International Journal of Trend in Research and Development, Volume 8(5), ISSN: 2394-9333
www.ijtrd.com

Twitter Sentiment Analysis Using Random Forest Algorithm for Live Twitter Data

[1]Lohith D S and [2]Nitin Raj,
[1,2]Department of Electronics and Communication Engineering, The National Institute of Engineering, Mysore, India

***Abstract*—**With the evolution of the internet, the magnitude of data present within the web has also amplified. There is also a large volume of information present within the web for its surfers. And not to forget, there is tons of information generated every second. A Social network is a platform for exchanging ideas and opinions. Twitter, Facebook, Instagram and Google are some of the social network giants of the present times. They are rapidly gaining popularity as they allow people to share and express their views about topics, have discussions with different communities, or post messages across the planet. The data generated by these sites gives ample opportunities for research on sentiment analysis. The Project focuses mainly on sentiment analysis of Twitter data. The tweets consist of sentences that are highly unstructured, heterogeneous, and are either positive or negative or neutral. During this project, we offer a comparative analysis of existing techniques for opinion mining like machine learning and lexicon-based approaches. With the help of various machine learning algorithms like Naive Bayes, Random Forest Classifier we provide research on live Twitter data. Based on the experiments conducted on ML algorithms, we show that performance of Random Forest ML model provides better accuracy of 80% compared to other ML algorithms

Keywords*—Equal Pay For Equal Work; Anti-Employment Discrimination; Labor Contracts; Worker Protection;*

I. INTRODUCTION

The internet has revolutionised the way in which people expressed their views and opinions. Blog posts, social media, product review websites etc are the new ways through which views and opinion are spread across the globe. Social networking sites like Facebook, Twitter, Instagram, etc. are being used to express one's emotion opinions and share views about issues and incidents taking place around the world daily. Through the online communities, we get interactive media where consumers influence co-consumers of the platform. Social media is fabricating a colossal volume of data rich with human behaviour and sentiments within the sort tweets, blog posts, comments, etc. The data extracted from the users could be used by the tech companies to advertise the products based on the former's search history. The amount of data generated by the users is quite broad for a manual research. So requirement arises to automate this. Various sentiment analysis techniques are widely used. A company's brand value depend upon the extent to which it satisfies the customers with its product and services. Sentiment analysis can be employed to work on the reviews given by the customers. And hence firms and businesses use this data to know about their products or services in such a way that they understand the user's demands[2]. Textual Information retrieval techniques mainly specialize in processing, searching, or analyzing the factual data present. Facts have an objective component but there are other textual contents that express subjective characteristics. These contents are mainly opinions, sentiments, appraisals, attitudes, and emotions, which form the core of Sentiment Analysis (SA). It offers many challenging opportunities to develop new applications, mainly thanks to the large growth of obtainable information on online sources like blogs and social networks. For instance, recommendations of things proposed by a recommendation system are often predicted by depending on user-generated content online to an excellent extent for deciding. For e.g. if someone wants to shop for a product or wants to use any service, then they firstly search its reviews online, discuss it on social media before making a choice, taking under consideration considerations like positive or negative opinions about those items by making use of Sentiment Analysis.

Big Data tech giants like Twitter and Facebook are now into the everyday life of people across the world. Messages and opinions are spread across the globe within seconds. The superpower of these websites has not always been sweet. Hate messages and fake messages on these web pages have paved the way for violent riots in many countries. Curbing these kinds of messages poses a big challenge to all social media sites. Our project is directed in this path to detect the sentiment of messages on these sites and differentiate them if they are spreading joy or hate. But looking at the consumers these sites have garnered over the years of their inception and views expressed every minute from all over the world makes it difficult to handle the amount of data generated. But to overcome this challenge would mean the success of the project.

