A Machine Learning Approach of Sentiment Analysis on Twitter

Mirza Tabish Hassan, Apurva Goel, Nazakat Ali Sofi
Department of Computer Science and Engineering, KIET Group of Institutions, Ghaziabad, UP, India

Article Info

Keywords:
Machine Learning
weblog
Sentiment Analysis
TF-IDF

Abstract

Social media has appeared as one of the platforms to uplift user's thinking and judgement. Twitter provides a way to share someone's emotion, thinking, opinion, feeling and mood with others in limited text, video or image. A good analysis of twitter data may be beneficial for any organization like political, commercial and may decide it's future course. This research helps in getting the opinions and thoughts of the masses about any hot topic prevailing in the present times like elections, covid-19 or any other crucial issues, as a data to analyze the support of masses towards the cause and as well their repulsion and reflection from it. In this work nine supervised machine learning algorithms are applied to analyze tweets. A dataset has been taken from the online repository to analyze tweets and extract the sentiments with the help of the machine learning algorithms. Further a comparative study of these machine learning algorithms on the basis of their performance parameter like precision, f1-score and recall.

1. Introduction

Online social media platforms are the way to share the opinions, thoughts, suggestions about any topic or product by the public. These platforms are used by different communities like political, commercial or social to share their ideas, work to the public for the promotion. The social media platforms are also used by different commercial companies to advertise their product, service and share their benefits to the public. The political sector also uses different social media platforms to promote their work in the public and get recognition. The public uses these platforms to share their feedback about any political work, commercial product or social issue. The feedback given by the public is used to enhance the product or service of any company. It is also used to develop new business strategies as per the public demand.

There are different social media platforms which is used to share the opinions, thoughts and emotions of the public like Facebook, Instagram, Twitter etc. The Facebook and Instagram are also used for sharing the status, picture of the someone's life along with there opinions while Twitter is a social media platform which is mainly used for the purpose sharing the opinions, thoughts, emotions and feedback. A proper analysis of content available on social media platforms may be beneficial for any organization or government.

The process of Sentiment Analysis is a technique of analyzing the tweets and determining the sentiment in the form of positive, neutral, or negative remarks. These categories of the tweets are classified on the basis of their polarity scaling from -1 to 1. The tweets having the polarity range from -1 to less than 0 are classified as negative tweets. The tweets having polarity scale equal to 0 are classified as neutral tweets. The rest of tweets having polarity ranging from greater than 0 to 1 is classified under positive tweets. This study provides a way of extracting sentiment of mass from the available tweets on Twitter platform. In order to analyze

the sentiment, this paper uses the data gathered from the reputed online repository Kaggle. After getting the tweets, nine supervised machine learning algorithms are used to classify the sentiments of the tweets gathered. Rest of the paper is divided in nine sections. Section 1 describes the brief introduction about the paper. Section 2 describes the flow of the methods used in the work done for retrieving the conclusion of the paper. Section 3 shows the study of related work which is available on the online repository. Section 4 describes about the source from which we have gathered the data. Section 5 refers to a detailed idea of the pre-processing of dataset. Section 6 refers to the detailed study of the different machine learning algorithms used in the paper. Section 7 is the model prediction which refers to the performance parameter and the result gathered from different algorithms of machine learning. Section 8 is the conclusion which is derived from this paper. Section 9 describes the future use of the research in detail.

2. Taxonomy of research

This section of the paper represents the workflow of the process used in the research in form of a flow diagram. The research starts with gathering the data from an online reputed repository Kaggle. The dataset contains 1.6M tweets from twitter. After that, the pre-processing of the data is done using NLP. There is various method of pre-processing of the data but the best among all is natural language processing which is being used in our research. The processed data is then splitted into two datasets (training set and testing set). Further a model of machine learning is selected. The model is trained over the training dataset. After training of the model, the testing dataset is provided to the model and performance parameter is evaluated. This step after the processing is repeated with multiple models of machine learning (in this paper 9 algorithms of machine learning is used). At the end a detailed comparative analysis of the results of different machine learning algorithm is done to obtain the conclusion. The workflow is visualized in the figure 1.

