SENTIMENT CLASSIFICATION ANALYSIS OF TWEETS ON TWITTER DATA USING MACHINE LEARNING ALGORITHM

[1]Ch Vinod Varma, [2]P Jahnavi, [3]K Vijaya Naga Valli, [4]K.Monica Sowmya, [5]S.Suryanarayanaraju, [6]M.Lahari

[1,2,3,5]Assistant professor, Dept. of CSE, SRKR Engineering College, Chinnamiram, Bhimavaram, Andhra Pradesh, India.

[4,6]Assistant professor, Dept. of CSE, Gokaraju RangaRaju Institution of Engineering and Technology, Kukatpally, Hyderabad, Telangana, India.

ABSTRACT: Technology advancement has network social media day-to-day expansion on the internet. Reviews, forums conversations, blogging, tweeting, remarks, and comments on social networking sites are all examples of people have used the internet to communicate their ideas and opinions. Twitter, one of the most popular micro-blogging platforms are where people communicate their thoughts as tweets, making it one of the best sources for sentimental analysis. Twitter's sentiment categorization procedure includes polarity analysis of the tweets emotions. This approach presents, Sentiment Classification Analysis of Tweets on Twitter Data using Machine Learning algorithm. The opinions expressed in tweets on Twitter are analyzed using feature selection for each score word. The characteristics of words are trained and tested using a Bayes Classifier (NBC), which is also used to forecast the sentiment orientation of each tweet. According to experimental findings, described methodology provides excellent categorization outcomes in terms of F1-Score, precision, recall, and accuracy.

KEYWORDS: sentimental analysis, Twitter, sentiment polarity, Machine Learning, social media, Naive Bayes Classifier (NBC).

I. INTRODUCTION

Many people have become attracted towards social media platforms such as Facebook, Twitter, and Instagram in recent years. Many people utilize social media platforms to share their feelings, convictions, and opinions on various subjects, locations, or people. Twitter serves as a data storage facility [1]. Therefore, this data is very helpful for forecasting the outcomes of political actions, new government initiatives, or research, as well as for choosing The raw data that we retrieved from tweets is the input to described model. To do anything, they automatically extract tweets and divide them into two categories: good and negative. The user-generated information on Twitter focuses on a number of goods, occasions, persons, and political issues [2].

Sentimental analysis is the method used to extract reliable information from texts. To put it another way, it is the process of creating structured data out of unstructured data [3]. This is used to needs of the customers thoughts, feedback, and product reviews. Unstructured data includes information from the internet, such as chats, emails, pdfs, word documents, e-commerce websites, and social networking sites, additionally to the tables and figures from the company. Analyzing structured data allows for simple operation and simple result gathering. However, it can be challenging to make inferences from unstructured data sources like Twitter, email, and other social media because of a number of issues such the virtual noise effect and ambiguous information.

Document, sentence, and aspect levels of sentiment classification are all used. The original methodology is categorized sometimes as good, negative, or unbiased at the classification stage [4]. Neutral essentially indicates that there is no opinion.

Each statement is read by the statement scale, which decides whether it is presented in a positive, negative, or neutral manner. Analyses at the entity or aspect level are more extensive. It directs its attention to the opinion itself rather than the language structure [5]. Sentiment research has several uses in a wide range of industries, including politics, business, consumer advocacy, government, and brand reputation administration.

The three main types of sentiment analysis techniques are Machine Learning, Lexicon-based, and hybrid. Another classification is offered [6] with classifications for quantitative, knowledge-based, and hybrid techniques. It is possible to do difficult research in a number of fields by computationally evaluating people's attitudes and beliefs. In order to anticipate polls, for academic reasons, or for the corporate, communications, and advertising industries, information retrieval from social network data is a more common procedure.
Recognizing emotions and identifying polarities are the fundamental functions of sentiment analysis. In applications like sentiment mining for examining customer reviews and opinions about items as well as in popular subjects like political analysis, entertainment, and other hot issues, sentimental analysis has been widely used. Data analysis performed by Machine Learning algorithms contributes in the creation of analytical models. Many such technologies analyze data in real-time and forecast hidden insights by interpreting information. There are several industries that apply Machine Learning techniques, including data analytics, IoT, cyber security, and more.

