Social Media Sentiment analysis using Twitter Dataset

Amit Kumar
Bachelor of Engineering in Computer Science and Engineering (Hons.) IBM - BIG DATA AND ANALYTICS-CS206
Chandigarh University
Mohali, India 140413
amitkumarprasad55@gmail.com

Sachin Pareek
Bachelor of Engineering in Computer Science and Engineering (Hons.) IBM - BIG DATA AND ANALYTICS-CS206
Chandigarh University
Mohali, India 140413
pareeksachin71@gmail.com

Tushar Sharma
Bachelor of Engineering in Computer Science and Engineering (Hons.) IBM - BIG DATA AND ANALYTICS-CS206
Chandigarh University
Mohali, India 140413
tusharbhardwaj2311@gmail.com

Neeru Bala (E13122)
Asst. Professor (AIT-CSE)
Chandigarh University
Mohali, India 140413
neeru.apcse@gmail.com

***Abstract*— In our day-to-day lives, we use a lot of social media sites to communicate and share our views. A huge amount of data can be generated by the users. To extract the emotions, we are using the social media sentiment Analysis using the Twitter dataset. Our main goal is to extract the opinions of millions of users and get full insight into the thoughts behind the data. The analysis is basically performed on the text. We are using one of the biggest social media sites, Twitter, to analyses the tweets and identify the opinions of the users. Sentiment analysis, which involves analyzing text-based data, presents distinct challenges when compared to traditional text analysis. The sentiment outcome is categorized into optimistic, pessimistic, and impartial sentiments.**

Keywords - sentiment analysis, social media, twitter data, NLP, logistic regression, TF-IDF

I. Introduction

In the age of the internet, social media Websites are on trend, like Twitter, Facebook, and LinkedIn. Social media sites are one of the best places to express their emotions and creative thoughts. Our main goal of analysis of the data in the proper manner using 3D models. Lots of unstructured text data is available on social media. We extract this data and find the emotions of the user. Social media sites like Twitter, where people can share their thoughts and issues. People can share their opinions in different situations; some of the time these opinions will be positive, some will be negative, and some will be neutral. The way of sharing opinions is not always the same. Sometimes it's very difficult to understand because of a lack of context. The active user base on Twitter in 2023 will be 330 million. That's why we are using this social media site to perform sentimental analysis. Twitter sentiment analysis poses a challenge due to the concise nature of tweets. To extract valuable features, it is necessary to preprocess them, considering factors such as non-standard vocabulary, misspellings, and emoticons. As with other types of text, various techniques for feature extraction can be utilized to gather important information from tweets.. We are using scraping to get the data from Twitter, then we pre-process the data, perform integration on the data, and visualize the data.

II. LITERATURE REVIEW

Now as per the rapid growth of the business almost everyone needs analysis of data and guidance for surveys to know well about what's going on in the present days. As SA comes into existence after 2005, as many other social media or social networking site is them forgive us data about the public thought and feelings. After time to time many popular social media comes Infront of us like Twitter, Facebook, and LinkedIn were founded in 2006,2004 and 2003, respectively.

Sentiment Analysis is research to know about the opinion of others which can be related to their emotions and attitude shown in natural language with respect to an occurrence or event. Nowadays these analyses show that sentiment analysis has reached achievement which provides not only positive and negative responses but also deals with behavior and emotion for different languages and topics. In the research of sentiment analysis, different researchers use different methods or techniques to predict social opinion and emotion through text and language.[8]

In [1 performed sentiment analysis on movie data set using Hadoop framework and analyzed many tweets. Analyzed a substantial number of tweets to determine Positive, Negative, and Neutral emotions on various aspects of Twitter data.

In [2] techniques for combining audio-visual modalities. textual sentiment analysis should be considered in contrasting different SA techniques. Lexicon-based ML, hybrid methods, and

In [3] We used this research article as our starting point. as in analyzing literature for emotional content. They are machine learning and symbolic approaches. Identifying emotional keywords from tweets with several keywords can present some challenges. It is also challenging to deal with slang and misspelt words. To solve these problems, a two-step feature extraction approach is followed by suitable pre-processing to produce an effective feature vector. Twitter-specific characteristics are retrieved in the first phase and added to the feature vector. Then, these features are deleted from tweets, and feature extraction is once again carried out as if it were being done on regular text.