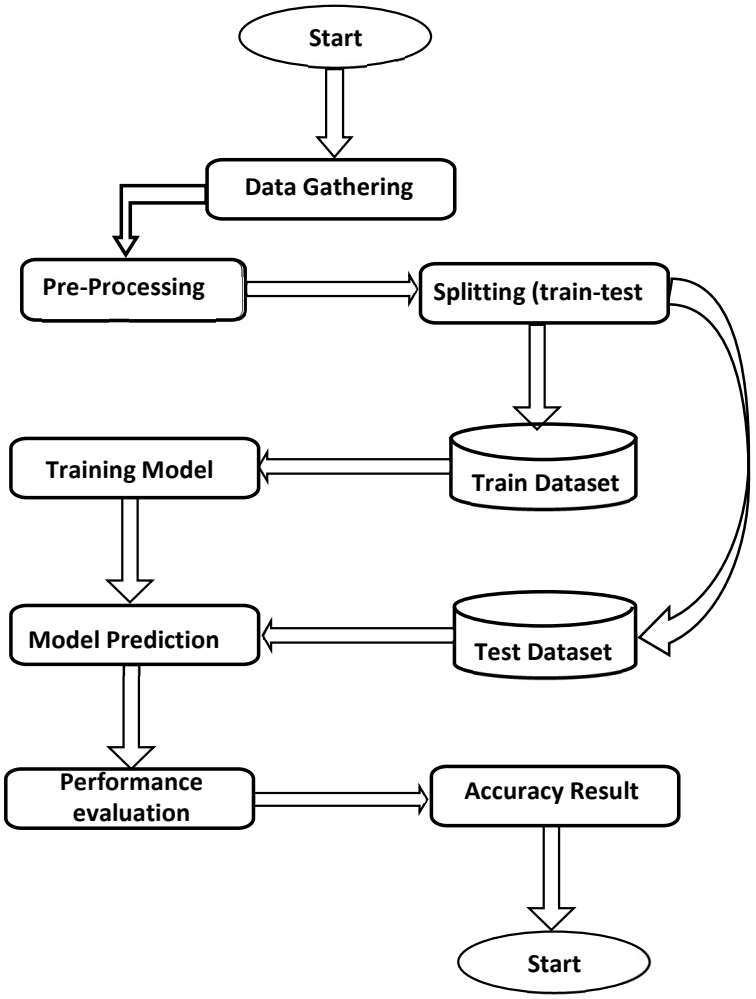

Figure 1

The steps which are elaborated in the further section are stated below:

- Data Gathering
- Pre-processing
- Splitting (train – test split)
- Model Training
- Model Prediction
- Conclusion

3. Related Works

[1]This paper describes about the different processes used for the sentiment analysis. The paper mentions the different approaches of machine learning which can be used for the implementation of sentiment analysis. The techniques the paper used are the lexicon-based sentiment analysis, machine learning approach and rule-based approach. The classification algorithms the paper had described are deep learning techniques. These techniques include bidirectional encoder representations from transformers (BERT), LSTM (Long Short-Term Memory Model) and GRU (Gated Recurrent Unit). The paper has also discussed about some of the machine learning algorithms that include Decision Tree, K-NN, Naive Bayes, SVM. The paper does a comparative study between all these algorithms and classifiers. The papers have taken different performance parameters like precision, f1-score, recall and accuracy to analyze the best algorithm and classifier for sentiment analysis.

[2]The paper had discussed about the sentiment analysis of tweets fetched from the twitter api using different machine learning algorithm. The aim of the research was to survey and inspect the feelings and emotions of public especially farmer community about their objection and opposition. In this research the data is gathered from social media platform, Twitter using twitter api. On daily basis, 200 tweets were fetched from twitter using hashtag like '#FarmerProtest'. The data collected from the twitter is appended in a list data structure in python for a duration of 5 months. The paper uses a set of 18000 tweets as a dataset. After data gathering, the pre-processing of the tweets is done to make the tweets free from unwanted tokens. The dataset is passed through the TextBlob to analyze the sentiments for supervised learning of algorithms used in the research. The vectorization of the data is done through 2 methods for fitting the data in machine learning algorithm. The paper uses both bag of words and TF-IDF for this process. After the tweets have been converted into the tokens, the research uses 4 machine learning algorithms for prediction of the sentiments. The dataset is splitted in 2:8 ratio for testing and training set respectively. The training set is used to train the algorithm like naïve bayes, support vector machine, decision tree classifier and random forest. The paper concludes that the support vector machine with bag of word vectorization has the highest accuracy.

[3]The paper focus on the analysis of the sentiments of any real time tweet using a website as interface. The paper uses natural language processing for the cleaning of the data. Naïve Bayes is used for the prediction of the sentiment of the tweet fetched from the twitter using api. The algorithm is trained using training dataset and integrated with the user interface. The training dataset used in the paper is a movie review from imdb. The user interface accepts the user handle, tweets of any user and fetch an appropriate number of tweets from twitter. After the tweets has been stored into a data frame, pre trained naïve bayes machine learning algorithm is used to predict the sentiment of the tweets. The classification of the tweets is plot on a graph using matplotlib. The graph is displayed on the website. Along with the graph, the user can see the positive, negative and neutral tweets separately. The result of the paper was a real time sentiment analysis on tweets fetched from the api of twitter on any specific topic, user, or type.

[4], [5]The paper has analyzed the accuracy and performance parameter like precision, recall and f1-score of naïve bayes supervised machine learning algorithm. The data is gathered from an online repository having tweets with their sentiments as target. The cleaning of the tweets was done by using natural language processing. The paper uses multiple techniques for the cleaning of tweets under natural language processing like stemming, lexicon analysis an tokenization. After the data has been preprocessed, bag of word is used to represent the

tweets into numerical from so that the machine can understand. The data is splitted into training and testing set in appropriate ratio. The machine learning algorithm is trained on the training set of data. Further, the model is used to predict the sentiment of the tweets in testing dataset. The result is compared with the original target and the performance parameter like accuracy and precision is calculated.

[4]There are millions of social media users generating large amount of data and information (big data) that can both structured and unstructured. Every person influenced from social networking sites feels great to share their views and thinking about the present action or state they had gone through. This information is expressed to put the views of users about any chosen topic or issue. This research derives the feelings behind these posts and selected Twitter platform as the data source to begin the work. This research examines and inspect the methods of pre-processing and classification of textual data using python to find the sentiment as outcome of this textual twitter data. The information and textual data are collected from the social networking site Twitter with the use of API of Twitter by using the access token and access token secret for the need of authorization and identification. After the data has been collected, TextBlob is used to compute the target value for training the algorithm. The target is computed in form of the polarity and subjectivity. On the basis of polarity and subjectivity, the tweets are classified in positive, negative and neutral. The NLTK is used for data processing and the machine learning technique k-nearest neighbor is used for the classification of data against the target. The classified tweets are compared with the target and performance parameter is computed.