II. LITERATURE REVIEW

A compilation of tweets from twitter either a static dataset or a live dataset using Twitter API is used as the primary corpus for sentiment analysis. This analysis hints at the use of opinion mining or NLP. Sentiment analysis can be performed by extracting polarity and subjectivity from semantic orientation which refers to the strength of words and polarity text or phrases[3]. Extracting polarity from sentences or tweets can be a complicated task as the sentences contain punctuation, emojis and other things that might result as a hindrance in calculating the polarity. These non essential terms in the sentences are called Stopwords. Stopwords can be responsible for deterring the accuracy, hence these stopwords need to be minimised[4]. Not only the stopwords, sentences meant for sarcasm also contribute in deterring the accuracy of the tweets. Algorithms do not understand the real meaning of the sarcasm and end up tagging the tweet to the wrong sentiment. Neither there is yet any means to identify sarcastic tweets among the rest of the tweets. Hence this issue remains persistent[6]. Classifiers had to play a major role in making the model to reach higher accuracy. Multinomial NB, SVM , Naive Bayes, and Random Forest Classifier are some known classifiers in the field of Natural Language Processing NLP[5]. Choosing the best among the known classifiers could be a major turning point in the latter part of the model. SVM and Multinomial NB had the ability to build a great model. The Naive Bayes algorithm also showed greater accuracy. But the accuracy varied greatly on using n-gram where the n could be varied from 1 up to 3 or 4.

Hence Naive Bayes wasn't reliable when the number of tweets or the size of the dataset increased. As the dataset we were using was live tweets, which would comprise an unending flow of data into the dataset,but Naïve Bayes method has a better accuracy level compared to using other methods, such as KNN and SVM [7].Sentiment analysis is helpful to find the sentiment of author behind his/her comment. Mining of tweets is done using a string search. The mined tweets are subjected to sentiment analysis using machine learning classifiers. These classifiers classify the tweets into positive, neutral and negative[8]. Applying Random Forest Classifier to the model and training and testing the dataset, brought high accuracy[1]. Random Forest Classifier[11] used trees and created a forest. The "forest" it builds, is an ensemble of decision trees, usually trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result[21].The amount of training data could impact the performance classification. More the training data is used, the better the performance of the classification obtained. Amount of data retrieved for the initial step should be great enough, so as to improve the performance in the latter steps of the classification process [16]. Random Forest classifier could be used for both classification and Regression models. Also the accuracy can be increased up to 4-5% using the Hybrid approach[17]. The next thing after sorting out a better algorithm was to implement a mechanism to find out the overall sentiment of the tweet. Vader[10] was the one with the best result. Vader sentiment is a lexicon and rule-based sentiment analysis tool. it is categorically adjusted to sentiments expressed on social networking sites. Vader's task is not only limited to tell about positivity or the negativity of the sentence given, rather it describes how positive or negative a tweet is by assigning a score to it. Vader was easily accessible as it is a built-in library of Python. Then we started with the training and testing of data. In the beginning, we started with a csv dataset from kaggle. We took it to the next level by accessing live tweets and analyzing their sentiments in real time. Now, coming to the data visualization part. It plays an important role in the whole process, you can do everything right but in the end if you are not able to impart the stats into others through a series of graphs and charts then the observer may find it difficult to understand the stats. Matplotlib and Word Cloud[9] are used for visualization. The accuracy of the model is 84%. Confusion matrix is used as an evaluation technique. The Web Page has been deployed using Heroku[14]. A web page has been developed using HTML and CSS.

III. METHODOLOGY

The fig 1 shows the block diagram of Twitter analysis using Random forest algorithm. In this block diagram

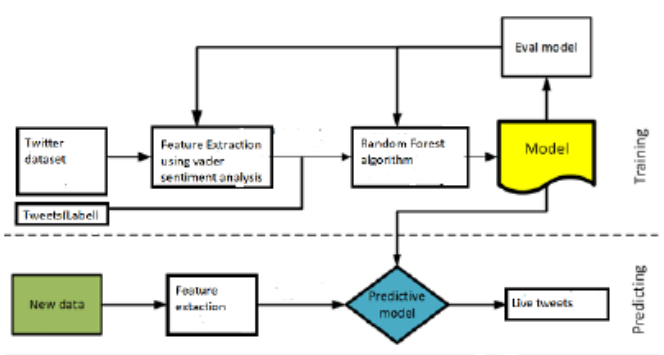


Fig 1. Proposed block diagram of Twitter sentiment analysis using RF algorithm for live data.