In [4], machine learning methods such as Support Vector Machines (SVM), Naive Bayes (NB) and Decision Making (DT) are used for classification. The article provides statistics for distribution accuracy, classification error, kappa statistics, mean error, root mean without error, relative error Errors and wrong relative bases to review videos and tweets.

In [5] techniques of system gaining knowledge of with semantic evaluation for classifying the sentence and product opinions based totally on twitter facts. The important thing aim is to analyse a massive number of reviews through using twitter dataset which are already classified. the naïve byes approach.

In [6], this article provides an overview of recent advances in sentiment analysis and classification and briefly discusses the challenges required to assess sentiment. We also found that most of the work done was based on machine learning techniques rather than narratives.

In [7] Classifying the polarity of a specific tweet feature is part of sentiment analysis. The three categories of polarity are positive, negative, and neutral. Different lexicons, like the Sent-WordNet and Bing Lui sentiment lexicons, are used to identify polarity and assess sentiment strength and sentiment score, among other metrics.

In [8] using a recurrent neural network (RNN) to categories tweets' emotional content. By analyzing the connections between words, the model classified tweets as either positive or negative and utilized the recurrent neural network for emotional prediction. This resulted in more nuanced categorization of emotional strength, ranging from weakly positive or negative to highly positive or negative, as opposed to just categorizing texts as positive or negative.

In [9], the analysis focuses on writing tweets in English for sentiment mining, different communication companies in the Kingdom of Saudi Arabia classify them using a supervised machine learning algorithm. They also use TF-IDF (Time Frequency-Reverse Document Frequency) to measure the importance of a word for a tweet.

In [10] devises a technique for gauging sentiments via Arabic tweets and Facebook remarks available to the public. To evaluate the effectiveness of different weighting factors on precision of measurements, they utilize supervised machine learning approaches such as support vector machine (SVM) and Naive Bayes, along with binary model (BM) and TF-IDF.

TABLE I COMPARATIVE ANALYSIS

Author(year)	**Focus of Study**	**Approach/Algorithm**	**Data Set**
Apoorv Agarwal et.(2011)	Sentiment analysis on Twitter	SVM, PPS (prior polarity score)	Twitter
Mohammad Rezwanul.et. (2017 IJACSA)	Sentiment Analysis	SVM, KNN, Grid Search	Twitter
Shreya Ahuja et. (2017)	Clustering and Sentiment Analysis	K-Mean Clustering, Fuzzy c Mean, Partition clustering	Twitter
Manju V. et.(©2015 IEEE)	Opinion Mining and sentiment analysis	Naïve Bayes, SVM and J48 Classifiers	Twitter
Amrita Shelar et. (©2018 IEEE)	Discovered the sentiment of people in form of polarity	NLTK (Natural Language Processing Toolkit)	Twitter
Sahar A. et.(©2019 IEEE)	Using sentiment analysis to categorize certain English tweets	Maximum entropy, Naive Bayes, Support Vector Machines (SVM), Decision Trees, Random Forests, and Bagging	Twitter API
Usman Naseem (2021 IEEE)	Covid Sentiment Analysis through social media	ML, NLP	Twitter
Adyan M. et. (2017 InAES)	Twitter sentiment using Deep Learning Method	Deep leaning methods and Deep neural network	Twitter, Facebook,etc.
Dr. Priyanka et. (2020 IEEE)	Text Classification on Twitter Data	RNN with LSTM	Twitter

III. PROPOSED MODEL

A dataset is a collection of data that is structured and organized in a specific way for analysis purposes. Our dataset includes multiple rows and columns where in rows are the single observation points or data points and column represents the variable or attribute of that observation.

In data science project, quality of data is very important because that directly affects the accuracy and analysis of the results drawn from the data.

A. *DATASET CREATION*

In this investigation, data is scraped from twitter using the Snscrape module which eases us to scrape any amount of data without the use of API key authentication of twitter and do not have the limiting condition of 3200 tweets of Twitter API.