[6]The paper aims to analyze the sentiments of the feedback of any product in different levels. The sentiments are classified into three category positive, negative and neutral having different levels of sentiments. The paper describes about different techniques of sentiment analysis and there use cases. The techniques used in the paper are deep learning, different machine learning algorithms and lexicon-based methods. The paper also discusses about the different problems faced during the analysis of sentiments using these techniques. The paper concluded that the deep learning method is an appropriate method for the classification of sentiments of any huge data. Along with this, the paper also concluded that if public gives the proper feedback of the merchandise or product, they buy from the market then sentiment analysis will help other to know about the cons and pros of the product.

[7]The paper discusses about the techniques of machine learning for the analysis of sentiment of a movie review. The data set use for the analysis of the sentiments is a movie review taken from an online repository imdb. Further, natural language processing is used for the cleaning of the review. Using different techniques like stemming, tokenization, stop-word removal and advance python programming, the tokens (words or fragments of any textual data) which does not contribute to the analysis is removed. The pre-processed data is further passed through different algorithms of machine learning one by one. The machine learning algorithm used for analysis are naïve bayes, support vector machine, random forest, decision tree classifier and k-nearest neighbor. The performance parameter and accuracy are written down in a table and a comparative analysis is done in the research.

[8] The paper discusses about the classification of the sentiments in form of positive, negative and neutral by using transfer learning approach. The data is gathered form different online repository like imdb, Kaggle and other. The data used for the research is the product review on amazon and movie review form imdb. The pre-processing of the data is done by passing the data through different natural language processing techniques. The preprocessing removes the unwanted words including punctuations and numbers which does

not affect the accuracy of the prediction. After the preprocessing, transfer learning mechanism of machine learning is used. In transfer learning, the algorithm is trained on a domain and then tested on different domain. In the paper, for fitting different domain in the model data has been fetched from 5 source like are IMDB, Stanford sentiment Treebank, YELP, SENTIMENT140 (STS) and Amazon product. The paper also discusses about the different techniques which can be used to analyze the sentiments of any text. The paper concluded that among different methodology AWD-LSTM has the highest accuracy of 95.40%.

[9]The paper discusses about the sentiment analysis of data using recurrent neural network in deep learning. The data used for the analysis is related to the covid-19 and fetched from the twitter api. The data is fetched on daily basis over a month and appended in a data frame. The pre-processing of the data gathered is done using stemming and tokenization. After the data has been pre-processed, the sentiment is classified using TextBlob, which is inbuilt sentiment analyzing library in python. The sentiment analyzed from the TextBlob is used as target for the machine learning algorithm. A recurrent neural network from deep learning is used as a machine learning model. The model is trained and tested over the data gathered. The paper gives the public sentiments about the covid-19 and how bad the pandemic hits the mass. The paper also mentions about a user interface for a real time analysis of sentiments using machine learning.

[10] The paper discussed about the way of improving the efficiency using feature ensemble learning. The paper has proposed a method of sentiment analysis which can improve the efficiency of the algorithms. The method proposed in the paper is started by fetching the raw tweets using the twitter api or any online resource. The data is cleaned by removing the punctuations, URL, hashtag, and other tokens which does not contribute in the performance analysis. The data tokenization and POS-tagging is done after the cleaning for the pre-processing. After the data has been pre-processed, features of the preprocessed data are extracted like N-gram, Negative words, Positive words etc. The features extracted is then converted into feature vectors. This vector is passed through convolution artificial intelligence layer for training of the network. When the model is trained, the sentiment is predicted in 5 categories as strong positive, positive, neutral, negative and strong negative. This provides, a brief classification of the sentiments with their levels.

[11]The paper discusses about the BERT i.e., bidirectional encoder representations from transformer for the analysis of sentiments. The convectional method uses unidirectional encoder for different nlp techniques like sentiment analysis. The paper discussed that the performance can be increased by using bidirectional encoder. The bidirectional encoder technique encodes the words form both directions i.e., right to left and left to right. The paper also used aspect learning method for the analysis of sentiment. In this method words having a sentiment in a sentence is related to the previous word for the analysis. The paper uses two levels of the aspect models i.e., sentence level and text level. The paper has analyzed the sentiment on both the levels and concluded that the performance of the natural language processing can be improve by using BERT with a dataset having more unique aspect.

[12] The paper proposed different sentiment analysis techniques using deep learning. The different datasets mentioned in the paper are IMDB, Yelp, Amazon review dataset, MOUD dataset, Getty Images dataset, Twitter image dataset, CMU-MOSI dataset, and Stanford sentiment Treebank dataset. The data is cleaned and preprocessed using convectional method. The different deep learning techniques used in the research are Convolutional neural networks (CNNs), Recursive neural networks, long short-term memory, Recurrent neural networks, Gated recurrent units and Deep Belief Networks (DBNs). These algorithms are used to train

different models and all the models are tested on testing set. The performance parameters are computed, and a comparative study of these model is done in the paper. The paper also discussed about the advantages and disadvantages of each algorithm and concluded that the model having CNN, LSTM, Attention, and late fusion has the highest accuracy of 96.40%.

