ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 3 | Issue 1 | February 2024

Journal Impact Factor 2023 (Quality Score): RPRI = 6.53 | SJIF = 3.647

www.ijirid.in

Twitter Sentiment Analysis Using Machine Learning Algorithm

Prof. Ms. V. P. Vaidya[1], Ritesh Korde[2], Om Chaudhari[3], Piyush Chirde[4], Sahil Deoda[5], Dipanshu Kadu[6]

[1]*Assistant Professor, Sipna College of Engineering and Technology, Amravati, (MS), India*

[2,3,4,5,6]*Undergraduate Student, Sipna College of Engineering and Technology, Amravati, (MS), India*

Abstract: *Sentiment analysis deals with the identification and classification of opinions or feelings expressed in source text. Social media generates a huge amount of sentiment-rich data in the form of tweets, status updates, blog posts, etc. Sentiment analysis of this user-generated data is very useful to find out the opinion of the crowd. Twitter sentiment analysis is difficult compared to general sentiment analysis due to the presence of slang words and typos. The knowledge base approach and machine learning approach are two strategies used for sentiment analysis of text. In this paper, we try to analyze Twitter posts about electronic products like mobile phones, laptops, etc. using the Machine Learning approach. By performing sentiment analysis in a specific domain, it is possible to identify the influence of domain information on sentiment classification. We present a new feature vector to classify tweets as positive, and negative and extract people's opinions about products.*

Keywords: *Sentimental Analysis, Twitter, Machine Learning, Naive Bayes, KNN Model, Linear Regression Model, Decision Tree Model, etc.*

I. INTRODUCTION

Twitter is a very popular microblogging platform where millions of people share their opinions in the form of tweets. Twitter users can express their thoughts in the form of tweets of up to 280 characters. People often use emoticons, slang, etc. As a result, it is common knowledge that the language of Twitter is unstructured. Sentiment analysis is used to extract relevant meaning from tweets, and the effect is expressed as the total number of positive, negative, and neutral tweets. A number of classifiers are used to obtain the result and the most accurate one is considered [1].

In this modern era of communication, everyone is connected to the Internet. Many analysts agree that social media has a big influence on election results. In the current scenario, a large section of the Indian population uses social media to share their views and opinions on government policies and other social issues. This post aims to get an opinion on the tweets of the political leaders that they tweeted during the election period. This paper also contributes to finding out the sentiment of the tweets in which these leaders are tagged. We then train our model after combining all the data sets, which helps us predict sentiment and what the accuracy is using machine learning algorithms.

II. LITERATURE SURVEY

Sentiment analysis deals with identifying and classifying opinions or sentiments that are present in the source text. Social media is generating a huge amount of sentiment-rich data in the form of tweets, status updates, reviews, blog posts etc. Sentiment analysis of this user-generated data is very useful in knowing the opinion of the crowd. Twitter sentiment analysis is arduous as compared to basic

 DOI: 10.5281/zenodo.10967535

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 3 | Issue 1 | February 2024

Journal Impact Factor 2023 (Quality Score): RPRI = 6.53 | SJIF = 3.647

www.ijirid.in

sentiment analysis due to the presence of slang words and misspellings. The maximum limit of characters that are allowed on Twitter is 140. A machine learning approach can be used for analyzing sentiments from the text. Some sentiment analyses are performed by analyzing Twitter posts about electronic products like cell phones, computers etc. using the Machine Learning approach. By performing sentiment analysis in a specific domain, it is possible to identify the effect of domain information in sentiment classification. They presented a new feature vector for classifying tweets as positive, negative, or neutral and extracting people's opinions about products [8].

Semantics is introduced as an additional feature to the training set for sentiment analysis in this paper [6]. The semantic features were fed into the Naive Bayes (NB) model training using the interpolation method. The Naive Bayes method, on the other hand, has several drawbacks, including data scarcity and zero frequency. As a result, other classifiers can be used to improve accuracy [8].

