SJIF Impact Factor (2023): 8.574| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 8 | Issue: 3 | March 2023 - Peer Reviewed Journal

ANALYSIS OF TWITTER DATA USING MACHINE LEARNING ALGORITHMS

Sanchana.R[1], Josephine Ruth Fenitha[2], Shanmughapriya.M[3], Bhavani Sree. Sk[4], Nithyadevi.S[5]

[1]*Assistant Professor in Department of Information Technology,Sri Sairam Institute of Technology*
[2]*Department of Information Technology,Sri Sairam Institute of Technology*
[3]*Department of Information Technology, Sri Sairam Institute of Technology*
[4]*Department of Information Technology,Sri Sairam Institute of Technology*
[5]*Department of Information Technology,Sri Sairam Institute of Technology*

Article DOI: https://doi.org/10.36713/epra12585
DOI No: 10.36713/epra12585

ABSTRACT

Sentiment analysis is one among the distinguished fields of knowledge and pattern mining that deals with the identification and analysis of sentiment within the text. The main challenges in sentiment analysis are word ambiguity and multi polarity. The problem of word ambiguity is to define polarity because the polarity for words is context dependent. The tweets are initially preprocessed. The preprocessing includes the removal of stop words, and lower case conversion. The tweets are then passed to the feature extraction techniques. Then the data is splitted as training and testing data. The trained data is passed to the different machine learning algorithm like Naive Bayes. Support Vector machine, Random forest, and Decision Tree and k-NN algorithm. The accuracy obtained using the Naive Bayes. Support Vector machine, random forest, and Decision Tree, k-NN and Logistic regression algorithm is 80%, 77%, 72%, 61% ,56% and 78%. The naïve bayes algorithm has achieved a better accuracy when compared to the other algorithm.

KEYWORDS: *SVM, Naive bayes, Decision tree, Random forest*

I. INTRODUCTION

Natural language processing utilizes two main principal strategies they are syntax and semantic analysis. The arrangement of words in a sentence to make grammatical sense is called syntax analysis. Natural language processing uses syntax analysis techniques to assess meaning from a language based on grammatical rules. Syntax structure procedures incorporate parsing, word segmentation, sentence breaking, morphological segmentation, and stemming. Semantic uses a calculation to comprehend the importance and structure of a sentence. The techniques that NLP uses with semantics include word sense disambiguation, named entity recognition, and natural language generation. The word sense infers the importance of the word dependent on the specific situation. Named entity recognition decides whether the words can be categorized into groups. Natural language generation utilizes a database to decide the semantics behind the word. NLP uses a rule-based approach where machine learning techniques are used to train the models. The models are trained in such a way that when words and phrases appear in the particular text and are given a response when those particular phrases appear in the text. The primary use case for NLP is sentiment analysis. The sentiment analysis used by the data scientist can assess comments via web-based networking media to perceive how their business image is performing.

The challenging task in natural language processing is semantic analysis. NLP does not pick up sarcasm easily since it requires an understanding of the words being used and the context in which the words are being used. Natural language processing is additionally a difficult task by the means that language, and therefore the manner in which individuals use it, is persistently evolving. NLP makes use of a data engine that processes language and facial expression in order to monitor president trump's emotional state. Sentiment analysis is combined with image recognition to enhance accuracy. Image recognition techniques break down the image connected with the tweet, and then machine learning processes them together with the language to tell the actual emotionality of an image. Sentiment analysis is the process of determining whether or not the polarity in particular text is positive, negative or neutral. Sentiment analysis is one among the distinguished fields of knowledge and pattern mining that deals with the identification and analysis of sentiment within the text. Sentiment analysis is referred to as opinion mining. Opinion mining combines natural language processing and

SJIF Impact Factor (2023): 8.574| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 8 | Issue: 3 | March 2023 - Peer Reviewed Journal

machine learning techniques to assign weighted sentiment scores to each and every entity. The types of sentiment analysis are generally classified into fine-grained, emotion detection, aspect-based sentiment analysis, and intent analysis. Fine grained sentiment analysis provides different flavors of polarity by identifying the actual text as positive or negative sentiment. The text is related to a particular feeling such as anger disappointment or worries (negative feelings) or happiness; love (positive feelings). Emotion detection is used to detect emotions like happiness, anger, and sadness. The words that generally express anger may also express happiness depending on the context of the words being used. The aspect-based sentiment analysis breaks down the text into aspects. Each aspect is assigned with sentiment values as positive negative and neutral**.**