Our dataset is on the recent Adani Vs Hindenburg Fiasco. We will fetch tweets in a date and time frame and those will total to 20,000 tweets in our dataset.

We install necessary dependencies and scrape the data using the below query command from 2023-02-01 to 2023-03-09 date and time respectively.

Our data in this paper is scraped from Twitter with the query= "Adani OR Hindenburg OR #AdnaiVsHindenburg OR #HindenburgReport".

DATASET ATTRIBUTE DESCRIPTION:

a) DATE: Date of tweet creation

b) ID: Unique Id no. of the user

c) URL: The link to the Twitter account of user

d) USERNAME: Name of user

e) SOURCE: Source of tweet creation (Android/Desktop)

f) LOCATION: User location at time of tweet

g) TWEET: Actual tweet content

h) NUM_OF_LIKES: Likes on the tweet.

i) NUM_OF_RETWEETS: Retweets on the tweets

These fields will later help us in making an interactive dashboard for this Twitter dataset for easy visualization.

TABLE II Statistics of Dataset Used

DATASET	**POSITIVE**	**NEGATIVE**	**NEUTRAL**
SCRAPED	2536	1697	5768

B. *Pre-Processing of Tweets*

We are fetching the data from Twitter social media platform since it is predominantly consisting of text data and we can easily employ NLP and text processing technique to it.

Twitter limits us to publish only 140 characters of content at a time in a tweet. It also has a lot of slang words, hashtags, @username handle mentions which creates a problem in processing of data. Hence pre-processing of data is very crucial here.

- STEP1: REMOVING @USER TWITTER HANDLES

  We could see that @user and URL of the tweets are of no use in the sentiment analysis of our tweets. So, we simply remove them by a pattern matching function in python.

- STEP2: REMOVING PUNCTUATION, NUMBERS, SPECIAL CHARACTERS

  We remove the full stop(.), exclamation mark(!) and other punctuations from our tweet data so that it can be processed. Also remove any number and special characters from the tweet which do not contribute to the sentiment analysis.

- STEP3: REMOVE STOP WORDS

  Remove stop words from the tweet dataset as they provide no sense to the sentiment mining in our research.

  For Ex: "This issue is for general public" herein the stop words are this, is, for which are to be removed and the final processed text is issue general public.

- STEP4: TOKENIZATION

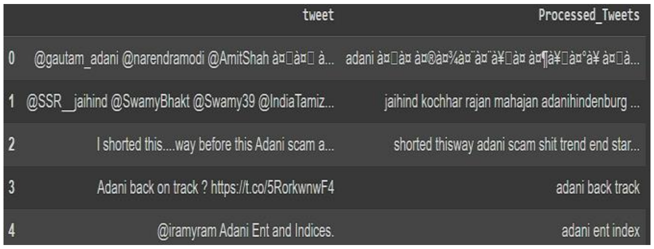
  Tokens are the individual words of the string text and we create tokens of the cleaned text and this process is Tokenization. We use the NLTK package of python.

- STEP5: STEMMING

  Stemming is the process of scrapping the suffixes of the word {"ing", "ly", "es"}.

  For Ex: All the words working, works, worked falls under the bracket of "word".

Fig. I Image of processed tweets

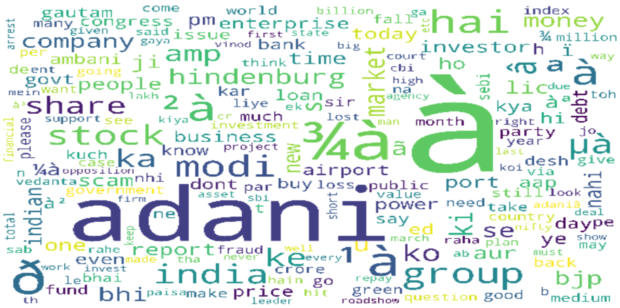


	tweet	Processed_Tweets
0	@gautam_adani @narendramodi @AmitShah à¤□à¤□ à...	adani à¤□à¤ à¤®à¤¾à¤¨à¤¨à¥□à¤ à¤¶à¥□à¤°à¥ à¤□à...
1	@SSR__jaihind @SwamyBhakt @Swamy39 @IndiaTamiz...	jaihind kochhar rajan mahajan adanihindenburg ...
2	I shorted this....way before this Adani scam a...	shorted thisway adani scam shit trend end star...
3	Adani back on track ? https://t.co/5RorkwnwF4	adani back track
4	@iramyram Adani Ent and Indices.	adani ent index