The following facts are the basis for sentiment analysis using Naive Bayes classification model. When working with huge volumes of data, it is quite simple to design and highly helpful. It performs better than the most sophisticated categorization techniques despite the easiest analytical procedure. It offers a system for categorizing and computing the probability of the next incident. This approach helps the user gain information about the data collected on Twitter users perceptions of the tweets.

The following sections consider making up the remaining portion of the paragraph. The literature review is covered in Section II, and the approach for sentiment classification analysis using a Machine Learning model is explained in Section III. Section IV explains the result analysis, and Section V concludes up the approach.

II. LITERATURE SURVEY

Le, H. Boynton, G. Mejova, Y. Shafiq, Z. Srinivasan, P. et. al. [7] with the intention of changing conventional research techniques into Twitter as a fresh source of news during election campaign and monitoring the evolution of public awareness over time, 2016 reports from the US on political, party, and psychological difficulties. For the presidential election, a significant amount of Twitter data from 6 presidential campaign was gathered, it was unable to establish the veracity of the accounts.

Sonosy, O. A., Rady, S., Badr, N. L.,Hashem, M. et. al. [8] It is necessary to collect a significant quantity of data from several sources in order to understand business trends develop. Location-based social networks can provide a large amount of data that can be analyzed to understand business behaviors. In order to uncover correlations between the properties of a data collection and the prediction of corporate activities, the authors carefully examined spatial regression models and completed expert analysis.

Sunny Kumar, Paramjeet Singh, Shaveta Rani et. al. [9] utilized lexicon-based methodology for social media sentiment analysis. In the lexicon-based method, the objective and subjective words are presented in the dictionary order. Each tweet receives a good and negative score, and the total is the emotional score. The combination of positive to negative words in a tweet determines the score. The size of the language affects emotions are classified. This will be more incorrect as the size of the dictionary increases.

Minara Panto, Nivya Johny, Muhssina K M, Vinay James, Mejo Antony, Aswathy Wilson et. al. [10] acquired customer reviews of the desired product and used sentiment analysis to rate the product. To get around the problem of acquiring feedback personally from people, the major focus is gathering data through Twitter. Dual prediction is utilized to increase accuracy and the unigram technique is used to find probabilities.

Xiao Sun, Fei Gao, Chengcheng Li, Fuji Ren et. al. [11] the microblog conversations surrounding a post can be extracted using a new convolutional autoencoder that is presented. The experimental results demonstrate that, with the right structure and parameters, the described Deep Learning method is more accurate on sentiment classification than advanced surface classification techniques like SVM or NB, demonstrating the suitability of DBN (Deep Belief Network). Classification of brief documents using a method for extending feature dimensionality.

Anurag P. Jain, Vijay D. Katkar et. al. [12] proposes a method for utilizing data mining classifiers to analyze user sentiment. Additionally, it contrasts how well a single classifier performs in a group of classifiers when analyzing sentiment. The results of the experiments show that the K-Nearest Neighbor classifier is extremely high predicted accuracy. The outcome also shows that a single classifier outperforms a classifier ensemble technique.

Anjaria, M. Guddeti, R.M.R et. al. [13] utilized a hybrid strategy that included Naive Bayes, maximum entropy, feed-forward Support Vector Machines (SVM), Artificial Neural Networks, and supervised classifiers like SVM. They proposed a strategy of user influence factor extraction to forecast election results. Identifying the influencing elements and using SVM to analyze the data is produced results with an accuracy of roughly 88% for both the 2013 Karnataka Assembly Elections and the 2012 US Presidential Elections.

Shulong Tan, Huan Sun, Ziyu Guan, Yang Li, Xifeng Yan et. al. [14] Latent Dirichlet Allocation (LDA) contextual factors, cause possibilities, and Reason Candidate and Background LDA (RCB-LDA) are some of the LDA-based models that have been suggested to explain sentiment changes on Twitter. These models attempt to identify the causes of people's changing feelings for the target. The benefit is that the noisy data is properly removed, and the foreground topics are processed out successfully. Using the RCB-LDA model, it determines the precise causes of sentiment variations in the Twitter data, which is highly helpful for making decisions. SentiStrength and Twitter Sentiment, which are less efficient than other sentiment analysis approaches are two methodologies used to determine sentiment.