For instance "The battery lifetime of this camera is simply too short" The above example is expressing a negative opinion regarding the camera, but precisely it describes the battery life which is the explicit feature of the camera. The intent analysis refers to the text rather than the people say with the text. For example **"**I would like to know how to replace the things". The above example can be inferred from the text, but in many cases, inferring the text requires some contextual knowledge. Multilingual sentiment analyses are often troublesome task since it needs a lot of preprocessing techniques and preprocessing makes use of wide range of resources.

The main challenges in sentiment analysis are word ambiguity and multi polarity. The problem of word ambiguity is to define polarity because the polarity for words is context dependent. For example "The story of the movie is unpredictable" [1] "The steering wheel of the automotive is unpredictable" [2]. The primary example defines the polarity of the word "unpredictable" as positive. The second example defines the polarity of the word "unpredictable" as negative.

For instance **"**The audio quality of my phone is so cool however the display colors are not too good" Multi-polarity sentiment analysis model can assign a negative or a neutral polarity to the particular instance defined above. The overcome this drawback, every sentence must be assigned with polarity; here "audio" is assigned with positive polarity and "display" is assigned with negative polarity. Sentiment analysis provides a quantitative and qualitative data through which we can assess the success of the market campaign. Sentiment analysis plays a vital role in Monitoring the customers and spotting the negative comments early which will help the business parties to overcome the critical situation. The biggest advantage of sentiment analysis is to increase the sales revenue which in turn improves the products/service quality and customers services. Sentiment analysis is employed across a variety of applications and for multiple purposes. One such application is the prediction of election results. Sentiment analysis has been utilized by political candidate to observe overall opinion concerning policy changes enabling them to fine-tune their approach and to better relate to voters. In brand reputation management application, sentiment analysis enables brands to identify peaks in overall brand sentiment so enabling companies to make improvement absolutely with client demands.

II. LITERATURE SURVEY

A Twitter sentiment analysis study by Go et al does a two-classed (negative and positive) classification of tweets. The training data was preprocessed before it was used to train the classifier. The Preprocessing techniques include removing the user names and actual URLs and converting the classes into equivalence classes like 'URL' and 'USERNAME' respectively. To select useful uni-grams, they used such feature selection algorithms. The feature selection algorithm is frequency, mutual information, and chi-square method. Multinomial Naive Bayes, maximum entropy and support vector machines (SVM) are the three supervised techniques used for analysis. The accuracy of 84% was obtained with multinomial Naive Bayes using uni-gram features selected on the basis of their MI score. The accuracy was low when the tweets are processed using the bi-gram approach. The experiment did not recognize and they were not able to handle neutral tweets. To take into account neutral tweets, they collected tweets about a term that do not have emoticons. Nearly 33 tweets were manually annotated and they are used as test data. The manually annotated tweets where labeled as neutral. They merged these two datasets with the training data and test data used in the above two-classed classification. They trained a three-classed classifier and tested it, and achieved an accuracy of 78% .

Sentiment analysis on ensemble learning classifier was proposed by Ankit et al., (2018), focusing on ensemble classifier which combines the base learning classifier to form a single classifier that aims in redesigning the performance and accuracy of sentiment classification. The Ensemble approach takes the positive and negative scores of the tweet. The positive score is higher than the negative score then the score of sentiment tweet is positive. The negative score is higher than the positive score then the sentiment score is negative. If the positive and negative scores are equal then the system calculates the cosine similarity of that tweets. The cosine similarity of the tweets is in contrast with the testing data and identifies the most similar tweets. Then calculates the positive and negative score identified tweets. It failed to calculate the values of neutral tweets which incorporate neither positive nor negative sentiment.