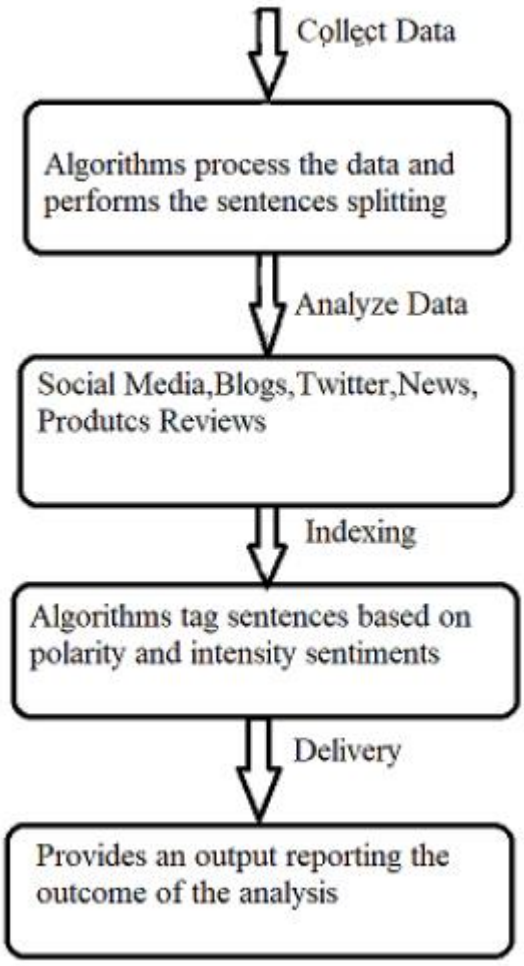


Fig 2.Flow chart of Twitter analysis using RF algorithm

1. Data Retrieval: Twitter data is mined using Twitter API. The data regarding a particular topic is being mined. We have selected Twitter API due to its flexibility in data mining.
2. Data Pre-processing: The part of data with stopwords, characters, articles, suffixes and prefixes are removed using tokenization and lemmatization.
3. Tweet correction: It's hard for a machine to understand human slang or sarcasm. We try our best to replace such words with standard words in the dictionary. But sarcasm is a part where the model often tends to make a mistake.
4. Polarity Detection: To detect what a tweet is trying to convey. Whether it stands on a positive side or negative side or on the neutral according to the society's standards. We have used the Vader sentiment library to assign a magnitude to each and every word.
5. Data Visualization: The visualization perfectly puts all things in a single bag and showcases it to the observer. It perfectly depicts all the stats. Matplotlib and Wordcloud have been used to implement this part.

IV. RESULT

Once the implementation of our model using Random forest algorithm(RF) is done , we got sentiment distribution for search on twitter about mahindra and toyota cars,which shows positive negative and neutral sentiment of peoples live post about mahindra and toyota.

Fig 3 shows 32.3% positive 11.3% negative and 56.5% neutral for mahindra. Fig4 shows 47.1% positive 36.8% negative 16.2% neutral for toyota .we have found that more positive

reviews found for mahindra related search than toyota .In fact negative tweets gives better insight for the stakeholder or for an organization to improve their service or product to their customers. This helps them to know their customers interest and requirements. comparison of both brands for sentiment distributions about the online In the word cloud of toyota (as shown in fig 5) containing the word e.g. Motor, hybrid, sale, vehicle.

In the word cloud of Mahindra (as shown in fig6), commercial, India, automotive, electric. Comparison of both brands for sentiment distributions about the online Twitter users, we can arrive at a conclusion that in both brand searches positive reviews are found to be greater than negative reviews. And more neutral reviews are found in both brand searches. neutral review is found more for Mahindra brand search.

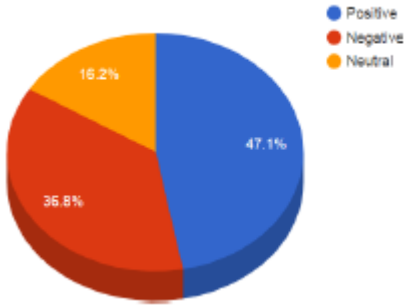


Fig 3

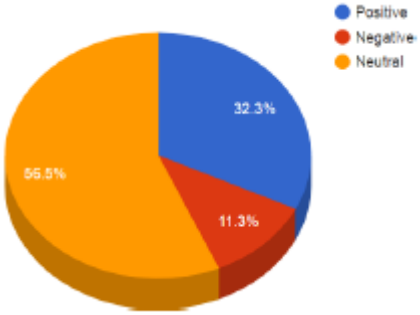


Fig4

Fig 5.word cloud for Toyota

Fig 6.word cloud for Mahindra

The model has been successfully implemented to analyse live tweets with the accuracy of 84%.The web application has also been developed using Heroku Web application deployment. Below are some glimpses of the web page searching for a current debatable topic “Cryptocurrency”.