H. Tang, S. Tan, and X. Cheng et. al. [15] discussed four problems with supposition mining, including classifying subjectivity, classifying words with emotion, classifying

sentiment in archives, and classifying feelings. The Naïve Bayes classifier, multiple Naïve Bayes classifier, and cut-based classifier were a few of the techniques that were then described.

III. SENTIMENT CLASSIFICATION ANALYSIS USING MACHINE LEARNING

Figure 1 demonstrates the block diagram of Sentiment Classification Analysis of Tweets on Twitter Data Using Machine Learning Algorithm.

They examine the tweets that users had uploaded with hashtags to indicate their perspectives. Subscribe to the Twitter API, authenticating it with the help of the access_ token, access_secret, consumer key, and consumer secret, and then start collecting data from Twitter. Use the R packages Twitter and Rouath to accomplish. Once the keys have been created, gathering information on the desired person, product, etc. is simple.

Multiple misspellings, symbols, Uniform Resource Locator (URL), and hashtags can be found in the twitter data that has been collected. These statistics produce substandard outcomes. The preprocessing phase is required to obtain correct findings in order to prevent this. All link tags, punctuation marks, hash tags, special characters, emoticons etc. are removed during preprocessing. The tweets are grammatically corrected if necessary. Stop words that don't modify the tweet's are eliminated.

Removal: For the purposes of processing, user names and URLs included in the data are irrelevant. Therefore, most identities and URLs are either deleted or changed to standard tags.

Stemming: This refers to the process of reducing multiple types of words with the same or comparable meanings by replacing them with their roots. To make feature extraction easier, the included procedures must strive to make the data machine readable.