[13]The paper discusses about document level and sentence level sentiment analysis along with twitter sentiment analysis. In document level approach, document is given as an input and sentiment of each document is analyzed using machine learning. In sentence level approach, the sentiment of each sentence is analyzed using machine learning. The paper also discusses about the different classification techniques which can be implemented on these approached for prediction of sentiment. The classification techniques discussed are Naïve bayes, Support vector machine and maximum entropy method. The paper concludes that the support vector machine on data fetched from twitter api has an accuracy of 92%.

[14]The paper describes about the AI based sentiment analysis with an improved efficiency. The data used for the research is fetch from 3 online repository which are google scholar, ieee xplore and acm digital library. The data used in the research has 52 list (26 from google scholar, 25 from ieee xplore and 1 from acm digital library). After the data has been gathered, different classifiers with CNN are used for the model training and prediction. The model used in the paper are CNN+LSTM, CNN+BERT, CNN+RNN and CNN+PNN(BiGRU). The method followed by the research is started with the extraction of word features. After the word features are extracted, the data is splitted into training and testing set which is used for the training and prediction of the sentiments. The paper concludes that the model having CNN with LSTM has the highest accuracy of 92%.

[15]In traditional sentiment analysis, the sentiment words and aspect sentiment pairs are considered while training the model. In new approach of pre training, these features are neglected which can decrease the efficiency of the algorithm. The paper discusses about a technique known as sentiment knowledge enhance pre-training. In the paper sentence level approach is followed. The data is gathered from two online source that are Standford Sentiment Treebank and Amazon-2. The data is cleaned and preprocessed. After the data has been prepared for the model training, neural network is used as model. The model training uses around 10 epochs on the sst-2 dataset and 3 epochs on amazon dataset. The model is then evaluated on the testing set and the performance of each is compared.

[16]The paper discusses about the process of sentiment analysis on any textual data. In this paper the sentiment is analyzed on the concept level of the data. The textual data gathered from any source is passed through a cleaning process. The words, numbers and punctuations which does not contribute to the analysis is removed in cleaning. The cleaned data is then passed through a stemmer in which the words in the sentence are converted into there root form. The sentence is then tokenized, and the feature is extracted from the data. Based on the features extracted from the data the sentiment is determined. The paper discusses about some of the machine learning techniques such as naïve bayes, support vector machine, Bayesian network, decision tree classifier and neural network. These algorithms are used for determining the sentiments on different domains of data. The paper mentioned about social media, marketing and product review

[17] During Covid 19 pandemic, people were started sharing there opinion on online platform as they were not able to meet other. The paper focused on the sentiment analysis of the tweets done by the people on twitter on covid 19. The data in the research is fetched using twitter api on daily basis and appended in a data frame.

In a week, 530232 tweets have been gathered for the further process. After the data has been fetched, cleaning and tokenizing of the dataset is done. The tweets are categorized in three categories (positive, negative and neutral) with the help of TextBlob. After the classification, the data is splitted into training and testing sets for training and performance evaluation respectively. Naïve bayes algorithm is used as a model in the paper. The model is trained on training set under supervised learning algorithm and the testing set is used form evaluating the performance parameters.

[18]The research discuss about the sentiment analysis of textual data in multiple language. The main focus of the research is determining the sentiments of under-resourced language. The under resourced language refers to the language having minimal resource for the digitalization. The research uses two methodology one is machine translation and other is machine learning for the conversion of the language into English. After the language has been translated in English, the preprocessing of the data is done. After preprocessing, the sentiment is analyzed by using co-training model. In co-training model, the trading is continued while the sentiment is analyzed. The paper concluded that using co-training method for multilinguistic sentiment analysis can improve the accuracy of the model.

[19]The paper discusses about the product review of any e-commerce product in Chinese. The model used for the product review in the research is deep learning. In the research, the advantages of sentiment lexicon, CNN model, GRU model and attention mechanism is combined to improve the accuracy. A sentiment lexicon is constructed which is use to give a weight corresponding to words. After the construction of sentiment lexicon, different layers are combined in series to improve the accuracy. The layers combined are embedded layer followed by convolution layer, pooling layer, BiGRU layer and attention layer. The model is trained on the dataset containing 100000 reviews fetched from Dangdang online repository. The data is pre-processed using natural language processing. After the preprocessing, the data is passed through the layers for training and testing. The performance parameters like precision, accuracy and f1-score are evaluated. The paper concluded that the accuracy of the model is 93.5% on 15 epochs on the model.

[20]The paper discussed about the sentiment analysis of tweets fetched from the twitter using supervised machine learning algorithm. The data for the research is gathered from the publicly available online source and a total of 250000 tweets have been used here. The dataset has the sentiment column where sentiment of each tweet is already given. This is used for the supervised learning of an algorithm. The data is cleaned and tokenized using natural language processing and python programming. The features like words and there frequencies, parts of speech tags, opinion words and phrases and negations are extracted. Then the dataset is splitted into training and testing sets. The training set is used for the training of the machine learning algorithm and the testing set is used for comparing the predicted value with expected value. The machine learning algorithm used here are naïve bayes, maximum entropy and support vector machine. The paper concludes with a comparative result between different algorithms in which support vector machine has got the highest accuracy of 86.40%.