Twitter sentiment analysis using the hybrid cuckoo search method was proposed by Avinash Chandra Pandey et al., (2017), a heuristic technique that is based on k-means and cuckoo search. The heuristic method is used to find the optimum cluster heads from the sentimental contents of the Twitter dataset. K-means data clustering method groups n data points in k clusters and distance can be calculated either by using Euclidean distance or cosine measures. K-means can be used for initial clustering which is the major

SJIF Impact Factor (2023): 8.574| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 8 | Issue: 3 | March 2023 - Peer Reviewed Journal

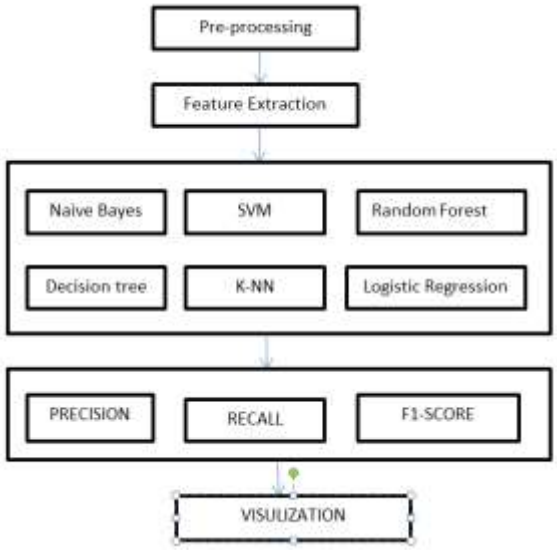
drawback. The generated cluster need to be further analyzed therefore cuckoo search method is used for further optimization of the cluster-heads. The cuckoo search approach makes use of a random initialization method that increases the number of iterations to converge and also stuck to some local solution. The process of the cuckoo search method results in faster convergence and better optimum solutions. The cuckoo search method modifies the initialization from k-means which resolves the problem of random initialization.

Twitter sentiment analysis was proposed by Barbosa and Feng (2010) used a two-phased approach for sentiment analysis. The two phases are: 1) classifying the dataset into objective and subjective classes (subjectivity detection) and 2) classifying subjective sentences into positive and negative classes (polarity detection). Suspecting that the use of n-grams for Twitter sentiment analysis might not be a good strategy since Twitter messages are short, they use two other features of tweets: Meta information about tweets and syntax of tweets. For meta-information, they use Parts of speech tags and mapping words to prior subjectivity (strong and weak), and prior polarity (negative, positive and neutral). The prior polarity is reversed when a negative expression precedes the word. For tweet syntax features, they use #, @, retweets, link, punctuations, emoticons, capitalized words, etc. SVM has achieved a better accuracy compared to the other machine learning techniques. For test data, 1000 tweets were manually annotated. The tweets are annotated as Positive, negative, and neutral tweet. The highest accuracy obtained was 81.9% on subjectivity detection followed by 81.3% on polarity detection.

III. ARCHITECTURE

The following figure 3.1 represents the architecture diagram of the Twitter data analysis using different machine learning algorithms. The dataset is collected from Twitter. The tweets are pre-processed to clean and transform the data for feature extraction. The tweets are initially pre-processed. The pre-processing includes the removal of stop words and lower case conversion. The tweets are then passed to the feature extraction techniques. Then the data is split as training and testing data. The trained data is passed to different machine learning algorithms like Naive Bayes. Support Vector Machine, Random Forest, Decision Tree, and k-NN algorithm. The precision, recall, and accuracy score are calculated and finally, the data are visualized in the graphical format.

Figure 3.1 Architecture Diagram of twitter data analysis



3.1. DATA PRE-PROCESSING

The data preprocessing includes the lower case conversion and stop words removal. The lower case conversion is done mainly because these words will be represented in two dimension vector spaces. The stop words are removed because these words occupy a unnecessary memory spaces and these words are not used during the analysis since they do not provide any meaning to the context.

3. 2. FEATURE EXTRACTION

Feature extraction plays a vital role in the part of text classification. It is based on vector space model. The feature extraction is an approach to minimize the quantity of resources required to describe a dataset. Analysis with a massive range of variables requires a large amount of memory and more computational power. The classification algorithm overfit the training sample and generalizes to new samples. Hence the feature extraction technique is essential while dealing with classification problems with a large number of variables. The multiple features are compared and different accuracy scores are achieved for different features. The vector space model views the text as a dot in the N-dimensional space. The different feature extraction techniques are bag of words, n-gram, TF-IDF. Bag of words is mainly used in the document classification. Document classification mainly uses each word as a feature for training the classifier. The main disadvantage of the model is that the order of occurrence of each word is lost because for each word it creates a vector of tokens in randomized order. In order to overcome this problem we go for an approach called N-gram. The N-gram approach works based on how often the word sequences occur in the corpus text.

