A Survey On: Classification of Twitter data Using Sentiment Analysis

Pratima Deshpande[1], Purva Joshi[2], Diptee Madekar[3], Pratiksha Pawar[4], Prof. M.D. Salunke[5]
Department Of Computer Engineering, Shri. Chhatrapati Shivaji Maharaj College Of Engineering, Ahmednagar, India, Savirtibai Phule Pune University
[1]pratima98.pd@gmail.com, [2]purvajoshi5374@gmail.com, [3]dipteemadekar1998@gmail.com, [4]pratikhapawar29192@gmail.com, [5]salunkemangesh019@gmail.com

***Abstract*— Sentiment means classify the opinions which are in the form of text .As we are classifying the text it must having different unstructured contains with information of particular subject. There are various social media sites which gives us information with their thoughts but twitter is biggest among them where any one who having account on twitter can tweet their opinions of any subject in any certain or uncertain way. Hence we get extra scope in mining of such data. The analysis is done with the help of machine learning algorithms on the dataset. Classification is done by the classifier algorithms. The reviews data is used for performing sentimental analysis. This paper gives idea about how the analysis is done on twitter data by using various algorithms and machine learning concepts. It is a survey of different papers for analyzing the sentiments of text.**

Keywords*— *machine learning; sentiment analysis; naïve bayes; support vector machine; random forest.

I. INTRODUCTION

Social media sites like Twitter, Facebook, blogs and many online forums produced large amount of unstructured data. Day by day social media micro-blogging is becoming very popular for users to express their opinions on different type of events, products, services, etc. Millions of users share their feelings on twitter. Twitter allow users to share their opinions in less than 140 words. So sentiment analysis is easy on twitter data. Large amount of unstructured data is generating from social media which is classified as positive, negative and neutral. Positive sentiments are like appreciation for other's tweets and also contains movies, products, etc. On the other hand negative sentiments are like bitter words or dissatisfactory comments on any product, movie, event, etc. Whereas neutral words are neither positive nor negative sentiment. Many manufacturing companies uses social media data for analysing the popularity of the product. Sentiment analysis is also useful for researchers for particular research area where the opinions of people are very important.

A. Sentiment Analysis:

Sentiment analysis is kind of information mining that recognizes the specific data from given content and characterize it as positive, negative or neutral. Sentiment analysis is otherwise called opinion mining which infers the assessments of individuals. Sentiment analysis gives us the intusion on user's opinions about items, occasions or movies. It utilizes natural language processing (NLP) which is utilized to separate and investigate the given information.

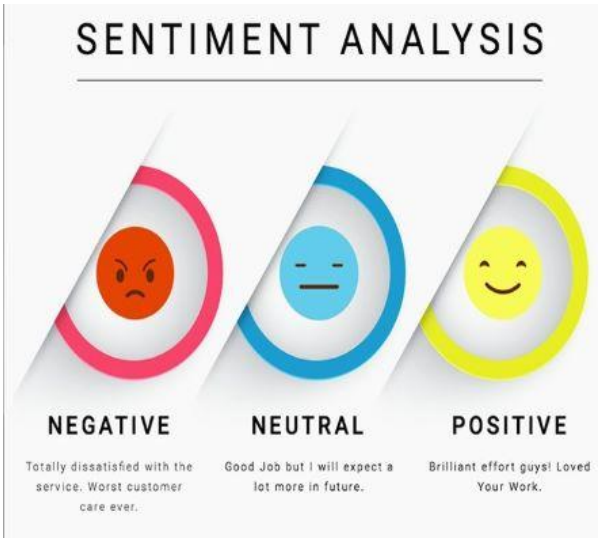

Fig. 1. Sentiment Analysis[12]

B. Natural Language Toolkit (NLTK):

The Natural Language Toolkit (NLTK) is a toolkit which manages the human language information used to develop the python program. It comprise of text processing libraries for tokenization, parsing, stemming, etc. NLTK is a standard library in python which works with human language. NLP is nothing but natural language processing which gives the interface among computers and human with the assistance of NLTK.

