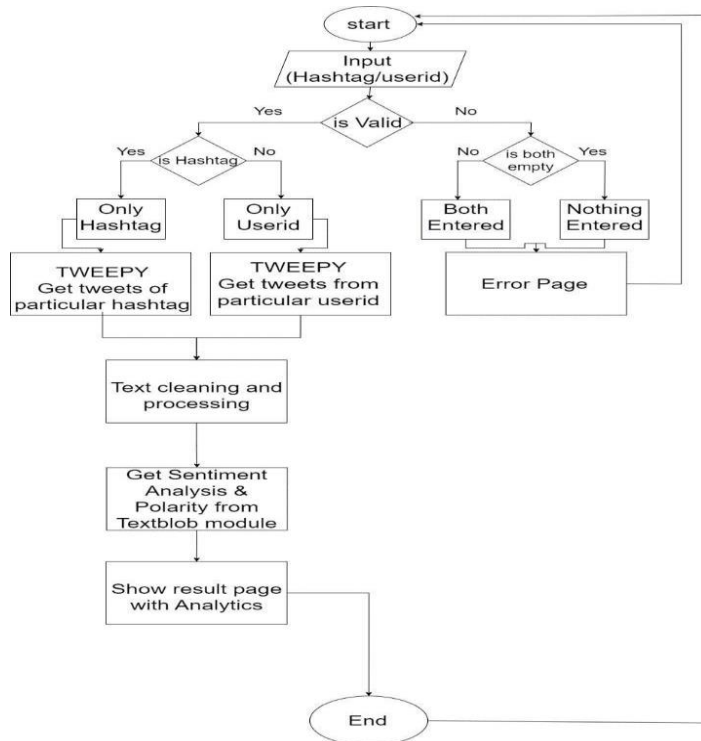
IV. FLOW CHART

Fig. 1: Flow chart

V. OUTPUT DICCUSSION

Output Discussion –

- The first user has to give the user id or hashtag of Twitter that they want to analyze their tweets sentiment analysis. If the user gives both user id and tag same time, then this will not work.
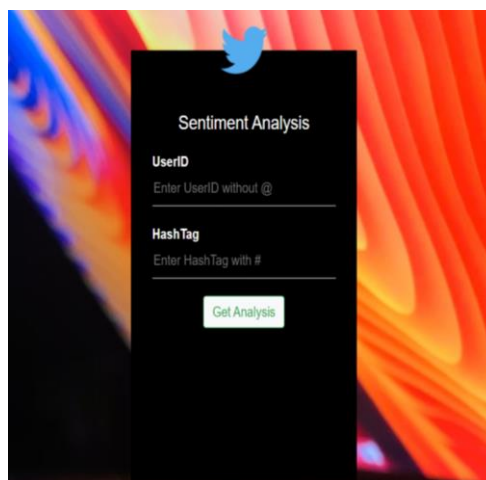

Fig. 2: interface where user have to provide *input.*

After clicking on the get analysis button get output with percentage how much percentage tweets are negative, positive, neutral.
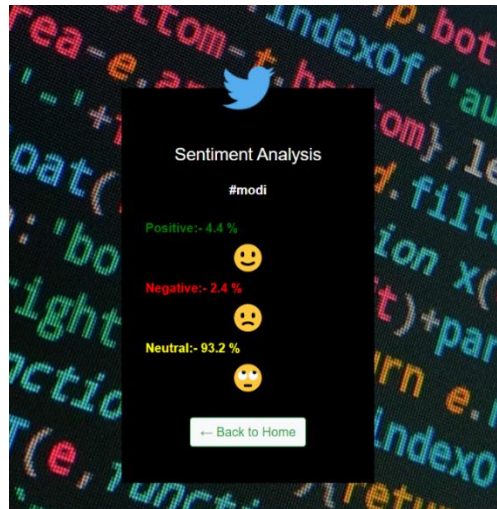

Fig. 3: Analysis output.

VI. FIGURES AND TABLES

Sr No.	Year & Author	Method or Algorithm or Technique	Accuracy	Advantage / Disadvantage
1	In 2017, Md Tabrez Nafis, Kirti Huda, Neshat Karim Shaukat	N-gram, SVM	90%	The proposed and current algorithms execution times are measured in terms of execution time. It has been determined that the upgraded pattern-based algorithm takes less time to execute.
2	In June 2017 Shivam Singh, Sonal Agarwal, And Sakshi Agarwal	Hadoop With NLP	85%	These Analysis can facilitate the process of decision making in various areas such as health care analysis, market analysis, weather forecasting, advertising analysis, fraud detection, traffic flow optimization etc.
3	In 2018, Ahmad Karim, Ali Hasan, SanaMoin, shahaboddin Shamshirband	Naive Bayes, SVM , Maximum Entropy, Decision Tree, Random Forest	85%	The feature vector's classification accuracy is tested using victimization. Thomas Bayes is a naive classification algorithm.
4	In 2018, Ahmad Karim, Ali Hasan, SanaMoin, shahaboddin Shamshirband	Textblob, Wordcloud	62.67%	In comparison to W-WSD and Text Blob, TextBlob has the highest accuracy, according to the experiment result data.
5	In Fab 2019, Faizan	Regular Expresion, K nearest neighbour algorithm	65.33%	Increase the accuracy of the model by using different deep learning techniques such as neural networks.
6	In March 2019 Hetu Bhavsar, Richa Manglani	SVM, Decision Tree and Adaboost Decision Tree based hybrid sentiment classification model	95%	It must also extract valuable text properties like bigrams that are more beneficial in sentiment analysis and extraction of properties like unigram not more accurate sentiment analysis.