SJIF Impact Factor (2023): 8.574| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 8 | Issue: 3 | March 2023 - Peer Reviewed Journal

Input: preprocessed tweet

Output: tokenized tweet
Step 1: Token = tokenization of the word
Step 2: Polygram = gram (token, 10)
Step 3: analysis = analyze the tweet using get sentiment function
Step 4: Gram input= [text, analysis]
Step 5: Append the input value with the gram input

s is Classification. It specifies the class to which the data element belongs to and it is used once the output has finite and distinct values. The sentiment analysis in social media has two potential outcomes, positive or negative.

"**The best economy in our lifetime"**

"The Obama administration built the cages not the Trump administration"

Classification techniques will categorize the data as positive, negative or neutral. The above example "the best economy in our life time" is categorized as a positive class label. "The Obama administration built the cages not the Trump Administration" is categorized as a negative label.

3.1 NAIVE BAYES ALGORITHM

Naive Bayes is a classification technique. The principle behind the Naive bayes algorithm is Bayes theorem. Bayes theorem defines the independence among predictors. There are two types of implementations – Bernoulli and Multinomial. The difference between the models is the way in which the features are extracted from the document. According to Bayes model the conditional probability can be calculated as

$$p\left(\frac{y}{x}\right) = p\left(\frac{x}{y}\right).p(y)/p(x) \quad [3.1]$$

[3.4] Represents a very large number of p(y/x) is the posterior probability of the class given attribute. The prior probability of the class is represented using P(y). P (x) is the prior probability of the attribute. P (x|y) is the likelihood which determines the probability of the attributes given class.

$$p(y|x1, x2 \ldots\ldots xn) = p(y)\pi p(x_i|y) \quad [3.2]$$

We have to find the probability of the given set of inputs for all possible values of the class y and the output which has the maximum probability. The main drawback of naive Bayes classifiers is highly scalable. The naive bayes model requires a large number of variables (features/predictors) in a learning problem. Naive bayes model is easy to build and particularly useful for handling large datasets. Naive Bayes is easy to implement and needs less training data.

Positive tweet analysis

$$p(tweet_{positive}) = \frac{p(totalcount_{positive})}{p(total_{tweetcount})} \quad [3.3]$$

Negative tweet analysis:

$$p(tweet_{negative}) = \frac{p(totalcount_{negative})}{p(total_{tweetcount})} \quad [3.4]$$

Neutral tweet analysis:

$$p(tweet_{neutral}) = \frac{p(totalcount_{neutral})}{p(total_{tweetcount})} \quad [3.5]$$

Naive bayes method is a popular basic method to categorize the text. It uses frequencies of words in document as the feature for classification. The advantage of Naive Bayes classifier is its high scalability as it requires parameters in proportion to the number of features or variables and that too linearly. Naive bayes requires a small set of data for training the classifier.

3.2 SUPPORT VECTOR MACHINE:

SVM is a supervised machine learning algorithm. Support vector machine may be used for both classification and regression. SVM performs classification by finding the hyper-plan that differentiates the categories that are premeditated in an n-dimensional area. The algorithm works comparatively well once there's a clear margin of separation between categories. SVM is more efficient in high dimensional space.

Prediction score

$$prediction_{svm} = predict(test\ ) \quad [3.6]$$

Accuracy score

$$accuracy_{score} = predict_{testy} * 100 \quad [3.7]$$

3.3 DECISION TREE

The decision tree is the most popular algorithm used for classification and prediction. The decision tree follows a tree-like structure that contains internal nodes, branch, and leaf nodes. The internal node represents an attribute, the branch represents an outcome and the leaf node holds the information about the class label. The decision tree is initially constructed by splitting the dataset into a subset based on the attribute. The process is repeated on the derived subset in a recursive manner. The main advantage of the decision tree is the construction of the tree does not require any domain knowledge. A decision tree can handle both continuous and categorical variables. The main disadvantage of using the decision tree - generates more errors in classification problems since we use a small number of training datasets.