Twitter is an online social networking site and contains a huge number of active users who enthusiastically share their thoughts and reviews on events, news, products, sports, and elections. These reviews, written by the users, express their sentiments towards the topics 8 they tweeted. Fishing out sentiments embodied in the user's written text, in the world of social media is known as Sentiment analysis or opinion mining. Firmino Alves et al. (2013) state that from the beginning of the 21st century, sentiment analysis is one of the most interesting as well as active research topics in the domain of Natural Language Processing. It helps the decision maker to understand the responses of people towards a particular topic and aids in determining whether the event is positive, negative, or neutral. Twitter has been considered a very important platform for data mining by many researchers. Hemalatha et al. (2014) discuss that the Twitter platform contains more relevant information on events with hashtags that have been followed and accepted by many popular personalities [6].

III. METHODOLOGY

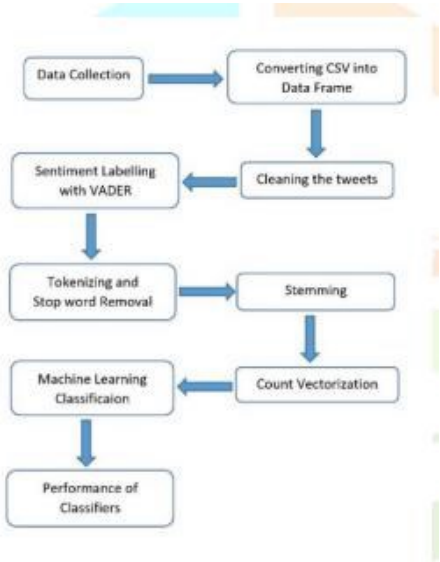

Figure 1: Methodology for the Twitter Sentiment Analysis

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 3 | Issue 1 | February 2024

Journal Impact Factor 2023 (Quality Score): RPRI = 6.53 | SJIF = 3.647

www.ijirid.in

A. Data Collection:

The Twitter API is considered a "Gold Mine of Data," so we decided to go with that. In contrast to other social media sites, almost every user's tweets are completely transparent and extractable, resulting in a broad database for analysis, as discussed in [2]. So, to extract Twitter data, one has to make a Twitter developer account. You have to provide some necessary details for creating an application that later will be used for extracting the data. After our application is created, we will get access to some keys such as customer keys, access token keys, customer secret keys, and access secret keys. When a user wants to get data, these keys are very important. To retrieve a tweet from Twitter, we created a Python script that uses the "Tweepy" Python library. Python has an open-source library called "Tweepy" that allows it to link to Twitter and collect data from their API, which we will use in our program.

B. Converting CSV into a Data Frame:

Only the related objects are kept from the text file, such as objects containing the entire text of a tweet, screen name, screen ID, date, and time at which time the tweet is created. The rest of the items will be discarded.

C. Cleaning the Tweets:

Twitter language is in an unstructured format. It may consist of emoticons, empty spaces, URLs, @tags, and hashtags. We have to pre-process the data first by using the Python libraries to retrieve the relevant parts of the data and remove the unwanted ones.

D. Sentiment Labelling with VADER:

Sentimental Analysis is a form of statistical analysis that determines whether a piece of text is negative, neutral, or positive and it is performed using one of two methods: Valence-based or Polarity-based. The text is graded as positive or negative in a polarity-based strategy. This ensures that the terms "good" and "superior" would have the same sentiment, i.e. positive. When analyzing a piece of text, VADER uses a valence-based approach. VADER analyses text sentiments using lexicons of sentiment-related words. It examines a text and determines if any of the words in the sentence are in its lexicon dictionary. It assigns them a positive, negative, or neutral rating [3].

VADER not only grades sentences based on the words in them but also the capitalization of the words and the sentence structure. For example, say the sentence "The weather is pleasant today, and I'm in great shape." is considered. In the sentence, "pleasant" and "great" are rated 0.51 and 0.62 respectively. VADER also assigns a score to the sentence depending on the use of exclamation points or emoticons. As a result, it is perfect for social media information. Not only that, but it also considers the usage of modifying terms preceding a sentiment term, such as" extremely,"" really,"" too," and so on. For example, "just good" reduces the positive intensity of a sentence, while "so good" increases the positive intensity.

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 3 | Issue 1 | February 2024

Journal Impact Factor 2023 (Quality Score): RPRI = 6.53 | SJIF = 3.647

www.ijirid.in

Another advantage of VADER is that it can accommodate changes in the sentiment of a sentence when it includes the word “but”. The sentiment of the sentence before and after the word “but” is considered by the rule, but the sentiment of the sentence after “but” is given greater weight than the one before it [4].