C. *Data Visualization*

Data visualization is used to graphically explain the data using charts, plots, Word-cloud for a better understanding of the data.
Word-cloud is used here which is a visualization tool wherein the most frequent word comes up in a large size and less frequent comes up in smaller size.
We can also mask any image into the frequent words of Wordcloud and customize it accordingly.

FIG. II WORLD CLOUD



D. *Extracting features from cleaned tweets*

- BAG-OF-WORDS FEATURES

Natural language processing (NLP) methods like the bag-of-words methodology frequently represent text as a collection of unorganized words or concepts without considering their order in the original text.

In this method, each word or phrase in a document or corpus is counted for frequency before a vector is created to represent the text or corpus.

Here is an example of how the bag-of-words technique works:

Let us say we have a group of three documents:
Document 1: “The quick brown fox”

Document 2: “Jumped over the lazy dog”

“The fox is quick, and the dog is lazy,” says document 3.

To use the bag-of-words method, we first compile a vocabulary of every distinct word found in the group of documents. Our vocabulary in this situation would be: [“The,”“quick,”“brown,”“fox,”“Jumped,”“over,”“lazy,”“dog,”“is,”“and”]

After that, each document is represented as a vector of word frequencies.

This vector shows that in Document 1, the term “the,”“quick,”“brown,” and “fox” all appear once each, whereas all other words appear zero times.

This vector shows that in Document 2, the phrases “Jumped” appears once, “over” appears once, “lazy” appears once, and “dog” appears once, while all other words appear zero times [TABLE III].

By representing each document as a word frequency vector, we can use these vectors for various NLP tasks such as text classification, pattern recognition, and data retrieval.

TABLE III VECTORIZATION OF VOCABULARY

	The	Quick	Brown	Fox	Jumped	Over	Lazy	Dog	Is	And
D1	1	1	1	1	0	0	0	0	0	0
D2	0	0	0	0	1	1	1	1	0	0
D3	2	1	0	1	0	0	1	1	2	1

- TF-IDF FEATURES

Term Frequency-Inverse Document Frequency is referred to by the abbreviation TF-IDF. It is a statistical method for determining the importance of a phrase inside a text or corpus (a collection of texts). The approach is based on the idea that a phrase that frequently appears in one document is crucial, but the term may not be crucial if it frequently appears in other papers. Here’s an example to illustrate the concept of TF-IDF:

Suppose we have a corpus of three documents:

EXAMPLE 1: “The cat in the house.”

EXAMPLE 2: “The cat saw the rat.”

EXAMPLE 3: “The dog ate the cat’s hat.”

We determine the importance of “cat” word in these documents using TF-IDF.

Step 1: Term Frequency (TF) for each “cat” word appearance

In ex 1, “cat” appears once.

In ex 2, “cat” appears once.

In ex 3, “cat” appears twice.

Step 2: Inverse document frequency (IDF) for “cat” appearance.

No. of example containing word “cat” =3

IDF formula: IDF = log(N/n), where N is the total number of documents/examples in the corpus and n is the number of documents/examples containing the term. In this case, IDF(“cat”) = log (3/3) = 0.

Step 3: TF-IDF score for “cat” in each document.

For ex 1, the TF-IDF score for “cat” is TF (“cat”, Document 1) * IDF(“cat”) = 1 * 0 = 0.

For ex 2, the TF-IDF score for “cat” is TF (“cat”, Document 2) * IDF(“cat”) = 1 * 0 = 0.

For ex 3, the TF-IDF score for “cat” is TF (“cat”, Document 3) * IDF(“cat”) = 2 * 0 = 0.