Fig 7. Homepage of web app

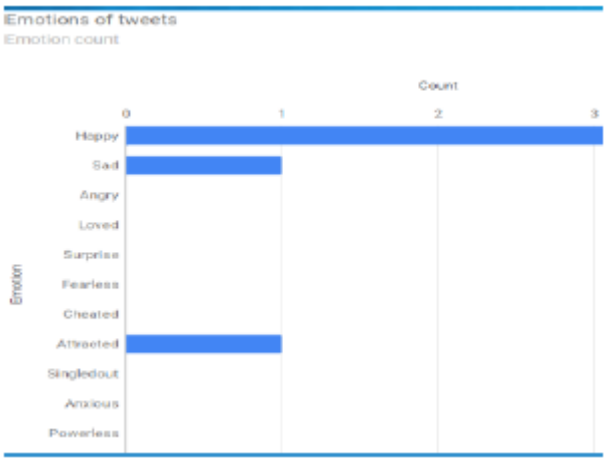


Fig 8. Depicts the count of words describing emotion(tweets)

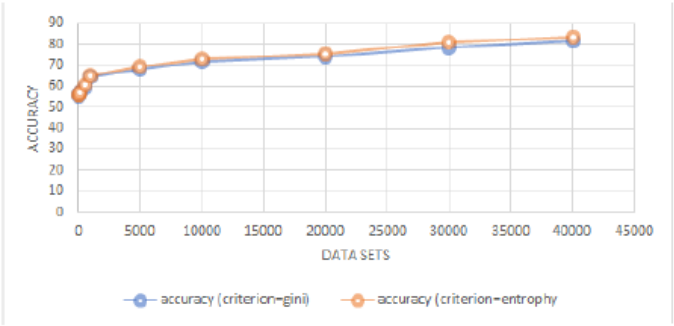


Fig 9.represents the difference accuracy obtained from RF algorithm

Table 1.accuracies of RF algorithm

	dataset	accuracy (criterion=gini)	accuracy (criterion=entrophy
2	10	55.42	55.76
3	50	56.75	55.86
4	100	57.2	57.4
5	500	59.48	60.75
6	1000	64.79	65.12
7	5000	68.46	69.34
8	10000	71.58	72.86
9	20000	74.56	75.55
10	30000	78.45	80.9
11	40000	81.76	83.76

Table 2.comparison between random forest and naive bayes algorithm.

Algorithm	Accuracy
Random Forest algorithm	84%
Naive Bayes algorithm	68%

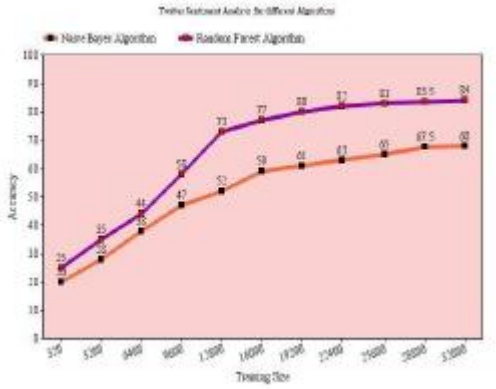
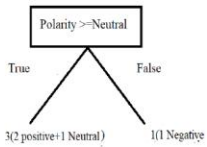


Fig10. Graphical comparison between random forest and naive bayes algorithm with respect to accuracy

Table 3.Gini index approach using Random forest

User ID	Tweets as Sentimental Analysis	Status
1	I got the vaccine.I'm pretty 'fit' i havenot experienced any covid symptoms despite working outside throughout the pandemic.	Positive
2	They Relied on chinese vaccine now they are battling outbreaks	Negative
3	covid lab-break theory:'rare' genetic sequance doesnt mean the virus was engineered	Neutral
4	At a White House briefing on COVID-19, Dr. Anthony Fauci said that 20.6% of new cases in the U.S. are due to the Delta variant.....	positive



Gini index= $1-\sum (P_i)^2$

=1-(summation of probability of neutral tweets)
=1-((3/4)^2+(1/4))
=0.375

CONCLUSION

On comparing naive bayes and Random forest algorithm we got better accuracy using random forest algorithm we have compared the gini and entropy criterion for splitting the nodes of a decision tree. On the one hand, the gini criterion is much faster because it has to do with less computational work,On the other hand, the obtained results using the entropy criterion are slightly better. We also observed accuracy obtained from both gini and entropy criterion increased as the number of data sets increased. So as we can observe the results are so similar, it does not seem to be worth the time invested in training when using the entropy criterion.