E. Stop Words Removal

Stop words are widely used words like ‘is’, ‘an’, ‘in’, ‘of’ etc. and others that are considered meaningless because they appear regularly in sentences and add no meaningful significance or weight to the sentiment. They unnecessarily increase the size of the data.

F. Stemming

Stemming reduces the words to their stems. Stemming is a raw method for removing the suffix from terms which also involves the elimination of derivational affixes. Stemming algorithms are majorly rule-based ones. For instance, they reduce all the ' consulting’ words to ‘consult’. It is important to note that the above two cleaning processes (Stop words Removal and Stemming) have not been done before (sentiment labelling) because doing so would result in irregularities in sentiment detection [4].

G. Count Vectorization

Count Vectorization is the transformation of any given text into a vector based on the frequency count of occurrence of any word in the text. It makes use of two features:

- min_df that defines the minimum frequency of a word to be used as a feature.
- ngram_range which is a tuple. It defines the minimum and maximum length of the sequence of tokens considered.

H. Machine Learning Classification

Classification is the method of constructing a model. Classification falls under the umbrella of supervised learning. A model can be thought of as a mathematical equation that is used to forecast a value by providing it with one or more values. It connects one or more independent variables to one or more dependent variables. The more appropriate the data and the greater the number of dependent variables, the more accurate the model would be. We divided our dataset into training and test datasets in our model by importing sklearn model selection.

Sklearn is a Python library that provides data processing functions such as clustering, classification, and model selection. Model selection divides the input data, which can be arrays, lists, or data frames, into random train and test datasets. The training set contains a known output and the model learns from this data to be generalized to other data later. We have the test dataset (or subset) to test our model’s prediction on this subset. We have then used the train test split function to make the split.

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 3 | Issue 1 | February 2024

Journal Impact Factor 2023 (Quality Score): RPRI = 6.53 | SJIF = 3.647

www.ijirid.in

The test size is 0.2, which means that 20% of the total data is test data, while the remaining 80% is training data. The function's output is stored in the variables X_train, y_train, X_test, and y-test. In our dataset, X_train and X_test are the real tweets, and y_train and y_test are the sentiments to which the tweet belongs. Since a classifier cannot operate directly on text data, we must first convert it to a vector using the Count Vectorizer function. Thus, X_train and X_test is vectorized and X_train (vector) and y-train are fed into the classifiers to train them. Following model training, X_test (vector) is fed into the model, and the model predicts each test data sent as input.

Our dataset is divided into three categories: positive, negative, and neutral. As the training data, the following classifier model is trained and fed with known positive, negative, and neutral tweets. After the classifier has been correctly trained, it can be used to detect an unknown tweet and automatically classify it [11]. The Sklearn library is used to import the different classifier models.

The following are the different classifier models that we have used:

1. **Decision Tree:**

   A decision tree model is a supervised learning algorithm that recursively partitions data into subsets based on features, creating a tree-like structure where leaf nodes represent class labels or predicted values. Splitting continues until stopping criteria are met, such as maximum depth or minimum samples per node. Decision trees are intuitive and handle various data types, but they may overfit and struggle with generalization. Despite limitations, they offer interpretability and automatic feature selection. Ensemble methods like Random Forests and Gradient Boosting enhance performance by combining multiple decision trees.

2. **Naive Bayes:**

   It is regarded as a generative learning model which is derived from the Bayes Theorem. Different class features are considered independent of one another in this model, regardless of their actual dependence on one another. All these factors contribute to the probability. Naive Bayes is useful for analysing large data sets and is easy to implement. Accuracy results that can be obtained are satisfactory given its simple model and the amount of data that can be handled using Naive Bayes as demonstrated in [5].

3. **Linear Regression:**

   A linear regression model is a fundamental supervised learning algorithm used for predicting a continuous target variable based on one or more input features. It assumes a linear relationship between the independent variables and the dependent variable, represented by a straight line in a multidimensional space. The model estimates the coefficients of the linear equation that best fits the data, minimizing the difference between the predicted and actual values using techniques such as ordinary least squares.