The results clearly show that “cat” is not a significant word in the corpus because it occurs quite frequently in all three documents.

Also,it’s IDF value is 0, which means it is not unique to any particular document.

In conclusion, TF-IDF is an effective method for locating the key phrases in a document or corpus. It can assist increase the accuracy of text-based applications like search engines and recommendation systems by assigning more weight to phrases that are specific to a document and less weight to terms that are widespread throughout the corpus.

E. MACHINE LEARNING MODEL

Machine learning models are created by training algorithms using labeled or unlabeled data. Therefore, machine learning algorithms can be trained and produced in three ways: a) Supervised learning. b) Unsupervised learning. c) semi-supervised learning method. We used supervised learning to treat the algorithm.

Supervised Machine Learning is the category in which we are going to solve the underlying problem. As we take all labeled data for analysis of the tweets.

With supervised learning, you have input variables (x) and output variables (Y) and you use an algorithm to learn the mapping function.

$$Y=f(X)$$

Ideally, you want your mapping function to be approximated sufficiently well so that you can correctly predict the output variables (Y) when you have new input data (x).

To predict results on the test data, we generally use different models to see which one fits the dataset best.

i. LOGISTIC REGRESSION:

The only first model we are going to use in this analysis is Logistic Regression. This method is for a statical approach where we learn for binary classification problems. The goal is to predict whether the input belongs to one of two classes. The result is usually defined as 0 or 1.

The sigmoid function, commonly known as the logistic function, has the following form:

$$f(x) = 1 / (1 + e^{(-x)})$$

where f(x) is the anticipated probability of the event occurring and x is the linear combination of predictor variables.

By determining the values of the coefficients that maximize the likelihood of witnessing the data, maximum likelihood estimation is used to estimate the coefficients in the linear equation.

The logistic regression likelihood function is expressed as

$$L = \Sigma (f(xi)yi * (1-f(xi)) (1-yi)).$$

Where prod_i stands for the sum of all observations, yi is the binary response variable (0 or 1) for the ith observation, and xi is the linear combination of predictor variables for the ith observation.

ii. DECISION TREES:

The second model we used is decision trees.

Decision tree might have a node that tests whether the text contains the word “good”, and if so, branches to a positive sentiment node. If the text contains the word “bad,” another node might branch to a negative sentiment node.

As a splitting criterion, the property with the biggest information gain or the lowest impurity measure is chosen. Entropy, Gini index, and classification error are the three impurity measurements that are most frequently used.

Entropy is an indicator of how disorderly or uncertain a collection of samples is, and it is defined as:

$$\Sigma (p_i * \log2(p_i)) = - H(S)$$

S stands for the sample set, pi represents the percentage of samples that correspond to class i, and log2 stands for the binary logarithm.

The difference between the set’s entropy before and after the attribute A split is used to define the information gain of an attribute A with regard to a set of samples S:

$$H(S) = \Sigma (|S_v|/|S| * H(S_v)) – IG(A, S)$$

where |S| represents the number of samples in S, S_v represents the subset of samples where attribute A = v, and v is the value of attribute A.

The Gini index is an indicator of sample impurity and is defined as:

$$\text{Sum}_i (p_{i}2) – 1 \text{ equals Gini}(S).$$

where pi represents the percentage of samples that correspond to class i.

iii. NAÏVE BAYES CLASSIFIER:

Naive Bayes can be learned quickly and is computationally effective on big datasets. Due to this, many applications, such as spam filtering and sentiment analysis, favor it.

It states that, normalized by the probability of E, the probability of a hypothesis H given some evidence of E is proportional to the sum of the prior probability of H and the likelihood of E given H:

$$P(H \mid E) = \frac{P(E \mid H) * P(H)}{P(E)}$$

By assuming that the features are conditionally independent given the class label, Naive Bayes may separately calculate the probability of each feature given the class label:

$$P(X\_1, X\_2, \ldots, X\_n \mid C) = P(X\_1 \mid C) * P(X\_2 \mid C) * \ldots * P(X\_n \mid C)$$