3.4 LOGISTIC REGRESSION

Logistic regression is the machine learning algorithm that comes under the supervising learning techniques. In logistic regression, the outcome must be a categorical or a discrete value. The output of the logistic regression is a categorical dependent variable. The dependent variable can be predicted from the given set of independent variables. The logistic regression uses a cost function called the sigmoid function. The sigmoid function is mainly used to

SJIF Impact Factor (2023): 8.574| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 8 | Issue: 3 | March 2023 - Peer Reviewed Journal

map the predicted values corresponding to their probabilities. The linear regression classifier combines the weight of the input features. The weighted input features are passed as an input to the sigmoid function. The sigmoid function transforms the input into a number. The transformed input values should lie between 0 and 1.

3.5 K-NN ALGORITHM

The k-nearest algorithm is supervised machine learning. K-NN is also known as a non-parametric or lazy learner algorithm. The k-NN does not learn from the training set. The K-NN stores the training dataset and then classifies the data. The classified data is then grouped. The K-NN algorithm works by selecting the k neighbors. Then the Euclidean distance is calculated for the K number of neighbors. Take the k nearest neighbors from the Euclidian distance calculated. In each category count the number of data points. The new data points which are calculated are assigned based on the number of neighbors is maximum. Finally, the model is ready for evaluation. The main advantages of using the K-NN algorithm are more effective when the training dataset is large. The disadvantage of using the algorithm is to determine the k value which may be complex. The cost of calculating the distance between the data points for all the training datasets is high.

3.6 RANDOM FOREST ALGORITHM

The Random forest is a machine learning algorithm that can be used to solve both classification and regression based problems. Random forest algorithm belongs to a supervised learning technique. Random forest works based on the concept of ensemble learning technique. The ensemble method solves a complex problem by combining multiple classifiers in order to improve the performance of the model. The random forest contains a number of decision trees. The decision tree is classified based on the various subsets of the given datasets. Finally the average is taken in order to improve the predictive accuracy of the dataset. The number of trees constructed in the forest is greater leads to higher accuracy and preventing the problem of overfitting. The random forest algorithm works by selecting the random data points K from the training dataset. The decision tree is constructed based on the selected data points. If there are only few data points then find the predictions of each decision tree. Then assign the new data point to the category which has majority votes. The advantage of the random forest model improves the accuracy and resolves the problem of overfitting. The random forest algorithm is not more suitable for the regression-based problem.

V. RESULT AND DISCUSSION

5.1 NAIVE BAYES ALGORITHM

The following table 5.1 represents the precision-recall and f1-score achieved using the Naive Bayes algorithm. The naïve Bayes algorithm is calculated against two values i.e. 0 or 1. The precision, recall, and f1-score achieved using the positive tweets are 80%, 69%, and 82%. The precision, recall, and f1-score achieved using the negative tweet is 78%, 87%, and 82%.

Table 5.1 Precision, Recall and F1-score using Naive Bayes

	Precision	Recall	F1-score	Support
0	0.78	0.87	0.82	4500
1	0.80	0.69	0.74	3543

The following figure 5.1 represents the confusion matrix obtained using the Naive Bayes algorithm. The confusion matrix is plotted against the true negative, false positive, false negative and true positive values. The true negative value is 3912, the false positive value is 588, the false negative value is 1116 and the true positive value is 2427.

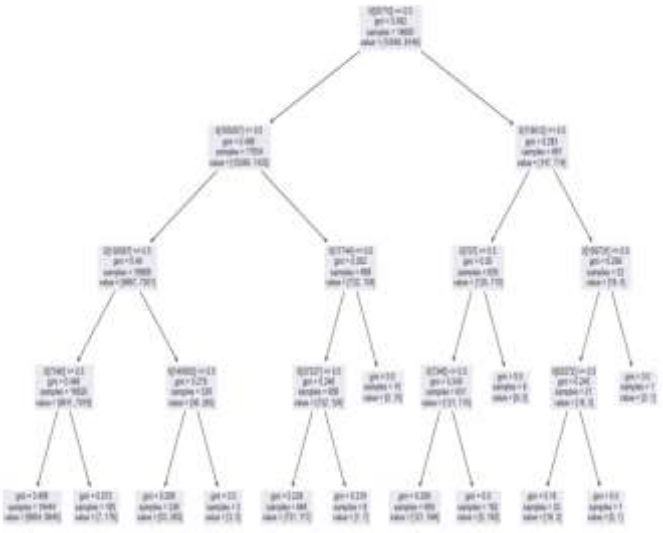
Figure 5.1 Confusion matrix using Naive bayes