References

[1] "C. Jose and G. Gopakumar, 'An Improved Random Forest Algorithm for Classification in an Imbalanced Dataset.,' 2019 URSI Asia-Pacifi Radio Science Conference (AP-RASC), 2019, Pp. 1-4,Doi:10.23919/URSIAP-ASC.2019.8738232."

[2] "Kumar, Ankit, et al. 'Sentiment Analysis Using Machine Learning For Twitter'. International Journal of Innovative Research in Technology, Vol. 6, No. 12, May 2020, Pp. 299–304."

[3] "P. Kłosowski, 'Deep Learning for Natural Language Processing and Language Modelling,' 2018 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), 2018, Pp. 223-228, Doi:"

[4] "M.Rambocas, and J. Gama, 'Marketing Research:The Role of SentimentAnalysis'. The 5th SNA-KDD Workshop'11. Universityof Porto, 2013."

[5] "Nihal G Bailur, Prof. Merin Meleet et.al 'Sentiment Analysis on Twitter Data Using ML', International Research Engineering and Technology."

[6] "Mohit Wadera et.al 'Sentiment Analysis of Tweets-A Comparison of Classifiers on Live Stream of Twitter' IEEE 2020."

[7] "Muhammad Zubair et.al (2017) 'Twitter Sentiment Analysis Classification Scheme.

[8] M. Wongkar and A. Angdresey, "Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter," 2019 Fourth International Conference on Informatics and Computing (ICIC), 2019, Pp. 1-5, Doi: 10.1109/ICIC47613.2019.898588

[9] "R. B. Shamantha, S. M. Shetty and P. Rai, 'Sentiment Analysis Using Machine Learning Classifiers: Evaluation of Framework Using Hybrid Performance,' 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), 2019, Pp. 21-25, Doi: 10.1109/CCOMS.2019.8821650."

[10] Luvsandorz, Zolzaya. "Simple Word Cloud in Python." Towards Data Science.

[11] "Sentiment Analysis Made Easy Using VADER" Https://Analyticsindiamag.Com/Sentiment-Analysis-Made-Easy-Using-Vader."

[12] "A. P. Jain and V. D. Katkar, 'Sentiments Analysis of Twitter Data Using Data Mining,' 2015 International Conference on Information Processing (ICIP), 2015, Pp. 807-810, Doi: 10.1109/INFOP.2015.7489492."

[13] "Ali, Jehad & Khan, Rehanullah & Ahmad, Nasir & Maqsood, Imran. (2012). Random Forests and Decision Trees. International Journal of Computer Science Issues (IJCSI). 9."

[14] "Xiang Gao, Junhao Wen, Cheng Zhang, 'An Improved Random Forest Algorithm for Predicting Employee Turnover', Mathematical Problems in Engineering, Vol. 2019, Article ID 4140707, 12 Pages, 2019."

[15] "Hutto, C.J. & Gilbert, Eric. (2015). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014."

[16] Rostad, Shane. "What Is Heroku? A Simple Explanation for Non-Techies." Trifin.

[17] "R. A. Ramadhani, F. Indriani and D. T. Nugrahadi, 'Comparison of Naive Bayes Smoothing Methods for Twitter Sentiment Analysis,' 2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS), 2016, Pp. 287-292, Doi: 10.1109/ICACSIS.2016.7872720."

[18] "R. Wagh and P. Punde, 'Survey on Sentiment Analysis Using Twitter Dataset,' 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018, Pp. 208-211, Doi: 10.1109/ICECA.2018.8474783."

[19] "Vishal A Khare. 'Sentiment Analysis of Twitter Data:A Survey of Techniques.' Docplayer.Net. Https://Docplayer.Net/19140608-Sentiment-Analysis-of-Twitter-Data-a-Survey-of-Techniques.Html." .

[20] "Prithika, Dasgupta. 'Sentiment Analysis: Personality Analysis from Tweets.' Coursehero.Com. Https://Www.Coursehero.Com/File/74169395/AI-Projectdoc"

[21] "A Complete Guide to the Random Forest Algorithm." Builtin.com.N.p.,n.d. Web.23 June 2021.