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 3 | Issue 1 | February 2024

Journal Impact Factor 2023 (Quality Score): RPRI = 6.53 | SJIF = 3.647

www.ijirid.in

Despite its simplicity, linear regression is widely used due to its interpretability, ease of implementation, and effectiveness in many real-world scenarios. However, it may not capture complex nonlinear relationships in the data. Regularization techniques like Lasso and Ridge regression can mitigate overfitting in situations with multicollinearity or high-dimensional data.[5]

4. K Nearest Neighbor:

KNN is a Lazy Learner classifier since it only uses training data to predict class. The training data is already labelled, so it learns to label new points using similarity measures and distance functions. It identifies the K nearest neighbours based on the distance function. Various distance functions that can be used are Euclidean distance, Manhattan distance, and Minkowski distance [6].

IV. PERFORMANCE EVALUATION OF CLASSIFIER MODEL

It is important to know which of the following classification algorithms works better for our dataset, i.e. which one makes the most accurate predictions after training them. A confusion matrix informs us about true positives, true negatives, false positives, and false negatives [7].

- TP (True Positive): when a case was found to be positive and predicted to be positive.
- TN (True Negative): when a case was negative and predicted to be negative.
- FN (False Negative): when a case was positive but predicted to be negative.
- FP (False Positive): when a case was found to be negative but predicted to be positive. The classification report is used to assess their prediction quality. Accuracy, Precision, and Recall are the most used performance evaluation metrics.
- Accuracy: It is the number of correct predictions over the total number of instances of data.
- Precision: It is the number of correct positive results over the total number of positive predicted results.
- Recall: It is the number of correct predicted results over the total number of actual positive results. Statistically accuracy, precision and recall are represented in the below equations.

$$Accuracy = TP + TN \,/\, TP + TN + FP + FN$$

$$Precision = TP \,/\, TP + FP$$

$$Recall = TP \,/\, TP + F\text{N}$$

V. RESULTS

The Twitter data set was used, which was obtained through the Twitter API. Around 100000 tweets were accessed from Twitter for training the classifiers Sentiment Analysis on Query “Election in India”.

 DOI: 10.5281/zenodo.10967535

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 3 | Issue 1 | February 2024