5.2 DECISION TREE

The following figure 5.2 represent the construction of the decision tree with the given set of input data.

SJIF Impact Factor (2023): 8.574| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 8 | Issue: 3 | March 2023 - Peer Reviewed Journal

Figure 5.2 Decision Tree Constructions



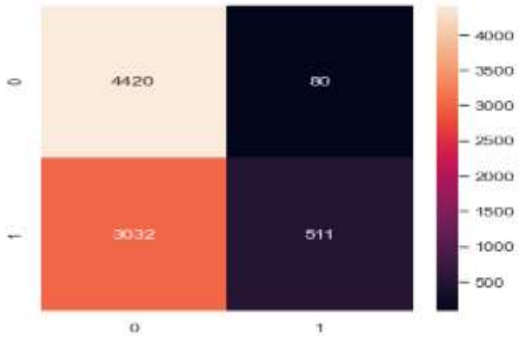
The following table 5.2 represents the precision recall and f1-scoreachieved using the decision tree algorithm. Decision tree algorithm is calculated against two values i.e. 0 or 1. The precision, recall and f1-score achieved using the positive tweets are 86%, 25% and 25%. The precision, recall and f1-score achieved using the negative tweet is 59%, 98% and 74%.

Table 5.2 Precision, Recall and F1-score using Decision tree

	Precision	Recall	F1-score	Support
0	0.59	0.98	0.74	4500
1	0.86	0.25	0.25	3543

The following figure 5.3 represents the confusion matrix obtained using the Decision tree algorithm. The confusion matrix is plotted against the true negative, false positive, false negative and true positive values. The true negative value is 4420, the false positive value is 80, the false negative value is 3032 and the true positive value is 511.

Figure 5.2 Confusion matrix using Decision tree



5.3 SUPPORT VECTOR MACHINE

The following table 5.3 represents the precision recall and f1-score achieved using the support vector algorithm. Support vector algorithm is calculated against two values i.e. 0 or 1. The precision, recall and f1-score achieved using the positive tweets are 80%, 64% and 71%. The precision, recall and f1-score achieved using the negative tweet is 76%, 87% and 81%.

Table 5.3 Precision, Recall and F1-score using Support vector machine

	Precision	Recall	F1-score	Support
0	0.76	0.87	0.81	4500
1	0.80	0.64	0.71	3543

The following figure 5.3 represents the confusion matrix obtained using the Support vector machine algorithm. The confusion matrix is plotted against the true negative, false positive, false negative and true positive values. The true negative value is 3918, the false positive value is 582, the false negative value is 1263 and the true positive value is 2280.

Figure 5.3 Confusion matrix using Support Vector Machine

SJIF Impact Factor (2023): 8.574| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 8 | Issue: 3 | March 2023 - Peer Reviewed Journal

5.4 LOGISTIC REGRESSION

The following table 5.4 represents the precision recall and f1-score achieved using the logistic regression algorithm. Logistic regression algorithm is calculated against two values i.e. 0 or 1. The precision, recall and f1-score achieved using the positive tweets are 81%, 65% and 72%. The precision, recall and f1-score achieved using the negative tweet is 76%, 88% and 82%.

Table 5.4 Precision, Recall and F1-score using Logistic Regression

	Precision	Recall	F1-score	Support
0	0.76	0.88	0.82	4500
1	0.81	0.65	0.72	3543

The following figure 5.4 represents the confusion matrix obtained using the Logistic Regression algorithm. The confusion matrix is plotted against the true negative, false positive, false negative and true positive values. The true negative value is 3969, the false positive value is 531, the false negative value is 1255 and the true positive value is 2288.