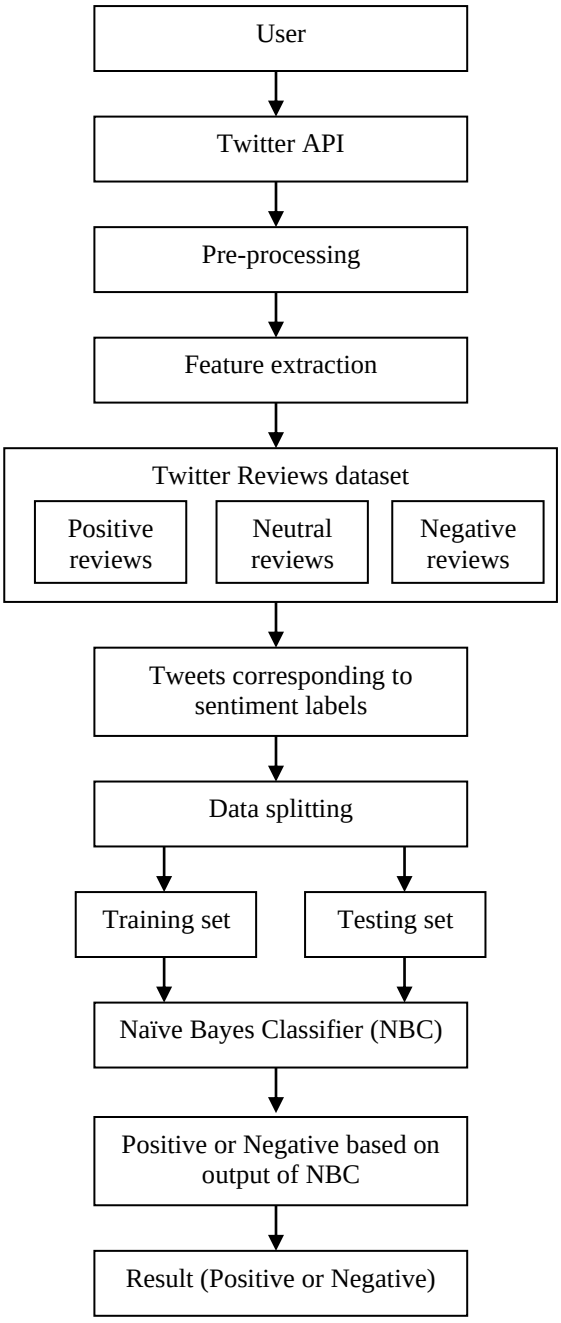

Fig. 1: BLOCK DIAGRAM OF SENTIMENT ANALYSIS USING MACHINE LEARNING

They need to extract relevant features for sentiment analysis once the tweets have been cleaned. The quantity and quality of features have an impact on the output that a model produces.

The dataset's aspect (adjective) is extracted using this procedure. This component is subsequently used to demonstrate positive and negative polarity in a statement, supporting in the classification of public opinion. Now that the data has been cleaned, the features are extracted to create significance data, which in turn creates a trained data set that includes both positive and negative data. Following feature extraction, this TextBlob approach is used to analyze the dataset to determine the sentiment ratings. TextBlob outperforms the original dataset according to noise removal.

A performance evaluation of the original and modified data using the TextBlob sentiment scores. A Python package called TextBlob is used for Natural Language Processing (NLP) activities like component tagging, sentiment classification, extracting noun phrases, translations, classifications, and more. To extract the emotions from Twitter tweets, we use TextBlob. Every tweet receives two properties from the TextBlob sentiment function: a polarization score between [-1, 1] and an objectivity score between [0, 1]. Positives, neutrality, and negatives polarization scores, correspondingly, represent positive, unbiased, and negative expressions.

After determining sentiment scores, the dataset is separated into a training set and a testing set using an 80:20 ratio, meaning that 80% of the original data is used for training and 20% for testing.

One Machine Learning algorithm that makes advantage of the Bayes algorithm and the strong independence among the features is the Naive Bayes classifier. A significant amount of information is typically utilized with a Naive Bayes algorithm because it is so simple to construct. It is utilized as one of the fundamental methods for classifying texts. Finding the positive, unbiased, and negative sentiments of messages taken from Twitter is done using sentiment analysis and the Naive Bayes classifier.

To deal with immediate inputs like tweets, the algorithm uses a Gaussian Naive Basis. It produces the class standardized distribution for a specified input. The provided test is then analyzed for classes, and an estimated score is calculated. The overall accuracy is the result of this method.

IV. RESULT ANALYSIS

People use Twitter more frequently than any other site to share their opinions and feelings through tweets. Subscribe to the Twitter API (Application Programming Interface), authenticating it with the help of the access_token, access_secret, consumer_key, and consumer_secret, and then start collecting data from Twitter. 2000 tweets were used to train the algorithms at first. This model is used with given pre-processed data after training. A total of 260 tweets were deleted out of 3000, of which 2680 are actual negatives and 160 are true positives. Training data is requires the 80% of original data and testing data requires the 20% of original data.

They have utilized four performance standards to assess the effectiveness of the Machine Learning model of the Naive Bayes classifier: accuracy, precision, recall, and F1-score.

Accuracy:

The overall correctness of the categorization produced is connected to the accuracy metric. The ratio of cases that were successfully categorized to all instances is known as accuracy.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \ldots . (1)$$

Precision:
The precision measure is the percentage of tweets successfully classified for the specified sentiment terms out of all tweets correctly categorized for this sentiment.

$$Precision = \frac{TP}{TP + FP} \ldots . (2)$$

Recall (Sensitivity):
The proportion of accurately anticipated positive observations to all of the actual class observations is known as recall. The proportion of accurate predictions to all true positives and false negatives is used to calculate recall. Precision is also known as Positive Predict Rate (PPR), and recall is also known as the true positive rate or sensitivity in this context.

$$Recall = \frac{TP}{TP + FN} \ldots . . (3)$$

F1-Score:
The weighted average of Precision and Recall is known as the F1-Score. Therefore, F1-Score accounts for both false positives and false negatives. The F1score is mostly valuable than the accuracy even though it is not as simple to read, particularly when the class distribution is unequal. The most effective cases for accuracy are those in which the cost of false positives and false negatives is similar.

$$F1 - Score = 2 * \frac{(Recall * Precision)}{(Recall + Precision)} \ldots (4)$$

Where,
True Positive (TP) Positive observations and evaluations of reviews are produced through classification.