Journal Impact Factor 2023 (Quality Score): RPRI = 6.53 | SJIF = 3.647

www.ijirid.in

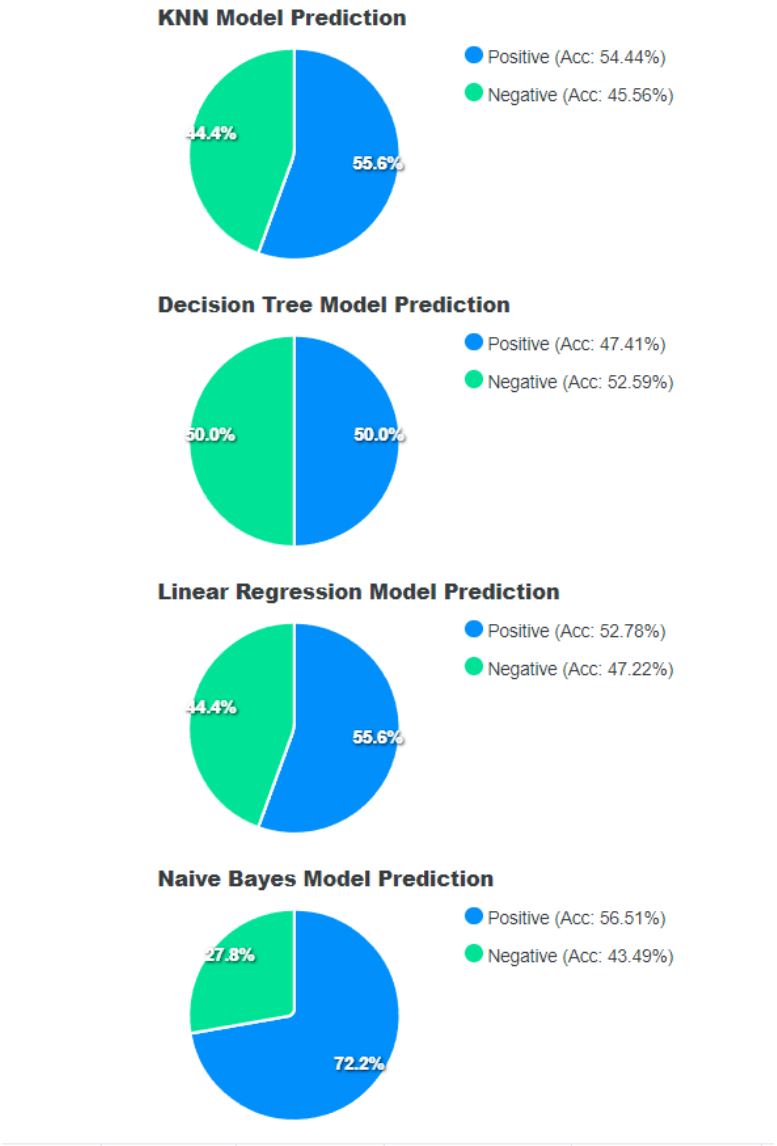

Figure 2: Different ML Models Prediction

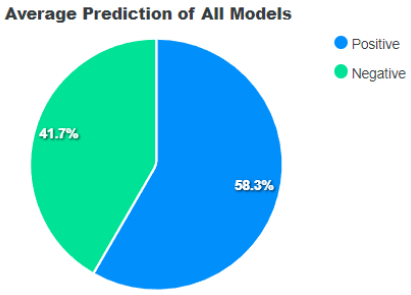

Figure 3: Average Prediction of All Models

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 3 | Issue 1 | February 2024

Journal Impact Factor 2023 (Quality Score): RPRI = 6.53 | SJIF = 3.647

www.ijirid.in

Table 1: Comparison of All Models

Model	Positive	Negative	Positive Accuracy (Avg)	Negative Accuracy (Avg)	Avg, Time Taken
KNN	12	8	52.00%	48.00%	0.11065s
Decision Tree	12	8	57.67%	42.33%	0.00104s
Linear Regression	11	9	52.08%	47.92%	0.00012s
Naive Bayes	15	5	57.49%	42.51%	0.00060s

VI. CONCLUSION

There are different Symbolic and Machine Learning techniques to identify sentiments from text. Machine Learning techniques are simpler and more efficient than Symbolic techniques. These techniques can be applied to Twitter sentiment analysis. There are certain issues while dealing with identifying emotional keywords from tweets having multiple keywords. It is also difficult to handle misspellings and slang words. To deal with these issues, an efficient feature vector is created by doing feature extraction in two steps after proper preprocessing. In the first step, Twitter-specific features are extracted and added to the feature vector. After that, these features are removed from tweets and again feature extraction is done as if it is done on normal text [9].

REFERENCES

[1] Sourav Das, Anup Kumar Kolya. "Sense GST: Text mining & sentiment analysis of GST tweets by Naive Bayes algorithm" International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), 2017.

[2] Pooja Dhede, Samruddhi Hagone, Gaurav Gaikwad, Gaurav Chaudhari, “Analyzing Awareness of Government Scheme using Swachh Bharat Tweets”, VJER-Vishwakarma Journal of Engineering Research, 2019.

[3] P. Singh, K. Singh Kahlon, R. Singh Sawhney, “Sentiment analysis of demonetization of 500- & 1000-rupee banknotes by the Indian Government”, 2018.

[4] Amolik, Akshay, Niketan, Jivane, Bhandari, Mahavir & Venkatesan, “Twitter Sentiment Analysis of Movie Reviews Using Machine Learning Techniques”, 2016.

[5] Kummer, O.; Savoy, J. Feature Weighting Strategies in Sentiment Analysis. In Proceedings of the First International Workshop on Sentiment Discovery from Affective Data, Bristol, UK, 24–28 September 2012.

[6] O’Keefe, T.; Koprinska, I. Feature selection and weighting methods in sentiment analysis. In Proceedings of the 14th Australasian Document Computing Symposium, Sydney, NSW, Australia, 4 December 2009.

[7] Chen, K.; Zhang, Z.; Long, J.; Zhang, H. Turning from TF-IDF to TF-IGM for term weighting in text classification. Expert Syst. Appl. 2016, 66, 245–260.

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 3 | Issue 1 | February 2024

Journal Impact Factor 2023 (Quality Score): RPRI = 6.53 | SJIF = 3.647

www.ijirid.in

[8] V.K. Chauhan, Dr. Amita Goel & A. Bansal, "Twitter Sentiment Analysis Using Vader," International Journal of Advance Research, Ideas, and Innovations in Technology, 2018.