Figure 5.4 Confusion matrix using Logistic Regression

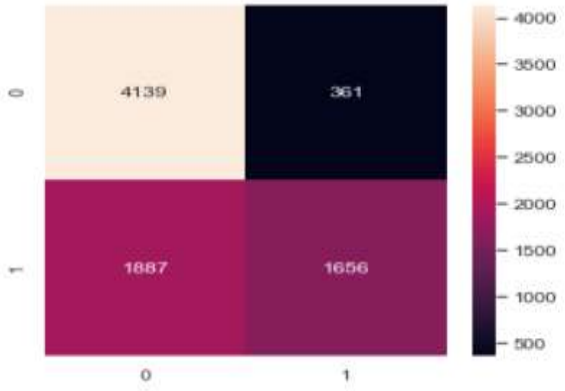
5.5 RANDOM FOREST

The following table 5.4 represents the precision recall and f1-score achieved using the Random forest algorithm. Random forest algorithm is calculated against two values i.e. 0 or 1. The precision, recall and f1-score achieved using the positive tweets are 82%, 47% and 60%. The precision, recall and f1-score achieved using the negative tweet is 69%, 92% and 79%.

Table 5.5 Precision, Recall and F1-score using Random *Forest*

	Precision	Recall	F1-score	Support
0	0.69	0.92	0.79	4500
1	0.82	0.47	0.60	3543

The following figure 5.5 represent the confusion matrix obtained using the random forest algorithm. The confusion matrix is plotted against the true negative , false positive, false negative and true positive values. The true negative values is 4139, the false positive value is 361,the false negative value is 1887and the true positive value is 1656.



5.6 KNN

The following table 5.4 represents the precision recall and f1-score achieved using the logistic regression algorithm. Logistic regression algorithm is calculated against two values i.e. 0 or 1. The precision, recall and f1-score achieved using the positive tweets are 66%, 50% and 25%. The precision, recall and f1-score achieved using the negative tweet is 57%,97%,72%.

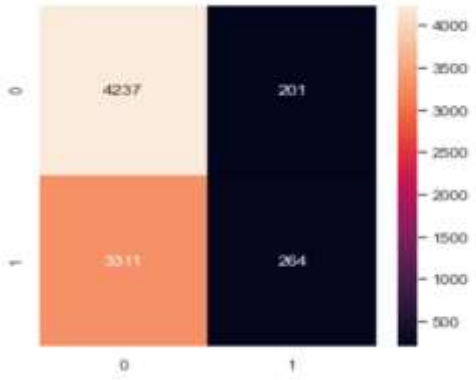
Table 5.6 Precision, Recall and F1-score using Logistic Regression

	Precision	Recall	F1-score	Support
0	0.57	0.97	0.72	4500
1	0.66	0.50	0.25	3543

The following figure 5.6 represent the confusion matrix obtained using the random forest algorithm. The confusion matrix is plotted against the true negative, false positive, false negative and true positive values. The true negative values are 4237, the false positive value is 201, false negative value is 3311 and the true positive value is 264.

SJIF Impact Factor (2023): 8.574| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 8 | Issue: 3 | March 2023 - Peer Reviewed Journal

Figure 5.6 Confusion matrix using Random Forest



5.7 TESTING ACCURACY

The input data is splitted as training and testing data. The training data is nearly 70% of the dataset which is used to teach the machine learning models. The test data is 20% of the given dataset where it is used to test the data for the given input, the outputs are derived correctly. The following table 5.7 represent the accuracy obtained using the test data.