II. LITERATURE SURVEY

Garg, et al.[2] gives information about the sentiment analysis on twitter data post Uri attack, in that they conclude that the positive tweets survive in low amount, because too much victorious negative information is spread from people to other peoples for instance might start targeting a community in nation and it can lead to civil restlessness. They also discussed the trends for retweets and number of favorites.

Ahuja, et al.[4] describe on the surveys different approaches of clustering with respect to sentiment analysis and presents a way to find relationships between the tweets on the basis of polarity and subjectivity. In that they forming a cluster of the results from both the tools' score, From that we are able to group 'definitely' positive and 'definitely' negative tweets.

K. Lavanya, et al.[3] explain about the Sentiment Analysis on Twitter Using Multi-Class SVM. In that the

proposed algorithm gives the better accuracy. The algorithm classifies the tweets of different topics as positive, negative, neutral. The proposed method is evaluated across different topics and it outperforms in terms of recall, precision and F-score.

Huma, et al.[6] describe about the Sentiment Analysis on Twitter Data-set using Naive Bayes Algorithm. They conclude that the increasing the capability of sentiment analysis using Hadoop MapReduce. Also the neutrality of tweets leaps if the emoticons in the tweets/retweets are fed into the analysis.

Kudakwashe Zvarevashe, et al.[7] design a framework for sentiment analysis with opinion mining for hotel customer feedback. For sentiment polarity, he proposed framework in which sentiment dataset automatically prepared for training and testing data to extract that opinion received from review. For classification components of the framework different machine learning algorithms like Naive Bayes Multinomial, Sequential minimal optimization Compliment Naive Bayes and composite hypercube used for comparative analysis to find out suitable machine learning algorithm for framework. Textual datalike hotel reviews are used for sentiment analysis. Natural Language processing technique used for sentiment analysis and for automatic classification of sentiments computational linguistic technique used. Because mislabeled dataset leads to incorrect decisions he describe about intuition model and sentiment polarity based model and also discuss about how sentiment analysis can perform on feedback collected from the customer.

N. AZMINA M. ZAMANI, et al.[8] used lexicon based approach for transform unstructured data into meaningful lexicon. All that lexicons stored in database when manual identification done. For this text information extracted and then clustered into emotions. After analysis categorized them into happy, unhappy and emotionless and then result displayed by giving percentage to sentiment categories after that it is decided whether Facebook post get positive or negative based on response or comments. Software tool developed in java for sentiment classification using lexicon based approach for that predefined word list need to be stored in database. For efficient data analysis lexicon based approach is more efficient. While text analysis emoticons also considered to classify emotions which are able to support voice inflections and facial expressions. In this post not include any images. Lexicon based approach has comparable accuracy issue. He also used web server for data preprocessing on comment extraction then related words collected and stored in database for sentiment analysis. JSON library are used for automatic extraction of Facebook post comments.

Charu C. Aggarwal, et al.[9] conduct survey on text classification on wide variety of domains in text mining. Some of domains are mentions here. News filter and organization ex. text filtering- there are many web portals are available for news so all of these data are available in electronic form. In single day, various amount of such data generated by many organizations which is not possible to handle manually. Hence automatic categorization technique is useful to organize such data on web portal. b. Document organization and retrieval ex. digital library- contains many scientific research papers, articles, web documents. In different domain various methods are used to organize data. Hierarchical method is useful for organization and retrieval of data. Opinion Mining contains users feedback, comments and review which need to be categorized to get appropriate result. In opinion data is categorized in positive, negative and in neutral form the Email classification and spam filtering- It is classify by determining the either mail is junk or not.

Harpreet kaur, et al.[10] conduct survey on sentiment analysis and told how useful it is in decision making, business application and prediction and trend analysis. It contains basic idea about the polarity of the text by classify it into positive, negative and in neutral form. The main purpose is to know about the attitude of the comment writer or speaker. He stated that Sentiment analysis can also perform on audio, images and videos. It include two concept subjectivity and objectivity in which subjective text are those text which contains emotions while objective text are those which contains only factual information. Sentiment analysis can perform on various level like document, sentence and phrase level. Sentiment analysis is helpful in decision making by collecting reviews about particular object that it is good or bad. In business application it is helpful for company to achieve desired goal by fulfilling customers requirement to satisfy their need. It is useful for prediction and trend analysis by using public review on current topic.