4. Data Gathering:

To evaluate the better results the quality and quantity of Dataset plays the key role and become one of the reasons for the success of research. The dataset for the training and testing purpose which we have used here in the project is retrieved from the Kaggle. It contains a total of 16,00,000 tweets. The dataset is being split in two different sets one for the training and other for the testing purpose.

Attributes of dataset:

1.Target 2-Ids 3-Date

4-Flag 5-User 6-Tweet.

5. Pre-processing:

In pre-processing, we have used multiple techniques to remove the unwanted words and make all the words to deflect to their root form. The process of preparation of dataset for the analysis is known as pre-processing. This process is done in order to make the algorithm more efficient by discarding the words which does not affect the accuracy of the algorithm. This process is applied with the help of natural language processing commonly known as NLP. A library of python which deals with the textual analysis of the data. This process removes the duplicate words and make the length of the tweets shorter in order to minimize the training time and efficiency.

For applying this process, we have passed the tweets from the following process:

5.1. The text has been splitted into tokens i.e., into smaller fragments of words and stored in a data structure for further processing.

5.2. All the words from that data structure are then converted to the lower case for a better understanding by the algorithm.

5.3. Using the regular expression for the python library re, all the punctuations have been removed from the data structure containing words.

5.4. After the removal of the punctuations from the word's data structure, we have removed the non-alphabetic tokens like numbers or emojis as they will not contribute to the prediction and training.

5.5. Then further, we have only words in the data structure, so we have used stemmer in order to deviate the words to their base form. The data structure will be updated here.

5.6. At the end of the process, we have re-joined the words into sentences for the analysis purpose. The sentence will be formed in form of tweets as thy were original but in a clean way.

6. Model Training:

Model training is a process of making an algorithm learn what it must do further. A Machine learning model is train over a pre-defined dataset in which the output of the task was defined. The model training is performed by three types: -

i. Supervised Learning
ii. Unsupervised Learning
iii. Reinforcement Learning

Training a model under supervision with the subject and its target in the dataset is known as supervised learning algorithm. In our model we have used 9 supervised learning algorithm to train and test the model. The model is train through process which include the learning, error tuning, and then validation. The model first learns from the training dataset that which type of statement is negative, which are positive and rest neutral. It understands the negative and positive words and there affect with other words in a sentence.

The model which are used in the paper for the comparative analysis on accuracy and performance are as said below.

6.1. Logistic Regression: -

Logistic Regression is a classifier in the machine learning which is used to classify the data into distinct groups. This is a supervised machine learning algorithm. The logistic regressing predicts the dependent data using the dataset of independent variable which will be given to the algorithm. The logistic regression gives discrete value instead of binary output as 0 or 1. The discrete value given by the logistic regression is a probabilistic value of being something and lies between 0 and 1. The 0 and 1 are both the extreme value for the logistic regression so we use sigmoid function to map the value in this range. Sigmoid function is a function which is used by different machine learning algorithms. This function is basically used when we need are given a value from negative infinity to positive infinity and the machine requires the value from a range of 0 to 1. This function maps the value by changing it to a range of 0 to 1 for further process.

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + \ldots\ldots\ldots + b_nx_n$$

The above equation will give the coordinates to be plotted on a plane on which the classification is done. To categorize the output in different category a threshold value is considered for the classification of the data which is in between the 0-1 range. Above the threshold value, the predicted value will be considered as one else zero.

6.2. Naïve Bayes: -

Naïve Bayes is a supervised learning algorithm which is used for the classification of the data. As the name, Naïve Bayes uses the methodology of the popular bayes theorem which is used to find out the best hypothesis from a given space. Here the space is the dataset given to the algorithm. The bayes theorem states:

$$p\left(h/_d\right) = \frac{p\left(d/_h\right) * p(h)}{p(d)}$$

In the above equation, 'h' is the best suited hypothesis on a given 'd' dataset. The 'p(d/h)' figures out the probability of 'h' with respect to 'd.'

It uses probabilistic calculation for the prediction by quantifying the dataset into likelihood table of a hypothesis. It has three ways for the classification as listed below:

1. Gaussian Naïve Bayes
2. Multinomial Naïve Bayes
3. Bernoulli Naïve Bayes

In our model, we have used Gaussian Naïve Bayes algorithm for which we have used sci-kit-learn library of the python to implement it in the code. The Gaussian Naïve Bayes algorithm works on the stated mathematical equation:

$$p(x/y) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

6.3. Support Vector Machine: -

Support Vector Machine commonly known as SVM is a supervised learning algorithm used for the classification of data. SVM is also used for the regression problem, but it is best suited for the classification problem. SVM use extreme points i.e., known as vectors which create a hyperplane. These vectors are known as support vectors. Hyperplane is the best decision boundary on a n-dimensional plane. There are two types of support vector machine: -

1. Linear SVM
2. Non-linear SVM

The python library sklearn is used to implement the support vector machine. The sklearn is commonly known as the sci-kit-learn. In our project we have used multiple kernels to classify the data for a comparative analysis of different models. Kernel helps to gather the input data and transform the non-linear decision surface into a linear one in training dataset by using different mathematical equations. The kernel used for the model is listed below with the equations used:

6.3.1. Gaussian Kernel Radial Basis Function:

$$K(x,y) = e^{-}(\gamma\left|\left|x-y\right|\right|^2)$$

6.3.2. Sigmoid Kernel:

$$K(x,y) = tanh(\gamma.x^Ty + r)$$

6.3.3. Polynomial Kernel:

$$K(x,y) = tanh^d(\gamma.x^Ty + r)^d, \gamma > 0$$

6.3.4. Linear Kernel:

$$f(x) = B(0) + sum\big(ai * (x,y)\big)$$

6.4. Decision Tree Classifier: -

Decision Tree Algorithm is used to solve both category of the problem of machine learning both classification and regression. It is also a supervised machine learning algorithm. This algorithm creates a decision tree based on the training data having three components: Root, Decision Node and Leaf Node as output. After the creation of the decision tree the model uses that

tree to predict the data of the test set. Decision tree uses Attribute Selection Method for the choice of the best attribute from the data set to recursively generate a tree. This process repeats until it cannot make further tree. The ASM uses two famous techniques for choice of the attributes.

6.4.1. Information Gain:

It uses the quantity of change in the randomness in the data after the breaking of the data through segmentation based on the attribute. The mathematical expression used for the calculation of the Information Gain is

$$I.G. = E(S) - [(W) * E(f)]$$

Here in the equation E(S) is the Entropy of the sample, W is the Weighted average and E(f) is the Entropy of each feature. The E(S) is determined using a mathematical expression as shown below.

$$E(S) = -P(y)log_2P(y) - P(n)log_2P(n)$$

where P(y) is probability of being yes and as P(n) as probability of being no.

6.4.2. Gini Index:

It uses the quantity of being pure or impure during the creation of the tree. The lower the gini index the most preferable attribute it is. The gini index (G.I.) uses a mathematical expression for determining the G.I. as listed

$$G.I. = 1 - \sum jPj^2$$

The python library sklearn is used to implement the Decision Tree Algorithm. The sklearn is commonly known as the sci-kit-learn. Here we have set the criterion as entropy for best result.

6.5. Random Forest Algorithm: -

It is the most famous machine learning algorithm which comes under the category of the supervised learning algorithm used for both regression and classification model. The basic concept of random forest algorithm is ensemble learning. The ensemble learning can be defined as the process of using multiple classifiers for a single complex classification problem. It combines different algorithm to improve the performance of the machine learning model. The name includes random forest which means that this algorithm is going to use a group of decision tree on the random basis. The prediction is done based on the average of the prediction of all the decision tree. The python library sklearn is used to implement the random forest algorithm. The sklearn is commonly known as the sci-kit-learn. The number of random trees is passed as an argument int the RandomForestClassifier class as a value of variable n_estimator. The value of the n_estimator is ten which is default, but we can override it by taking care of the overfitting of the data in that model.

6.6. KNN Algorithm: -

The KNN stand for the K^{th} nearest neighbor which is categorized in the supervised learning algorithm and is used only for the classification problem. KNN at the training phase only store data and uses the data at the time of classification so it is called lazy learner algorithm. KNN does not make any assumption based on the underlying data hence it is called non-parametric algorithm. The dataset is stored and when the prediction dataset is run in the model, it iterates over each row of the data used in training set and check for the k nearest neighbor and according to that it predicts the output. To find the distance between two rows of the data, it uses Euclidean distance mathematical equation as stated

$$d = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$$

In the above stated equation d is the Euclidean distance X_1 and Y_1 are the training dataset attributes value and X_2 and Y_2 are the testing dataset attributes value. The KNN is also implemented by using the KNeighborsClassifier

class form the sklearn library. The number of neighbors is passed as the argument of the class

as the value of variable n_neighbour. We have used five neighbors for the classification of our model.

6.7. Gradient Boosting Machine: -

Gradient Boosting Machine also known as GBM uses a combination of multiple simple models to boost the performance of the model. The GBM takes the training dataset to make it fit in the model and then assign an equal weight to all the points of data. If the data point classified has an error, it reassigns the weight until it gets a suitable result. It is one of the best and powerful technique for both the classification and regression problem. GBM is also a supervised learning algorithm but depends on the average of the output of multiple simple algorithms. GBM consist of the listed element for the computation.

1. Loss Function
2. Weak Learner
3. Additive Model

It is an ensemble learning algorithm so implemented by the help of GradientBoostingClassifier from the sklearn. ensemble library.

7. Model Prediction:

Model Prediction is a technique using Machine Learning and Data Mining to predict and forecast future results. It analyzes current and past data and scrutinize the learning model to forecast the favorable outcomes. We have split the dataset into two parts on a ratio of 8:2 with the use of train_test_split class from the sklearn library. The 80% of the dataset belongs to the training set and the rest 20% belongs to the testing set. We train theclassifier with the training set and make prediction by using the testing set on the trained model using different classifiers.