7	In 2019 Vishl A. Kharde, S.S. Sonawane	Machine Learning and Lexicon Based Method, SVM, Naive Bais	74.56%	The most accurate learning methods are naïve Bayes and SVM, which can be considered the baseline, whereas lexicon-based algorithms can be quite useful in particular instances.
8	In 2019, D.N.Vasundhara, P. Subhash, A. Brahmananda Reddy	Naive Bayes	80%	Victimization is used to test the feature vector's classification accuracy. Thomas Bayes is a naive classifier.
9	In February, 2020 Dr. KB Priya Iyer and Dr. Shakti Kumaresh.	Naïve Bayes classification, Machine Learning Algorithms	70%	As the virus is spreading vigorously, the study needs to be carried out every week to have a better understanding of the sentiments of the people.
10	In 2020 Ankita Sharmaa, Udayan Ghoseb	R language, Rapid miner	87%	All sentences are checked at the sentence level for polarity, which might be negative or Positive; mixed opinion may be or may not be considered for a sentence.

Table 1: Methods used by authors for solving similar problem and accuracy achieved.

VII. LIMITATIONS

The tweets that we collected for these project purposes were in the English language which is the limitation of this project because many of the tweets are in other languages. For extracting the tweets, we use Twitter API that only allows collecting of the last 7 days.

VIII. FUTRE WORK

In The Future, We will work on how this sentiment analysis accuracy can further be increased and users can get a 99.99% accurate result. As this result is only based on one language we can increase the number of languages for sentiment analysis.

IX. CONCLUSION

Sentiment analysis/opinion mining wide-area real-time applications have many research limitations. Since fast internet growth, internet-related applications, Sentiment analysis – most interesting research area natural language processing community. In our project, we analyzed sentiments Tweets extracted from Twitter, classify them according to polarities.

Major limitations -Sentiment analysis:
- Spam and fake News detection.
- Classification Filtering limitation.
- Limited language available

REFERENCES

[1.] A.Pak and P. Paroubek. „Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In Proceedings of the Seventh Conference on international Language Resources and Evaluation, 2010, pp.1320-1326

[2.] Pan S J, Ni X, Sun J T, et al. "Cross-domain sentiment classification viaspectral feature alignment". Proceedings of the 19th international conference on World wide web. ACM, 2010: 751-760.

[3.] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & amp;Stede, M..“Lexicon basedmethods for sentiment analysis”. Computational linguistics, 2011:37(2), 267-307.

[4.] A. Sharma, and S. Dey, “Performance Investigation of Feature Selection Methods and Sentiment Lexicons for Sentiment Analysis,” Association for the advancement of Artificial Intelligence, 2012.

[5.] M. Hu, and B. Liu, “Mining and summarizing customer reviews,” 2004.

[6.] S.Lohmann, M. Burch, H. Schmauder and D. Weiskopf, “Visual Analysis of Microblog Content Using Time-Varying Co-occurrence Highlighting in Tag Clouds,” Annual conference of VISVISUS. Germany: University of Stuttgart, 2012.

[7.] M.Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, “ Lexicon-Based Methods for Sentiment Analysis,” Association for Computational Linguistics, 2011.

[8.] H. Saif, Y. He and H. Alani, “Alleviating Data Scarcity for Twitter Sentiment Analysis”.Association for Computational Linguistics, 2012

[9.] Neethu M,S and Rajashree R,” Sentiment Analysis in Twitter using Machine Learning Techniques” 4th ICCCNT 2013,at Tiruchengode, India. IEEE –31661

[10.] J. Silge and D. Robinson, Text Mining with R: A Tidy Approach. O'Reilly Media, 2017.

[11.] P.-W. Liang and B.-R. Dai, "Opinion mining on social media data," in Mobile Data Management (MDM), 2013 IEEE 14th International Conference on, 2013, vol. 2: IEEE, pp. 91-96.

[12.] M. Hagen, M. Potthast, M. Büchner, and B. Stein, "Webis: An ensemble for twitter sentiment detection," in Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015), 2015, pp. 582-589

[13.] C. Wu, L. Shen, and X. Wang, "A new method of using contextual information to infer the semantic orientations of context dependent opinions," in Artificial Intelligence and Computational Intelligence, 2009. AICI'09. International Conference on, 2009, vol. 4: IEEE, pp. 274-278.

[14.] R. Sharma, S. Nigam, and R. Jain, "Opinion mining of movie reviews at document level," arXiv preprint arXiv:1408.3829, 2014

[15.] I. Twitter, "Second Quarter 2016 Report," ed, 2016

[16.] I. Twitter, "Twitter IPO Prospectus," ed, 2013.

[17.] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," Business horizons, vol. 53, no. 1, pp. 59-68, 2010.

[18.] F. M. Kundi, S. Ahmad, A. Khan, and M. Z. Asghar, "Detection and scoring of internet slangs for sentiment analysis using SentiWordNet," Life Science Journal, vol. 11, no. 9, pp. 66-72, 2014.

[19.] "A. K. Jose, N. Bhatia, and S. Krishna, "Twitter Sentiment Analysis". National Institute of Technology Calicut, 2010".

[20.] "Tumasjan, A., Sprenger, T., Sandner, P., and Welpe, I. "Predicting elections with twitter: What 140 characters reveal about political sentiment.." In Proceedings of the Fourth International AAAI Conference onWeblogs and Social Media (2010), pp. 178–185, 2010".