jie Li, et al.[11] stated that sentiment of each statement can be calculated by the opinion structure which is obtain by dependency parsing. They also told how to calculate sentiment of sentiment sentence and short text. They proposed to use dependency parsing. He used short text from microblog to perform sentiment analysis in which statement dependency is analyzed and then dependency relation between the words considered. He also sentiment structure which include dependency parsing, sentiment relationship, sentiment relationship migration and modified distance. Microblog has word limitation of 140 words which makes microblog short text. Sentiment calculation done by using sentiment value of a sentence.

III. METHODOLOGY

A. Fetching data from twitter

Twitter is collection of large dataset, for performing the sentiment analysis on twitter data the data extraction processing is important. In compare to the other networking sites twitter gives users to share their views openly . With the help of twitter API twitter gives the expansive access to tweets.

To fetch data from twitter steps:

- The twitter app is created by twitter to access the twitter developer account which having identical user name and password. By this the credentials are obtain to get string form of tweets from the API.
- After getting API the tweepy library is used which is python library. With the help of tweepy the interactions with tweets for performing analysis become simple.

As we get text dataset then further process is start.

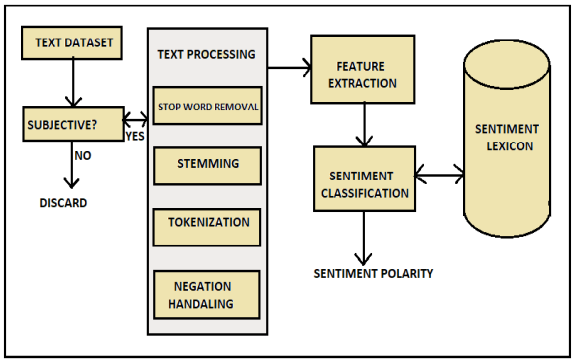

Fig. 2. Sentiment Analysis Process[10]

B. *Text Processing*

The data is in the form of text. To perform analysis the text processing is very important. Because the text contain lot of words so text processing is must.

- STOPWORD REMOVAL
  Sentence may contain pronouns (he, she, It), nouns, verbs, adjectives, articles (a, the), prepositions (here, near) these doesn't make any sense. Hence their removal is important.
- STEMMING
  It is nothing but removal of prefix and suffix. Consider example 'playing', 'played' where play is actual word and 'ing' is suffix which is not important.
- TOKENIZATION
  It is the most important process it also called as chopping. It is used for breakdown the sequence of text sentence to tokens, phrases or words.
- NEGATION HANDLING
  Consider example "Movie was not so good" here the positive and negative both words are occurred but the polarity of sentence is totally indicates the negation[10].

After processing on text next is extraction of useful words.

C. *Feature Extraction*

As the processing on text is takes place then the extraction of useful words is important. The removal of unnecessary words takes place by text processing. In feature extraction the actual words which are useful to perform analysis are extracted.

D. *Sentiment Classification*

Classification takes place in two ways first is supervised and second is unsupervised.

- Naïve Bayes
  This is important algorithm of machine learning which is mainly used for classification. It is very effective and highly accurate. Its classification rate is [7,8] [1]. It comes under supervised learning algorithm with probabilistic classifier[10]. It gives probability in the form of positive negative and neutral.
- SVM
  Support vector machine is supervised machine learning technique under the linear classification which performs the classification based on linear characteristics[10]. It uses hyperplane to make classes separate. It is a non-probabilistic classification algorithm.

E. *SENTIMENT LEXICON*

It is collection of words in which contain different word having score that indicates the positive, negative and neutral nature of sentence. The sentence which having high score it will define its polarity.

IV. CONCLUSION

Twitter is an amazing data source which individuals over the world meet up to share their opinion on different issues. Consequently, it gives a huge platform to researchers to get a huge measure of crude information. This crude information processing serves to analyse the opinions of users. From this survey, we have studied the different types of classification algorithms for text analysis. The data mining technique is used for text extraction from given data. From this the text is categorised as positive, negative or neutral.