Algorithms		**Positive**			**Negative**			**Neutral**		
		Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Support Vector Machine	Linear	0.92	0.91	0.87	0.87	0.84	0.89	0.91	0.87	0.89
	Polynomial	0.89	0.88	0.94	0.88	0.84	0.89	0.91	0.88	0.94
	Sigmoid	0.91	0.84	0.76	0.85	0.88	0.87	0.86	0.81	0.84
	RBF	0.93	0.87	0.87	0.88	0.84	0.91	0.88	0.87	0.83
0Decision Tree		0.88	0.99	0.93	0.87	0.81	0.84	0.90	0.83	0.86
Random Forest		0.90	0.86	0.88	0.83	0.69	0.75	0.83	0.96	0.89
KNN		0.92	0.25	0.40	0.74	0.18	0.28	0.42	0.99	0.59
Gradient Boost Machine		0.92	0.86	0.75	0.68	0.95	0.75	0.91	0.89	0.89
Logistic Regression		0.94	0.90	0.92	0.89	0.75	0.82	0.86	0.98	0.92

Table 1

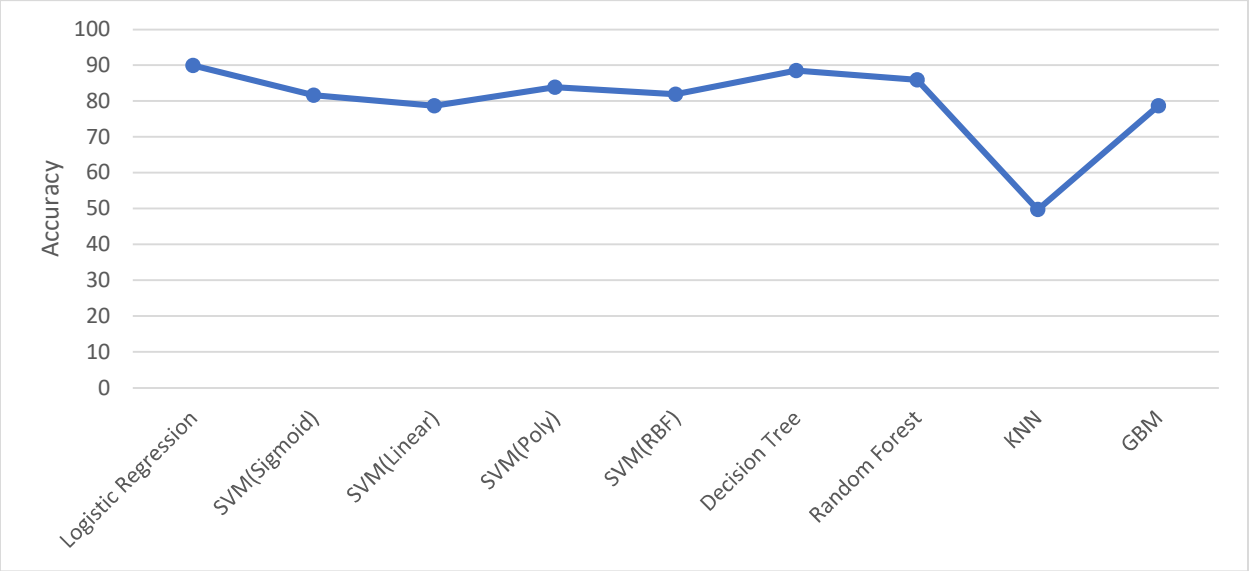

Figure 1

No.	Algorithms		Accuracy
1	Logistic Regression		89.94
2	Support Vector Machine	Linear	78.65
3		Polynomial	83.84
4		Sigmoid	81.59
5		RBF	81.86
6	Decision Tree		88.52
7	Random Forest		85.93
8	KNN		49.73
9	Gradient Boost Machine		78.68

Table 1

Table 1 shows the comparison of the accuracies of different classification algorithms used in the research.

Table 2 shows the precision, recall and f1-score of all the category i.e., negative, positive, and neutral on each classifier used for the comparison. This helps in the analysis of the difference in the prediction of sentiments by using logistical regression, naive bayes, support vector machine with different kernels (i.e., sigmoid, polynomial, radial basis function and linear kernel), decision tree, random forest, knn and gradient boosting machine.

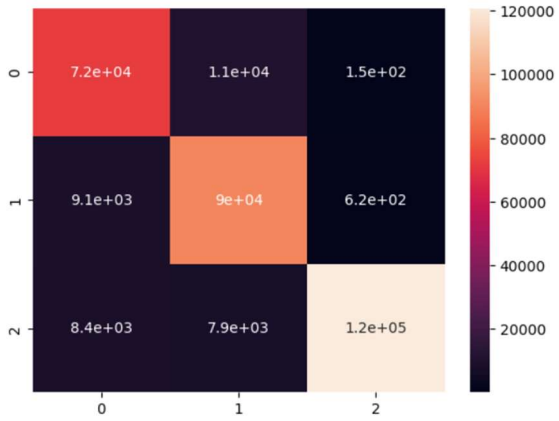

Figure 2

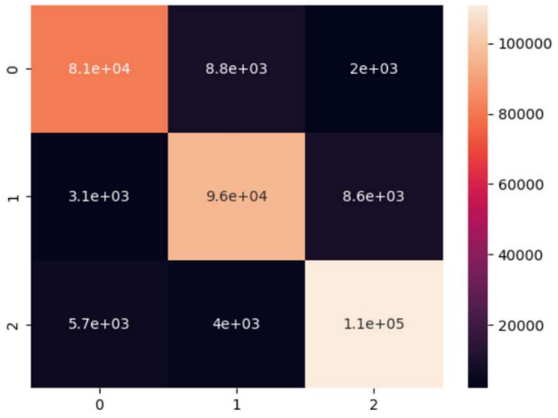

Figure 3

8. Experimental Result and Analysis:

After the prediction made by the

different machine learning model, we can analyze the performance of each of the classifiers used in the machine learning model. From the table 2 we can get a brief idea of the performance of the classifiers. Figure 2 shows the confusion matrix of the Logistic regression model for an analysis of the true prediction as target is labeled on the y-axis and the predicted value is labeled on the x-axis. The confusion matrix is used to get the performance of the algorithm. In Figure 2 the number of tweets whose sentiments are predicted correctly are placed in the diagonal of the matrix. The 0, 1 and 2 in the axis denotes the Negative, Neutral and Positive tweets classification. To compare between different algorithms, we have three parameters here precision, recall and f1-score. The mathematical formula for calculating each are as follows:

$$precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$f1 - score = 2 * \frac{precision * recall}{precision + recall}$$

So, from the confusion matrix we can analyze that the sum of correct predictions is 287808 (72214 negative, 90262 neutral and 120788 positive) out of 320000 tweets. This gives an accuracy of 89.94% which is highest, among others. The other parameters are also derived from the confusion matrix.