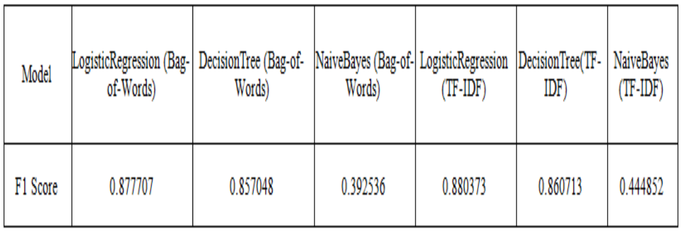
IV. EVALUATION AND RESULTS

In this paper we have used a dataset which consists of 10,000 tweets generated on our specific query. We later go through the data science lifecycle of data cleaning, data pre-processing, EDA, feature engineering techniques used like Bag-of-Words and TF-IDF, then building three supervised machine learning models – Logistic Regression, Naïve Bayes and Decision tree for training and testing our data.

We first trained our models in bag of words technique and later used TF-IDF technique.

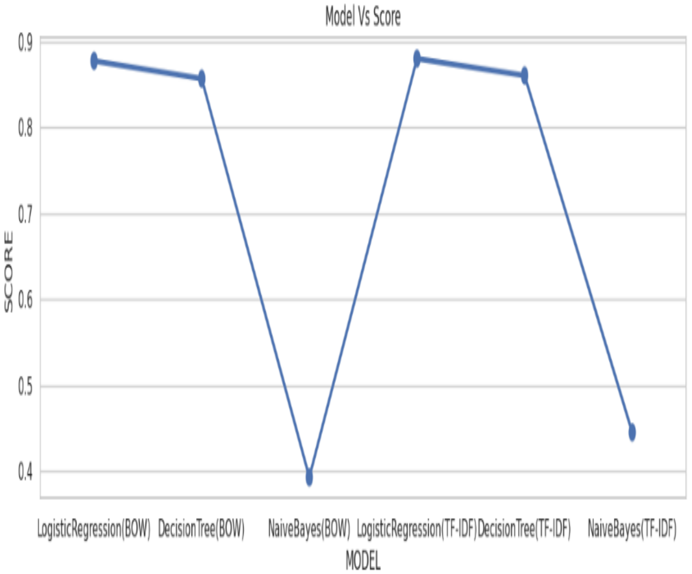
For logistic regression we obtained the F1 score as0.877 in Bag of words technique and F1 score as 0.8803 in TF-IDF technique. Similarly, f1 for naïve bayes were 0.392 and 0.444, f1 score for decision tree were 0.857 and 0.860 respectively [TABLE IV].

TABLE IV COMPARATIVE ANALYSIS OF F1 SCORES



Model	LogisticRegression (Bag-of-Words)	DecisionTree (Bag-of-Words)	NaiveBayes (Bag-of-Words)	LogisticRegression (TF-IDF)	DecisionTree(TF-IDF)	NaiveBayes (TF-IDF)
F1 Score	0.877707	0.857048	0.392536	0.880373	0.860713	0.444852

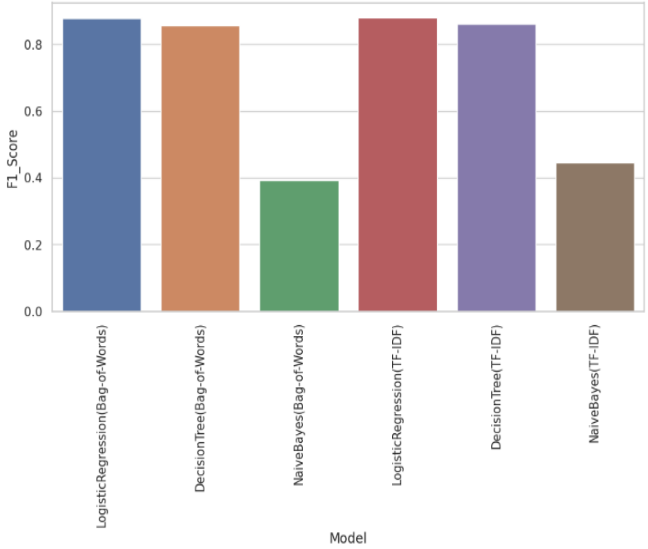
Fig. III MODEL vs SCORE



Herein we used F1 score as the resultant metric. Depending on the problem you are trying to solve, you could often either assign a larger premium to optimizing precision or recall. However, there is a more straightforward statistic that generally accounts for both recall and precision, so you can attempt to increase this number to improve your model. The harmonic mean of Precision and Recall is known as the F1-score, which is the measure in question.