REFERENCES

[1] Medha Khuran, et al. "Sentiment Analysis Framework of Twitter Data using Classification", 5th IEEE International Conference,2018.

[2] P. Garg, H. Garg, and V Ranga, "Sentiment analysis of the Uri terror attack using Twitter," Computing, Communication and Automation (ICCCA), 2017

[3] K. Lavanya and C. Deisy. "Twitter sentiment analysis using multi-class SVM," Intelligent Computing and Control (I2C2), International Conference on. IEEE, 2017.

[4] Ahuja, Shreya, and G. Dubey, "Clustering and sentiment analysis on Twitter data," 2nd International Conference on Telecommunication and Networks (TEL-NET), IEEE, 2017.

[5] M. Trupthi, , S. Pabboju, and G. Narasimha. "Sentiment analysis on twitter using streaming API," Advance Computing Conference (IACC), IEEE 7th International. IEEE, 2017.

[6] P. Huma, and S. Pandey, "Sentiment analysis on Twitter Data-set using Naive Bayes algorithm," Applied and Theoretical Computing and Communication Technology (iCATccT), 2nd International Conference on. IEEE, 2016.

[7] Zvarevashe, Kudakwashe, and Oludayo O. Olugbara, "A framework for sentiment analysis with opinion mining of hotel reviews," In Information Communications Technology and Society (ICTAS),2018 Conference, pp. 1-4. IEEE, 2018

[8] N. Zamani, M. Azminam, "Sentiment analysis: Determining people's emotions in facebook," University Teknologies MARA,Malaysia 2013.

[9] C. Aggarwal, and C. Xiang Zhai. "A survey of text classification algorithms," Mining text data. Springer, Boston, MA, pp. 163-222, 2012.

[10] Kaur, Harpreet, and Veenu Mangat, "A survey of sentiment analysis techniques," I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(ISMAC), International Conference, pp. 921-925, IEEE, 2017.

[11] Li, Jie, and Lirong Qiu, "A Sentiment Analysis Method of Short Texts in Microblog," Computational Science and Engineering (CSE) and Embedded and Ubiquitous Computing (EUC), IEEE International Conference, vol. 1, pp. 776-779, IEEE, 2017.

[12] https://www.kdnuggets.com/2018/03/5-things-sentiment-analysis-classification.html

[13] Sheeba Naz , Aditi Sharan , Nidhi Malik "sentiment classification on twitter data using support vector machine", 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)

[14] Huma Parveen , Prof. Shikha Pandey "Sentiment Analysis on Twitter Data-set using Naïve Bayes Algorithm", IEEE 2016.

[15] Rasika Wagh, Payal Punde, "Survey on Sentiment Analysis using Twitter Dataset", Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA 2018) IEEE Conference IEEE.

[16] Alaa S. Al Shammari, "Real-time Twitter Sentiment Analysis using 3-way Classifier", 2018 IEEE.

[17] Sahar A. El_Rahman, Feddah Alhumaidi AlOtaibi, Wejdan Abdullah AlShehri, "Sentiment Analysis of Twitter Data", 2019 IEEE.

[18] Rincy Jose, Varghese S Chooralil,, "Prediction of Election Result by Enhanced Sentiment Analysis on Twitter Data using Classifier Ensemble Approach", IEEE 2016.

[19] Ali Hasan 1, Sana Moin 1, Ahmad Karim 2 and Shahaboddin Shamshirband "Machine Learning-Based Sentiment Analysis for Twitter Accounts", Math. Comput. Appl. 2018, 23, 11.

[20] M. Saif, "Sentiment analysis: Detecting valence, emotions, and other affectual states from text," Emotion measurement, pp. 201-237, 2016.

[21] N. Sagar "A comparative study of classification techniques in data mining algorithms," Oriental Journal of Computer Science and Technology 8.1, pp. 13-19, 2015.

[22] M. Kumar and A. Bala, "Analyzing Twitter sentiments through big data," Computing for Sustainable Global Development (INDIACom), 3rd International Conference on. IEEE, 2016.

[23] M. Mittal,, et al. "Monitoring the Impact of Economic Crisis on Crime in India Using Machine Learning," Computational Economics, pp. 1- 19, 2018.