True Negative (TN) Reviews that have been seen are not positive and are rated as one and the same.

False Positive (FP) Reviews were observed, however they were evaluated positively.

False Negative (FN) Identified reviews show that they are favourable, but the classifier rates them as undesirable.

The performance analysis of Sentiment Classification Analysis of Tweets using Naive Bayes Classifier (NBC) is compared with other classification models as K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) represented in below Table 1.

Table 1: PERFORMANCE OF DIFFERENT CLASSIFIERS BASED SENTIMENT CLASSIFICATION

Parameters	KNN	SVM	NBC
Accuracy	78.2	85.3	98.1
Precision	76.3	86.6	97.6
Recall	77.8	84.9	96.4
F1-score	79.1	87.4	98.5

Fig. 2 shows a graphical representation of accuracy and precision.

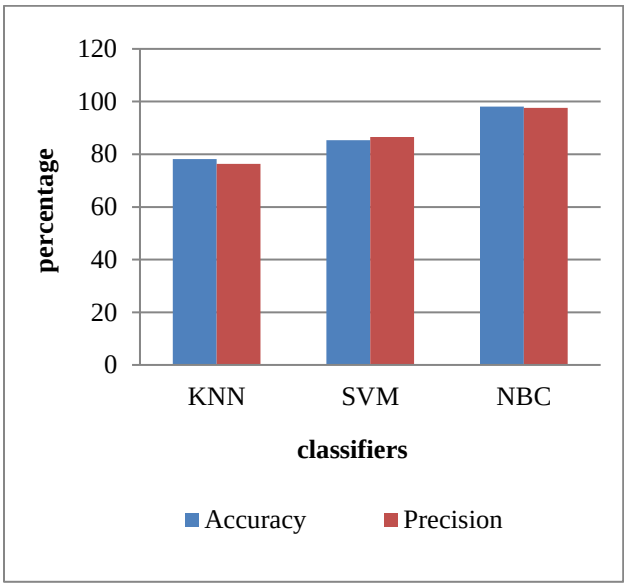

Fig. 2: COMPARATIVE ANALYSIS OF DIFFERENT CLASSIFIERS ACCURACY AND PRECISION PARAMETERS

Figure 3 shows a graphical representation of recall and the F1 score.

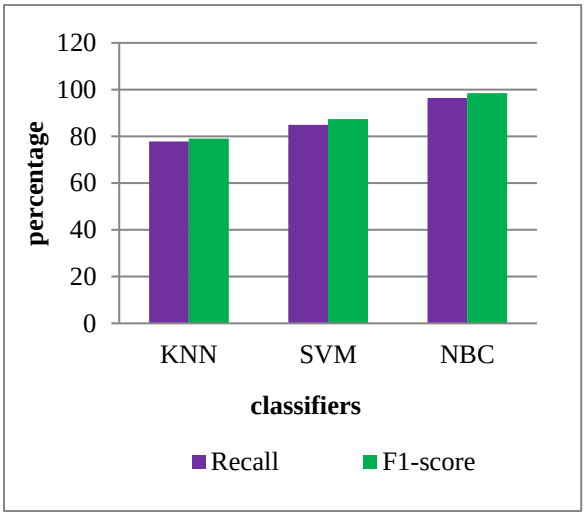

Fig. 3: COMPARATIVE ANALYSIS OF DIFFERENT CLASSIFIERS RECALL AND F1-SCORE PARAMETERS

From Table 1, it is clear that the performance of the Naive Bayes Classifier (NBC) is better than the remaining two classifications. Obtained performance parameters values as Accuracy 98.1%, Precision 97.6%, Recall 96.4% and F1-Score 98.5%.