$$\text{F1 Score} = \frac{2 * \text{Precision} * \text{Recall}}{(\text{Precision} + \text{Recall})}$$

Fig. IV F1 SCORE WITH TF-IDF



From our experiments with various models and feature engineering techniques we concluded that Logistic Regression gives the best result or best F1 Score with TF-IDF technique followed by Bag-of-Words technique followed by decision tree model and the worst result was obtained through Naïve Bayes model.

V. CONCLUSIONAND FUTURE SCOPE

We learned from this study about the various groups to which our sentiment ratings belong, both polarity- and subjectivity-wise. We can aggregate tweets that are positive, negative, and neutral by creating a cluster of the results from both tools' scores since pre-defined dictionaries and sentiment technologies can't accurately score every word in the context of a phrase. Some tweets only contain news or other people's opinions on a particular expression, but once we identify the subjectivity (opinion, feeling, or emotion of a specific person expressing it) of a tweet, we can group it using sentiment scores to determine the ratio of the true positive, negative and neutral "opinions" and not just sentiments of a sentence. Future development and growth in the field of sentiment analysis are highly anticipated. More complex sentiment analysis models are required to analyse sentiment accurately in various languages, situations, and data sources due to the growing demand for sentiment analysis across numerous industries.

REFERENCES

[1] Parveen, H., & Pandey, S. (2016),"Sentiment analysis on Twitter Data-set using Naive Bayes algorithm", 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (ICATccT). https://doi.org/10.1109/icatcct.2016.7912034 J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Jindal, K., & Aron, R. (2021, February),"WITHDRAWN: A systematic study of sentiment analysis for social media data", Materials Today: Proceedings. https://doi.org/10.1016/j.matpr.2021.01.048 K. Elissa.

[3] Neethu, M. S., &Rajasree, R. (2013, July),"Sentiment analysis in twitter using machine learning techniques" , 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT). https://doi.org/10.1109/icccnt.2013.6726818

[4] Das, D. (2021, May 31),"Social Media Sentiment Analysis using Machine Learning : Part — I" Medium. https://towardsdatascience.com/social-media-sentiment-analysis-49b395771197

[5] Das, D. (2021, May 31),"Social Media Sentiment Analysis using Machine Learning : Part — II", Medium. https://towardsdatascience.com/social-media-sentiment-analysis-part-ii-bcacca5aaa39\

[6] Patil, H. P., &Atique, M. (2015, December),"Sentiment Analysis for Social Media: A Survey", 2015 2nd International Conference on Information Science and Security (ICISS). https://doi.org/10.1109/icissec.2015.7371033

[7] Wagh, R., &Punde, P. (2018, March),"Survey on Sentiment Analysis using Twitter Dataset", 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA). https://doi.org/10.1109/iceca.2018.8474783

[8] Nemes, L., & Kiss, A. (2020, July 14),"Social media sentiment analysis based on COVID-19",Journal of Information and Telecommunication, 5(1), 1–15. https://doi.org/10.1080/24751839.2020.1790793

[9] Qamar, A., Alsuhibany, S. and Ahmed, S. (2017),"Sentiment Classification of Twitter Data Belonging to Saudi Arabian Telecommunication Companies", (IJACSA) International Journal of Advanced Computer Science and Applications, [online] 8. Available https://thesai.org/Downloads/Volume8No1/Paper_50-Sentiment_Classification_of_Twitter_Data_Belonging.pdf [Accessed 1 Feb. 2018].

[10] R. M. Duwairi and I.Qarqaz, "A framework for Arabic sentiment analysis using supervised classification" , Int. J. Data Mining, Modelling and Management, Vol. 8, No. 4, pp.369-381 , 2016