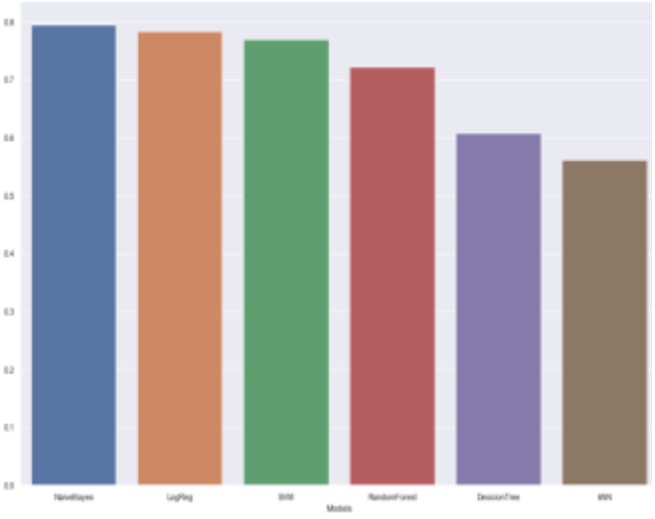
Table 5.7 Accuracy of test data

ALGORITHM	ACCURACY OF TEST DATA
Naive Bayes	80%
K-NN	56%
Random Forest	72%
Decision Tree	61%
Logistic Regression	78%
Support Vector Machine	77%

5.8 GRAPHICAL REPRESENTATIO

The graph is plotted against the different machine learning model and their corresponding accuracy. The x axis contains the accuracy and the y-axis is plotted against the different machine learning models. The following figure 5.7 represents the graphical representation of different machine learning model.

Figure 5.7 Graphical representations of different machine learning models



VI CONCLUSION

The proposed work analyzed the information spread over social networks like Twitter. Sentiment analysis is a uniquely powerful tool for the analysis of election results that is looking to measure attitudes, feelings, and emotions. Nevertheless, the evolution of social media analysis regarding politics is still considered to be in its infancy, despite the importance of the topic for political science. The dataset is collected from Twitter. The tweets are pre-processed to clean and transform the data for feature extraction. . The tweets are then passed to the feature extraction techniques. Then the data is split as training and testing data. The trained data is passed to different machine learning algorithms like Naive Bayes. Support Vector Machine, Random Forest, Decision Tree, and k-NN algorithm.

The accuracy obtained using the Naive Bayes. Support Vector machine, random forest, and Decision Tree, k-NN and Logistic regression algorithm is 80%, 77%, 72%, 61% ,56% and 78%. The naïve bayes algorithm has achieved a better accuracy when compared to the other algorithm.

VII. FUTURE WORK

There is a vast future scope in this research area as it is not only interesting but also challenging in the real world. In the future, sarcasm can be detected in text. Many tools and algorithms rely on the polarity of the words and the scoring is dependent on polarity. This means that accuracy drops since the semantics of the complete sentence are lost. To measure the polarity of the sentence on each and every individual words the semantics of the sentence. The semantics of the sentences make it difficult to identify the polarity. For instance, the financial industry has its own language which means completely differs from the Entertainment industry. This makes it hard for the tool to predict the emotion or semantics of the sentence.

SJIF Impact Factor (2023): 8.574| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 8 | Issue: 3 | March 2023 - Peer Reviewed Journal

REFERENCES

1. *Ahmad S, Asghar MZ, Alotaibi FM, Awan I(2018), "Detection and classification of social media-based extremist affiliations using sentiment analysis techniques" Vol.9,pp.1-24.*
2. *Alotaibi FM, Awan I (2016), "American presidential election. Journal of Internet Services and Applications" Vol.9, pp.1-18.*
3. *Budiharto & Meiliana, M. (2018), "Prediction and analysis of Indonesia Presidential election from Twitter sentiment analysis".Vol.5, pp.1-51.*
4. *Budiharto W, MeilianaM (2018),"Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis". Journal of Big Data. Vol.1, pp.5-51.*
5. *Caetano JA, Lima HS, Marques-Neto HT(2017), "sentiment analysis using twitter" Vol.5, pp.8-24.*
6. *Santos MF (2018), "Using sentiment analysis to define twitter political users classes and their homophily during the American presidential election"-Journal of Internet Services and Applications Vol.8,pp.70-77.*
7. *Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., & Stoyanov, V. (2016),"Sentiment analysis in Twitter. In Proceedings of the 10th international workshop on semantic evaluation Vol.5, pp.1-18.*
8. *Pandey, A.C., Rajpoot, D.S. and Saraswat, M (2017), "Twitter sentiment analysis using hybrid cuckoo search method. Information Processing & Management" Vol.53 (4), pp.764-779.*
9. *Saif, H., He, Y., & Alani, H. (2012), "Semantic sentiment analysis of twitter. In International semantic web conference" Vol.5,pp. 508-524.*