V. CONCLUSION

In this approach, Sentiment Classification Analysis of Tweets on Twitter Data using Machine Learning algorithm is described. In order to decide which features are best. For training and testing word features as well as determining the sentiment polarity of each tweet, Naive Bayes Classifier (NBC) is utilized. Sentiment analysis is a new and quickly growing area of the decision-making process. The project's purpose is to evaluate the sentiments on a topic that are taken from Twitter and determine whether they are good, negative, or neutral. Information obtained from the Twitter API by approved users. The TextBlob approach was used on the information to find sentiment ratings after feature extraction. Comparative performance analyzes of different classifiers are described in result analysis phase in terms of accuracy, precision, recall and F1-score. Obtained performance parameters values as Accuracy 98.1%, Precision 97.6%, Recall 96.4% and F1-Score 98.5%. Future study should include the addition of extra features, which will raise prediction accuracy.

VI. REFERENCES

[1] Yuning Guo, Jianxiang Cao, Weiguo Lin, "Social network Influence Analysis", 2019 6th International Conference on Dependable Systems and Their Applications (DSA), Year: 2020

[2] Amir Karami, Morgan Lundy, Frank Webb, Yogesh K. Dwivedi, "Twitter and Research: A Systematic Literature Review Through Text Mining", IEEE Access, Volume: 8, Year: 2020

[3] Gaurav Jariwala, Harshit Agarwal, Vrai Jadhav, "Sentimental analysis of News Headlines for Stock Market", 2020 IEEE International Conference for Innovation in Technology (INOCON), Year: 2020

[4] Pankaj, Prashant Pandey, Muskan, Nitasha Soni, "Sentiment analysis on Customer Feedback Data: Amazon Product Reviews", 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Year: 2019

[5] Satuluri Vanaja, Meena Belwal, "Aspect based Sentiment analysis on E-Commerce Data", 2018 International Conference on Inventive Research in Computing Applications (ICIRCA), Year: 2018

[6] Hamdullah Karamollaoğlu, İbrahim Alper Doğru, Murat Dörterler, Anıl Utku, Oktay Yıldız, "Sentiment Analysis on Turkish Social Media Shares through Lexicon based Approach", 2018 3rd International Conference on Computer Science and Engineering (UBMK), Year: 2018

[7] Le, H. Boynton, G. Mejova, Y. Shafiq, Z. Srinivasan, P. "Bumps and bruises: Mining presidential campaign announcements on twitter", In Proceedings of the 28th ACM Conference on Hypertext and Social Media, Prague, Czech Republic, 4–7 July 2017

[8] Sonosy, O. A., Rady, S., Badr, N. L.,Hashem, M. "A study of spatial machine learning for business behavior prediction in location based social networks", 2016 11th International Conference on Computer Engineering & Systems (ICCES)

[9] Sunny Kumar, Paramjeet Singh, Shaveta Rani "Sentimental Analysis of Social Media Using R ", 2016 5th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), Sep. 7-9, 2016, AIIT, Amity University Uttar Pradesh, Noida, India.

[10] Minara Panto, Mejo Antony, Muhssina K M, Nivya Johny, Vinay James, Aswathy Wilson "PRODUCT RATING USING SENTIMENT ANALYSIS ", In International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)-2016

[11] Xiao Sun, Fei Gao, Chengcheng Li, Fuji Ren, "Chinese microblog sentiment classification based on convolution neural network with content extension method", 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), Year: 2015

[12] Anurag P. Jain, Vijay D. Katkar, "Sentiments analysis of Twitter data using data mining", 2015 International Conference on Information Processing (ICIP), Year: 2015

[13] Anjaria, M. Guddeti, R.M.R. "Influence factor based opinion mining of twitter data using supervised learning", In Proceedings of the 2014 Sixth International Conference on Communication Systems and Networks (COMSNETS), Bangalore, India, 6–10 January 2014; pp. 1–8

[14] Shulong Tan, Yang Li, Huan Sun, Ziyu Guan, Xifeng Yan, "Interpreting the Public Sentiment Variations on Twitter", IEEE Transactions on Knowledge and Data Engineering, VOL. 26, NO.5, MAY 2014.

[15] H. Tang, S. Tan, and X. Cheng, "A survey on sentiment detection of reviews" Expert Systems with Applications 36, no. 7, 2009