Similarly, the Figure 3 shows the confusion matrix of the Decision Tree algorithm in which the correctly predicted tweets are 283264 out of 320000 tweets with an accuracy of 88.52%. After the overall analysis of the different model, we have analyzed that the best model which fits to our dataset for the classification is Logistic Regression. The runner up for the dataset is the Decision Tree Algorithm. This can change for the different dataset or different problem because the machine learning is all about the prediction of the data. Figure 4 shows a plot of the accuracy of different algorithms on a linear graph for a visualized idea of the performance of each classifier.

9. Future Scope:

Tweeter is a famous platform for opinion sharing about any topic, product or anything going on in mass. So, to get an idea about anything going in the public people uses twitter and read the opinion about that topic. In recent years, most of the public uses twitter for sharing their review and opinion over any protest going on, product launched by any company or any public problem so that their opinion can reach in mass. The analysis we have done here classifies some of the best algorithms for sentiment analysis of any textual opinion.

The model trained here can be used for the analysis of the sentiment of the public or mass about any product in commercial field so that the company can extract the negative opinion among all the opinions and enhance the product. This will give a boost in the marketing of the product and helps the company to find out the negativity in the product. In commercial area, when a company wants to introduce any new product, he can analyze the feedback of the similar past product to take a brief idea of the opinion of the public.

Similarly, in political context by designing a UI a person can get the opinion about any one or any party to analyze their image in public. One drawback of our algorithm is that we have used classifiers for the analysis of sentiments. The improvement can be done by using reinforcement learning algorithms by implementing TensorFlow library from python. TensorFlow uses perceptron model to predict the result giving a better accuracy with a less training time needed.

References:

[1] Prajval Sudhir, Varun Deshakulkarni Suresh, "*Comparative Study of Various Approaches, Application and Classifiers for Sentiment Analysis*", in *Global Transitions Proceedings,* 2021.

[2] Ashwin Sanjay Neogi, Kirti Anilkumar Garg, Ram Krishn Mishra, Yogesh K Dwivedi, "*Sentiment Analysis and Classification of Indian Protest Using Twitter Data*", in International *Journal of Information Management Data Insights,* 2021.

[3] Anupama B S, Rakshith D B, Rahul Kumar M, Navaneeth M, "*Real Time Twitter Sentiment Analysis using Natural Language Processing*", in *International Journal of Engineering Research and Technology,* Vol.09, Isuue 07, ISSN:2278-0181, 2020.

[4] Saurabh Singh, "*Twitter Sentiments Analysis Using Machine Learning*", in *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, ISSN:2456-3307, 2020.

[5] Faizan, "*Twitter Sentiment Analysis*", in *International Journal of Innovative Science and Research Technology*, ISSN No:2456-2165, 2019.

[6] Vishal Jain, Mahesh Parmar, "*A Review on Emotion and Sentiment analysis Using Learning Techniques*", in *International Journal for Research in Applied Science & Engineering Technology*, Vol 10, ISSN: 2321-9653, 2022.

[7] Ayushi Mitra, "*Sentiment Analysis Using Machine Learning Approaches (Lexicon based on movie review dataset)*", in *Journal of Ubiquitious Computing and Communication Technology*, Vol.02/ No.03, pp 145-152, 2020.

[8] Ruijun Liu, Yuqian Shi, Changjiang Ji, Ming Jia, "*A Survey of Sentiment Analysis Based on Transfer Learning*", supported by *National Natural Science Foundation of China*, 2019.

[9] Laszlo Nemes, Attila Kiss, "*Social media sentiment analysis based on COVID-19*", in *Journal of Information and Telecommunication*, DOI: 10.1080/24751839.2020.

[10] Huyen Trang Phan, Van Cuong Tran, Ngoc Thanh Nguyen, Dosam Hwang, "*Improving the Performance of Sentiment Analysis of Tweets Containing Fuzzy Sentiment Using the Feature Ensemble Model*", supported by *Basic Science Research Program* through *National Research Foundation of Korea*, 2020.

[11] Mickel Hoang, Oskar Alija Bihorac, Jacobo Rouces, "*Aspect-Based Sentiment Analysis Using BERT*".

[12] Ashima Yadav, Dinesh Kumar Vishwakarma "*Sentiment analysis using deep learning architectures: a review*", Springer Nature B.V. 2019.

[13] Abdullah Alsaeedi, Mohammad Zubair Khan, "*A Study on Sentiment Analysis Techniques of Twitter Data*", in *International Journal of Advanced Computer Science and Application*, Vol.10/ No.02, 2019.

[14] A.M. Johm-Otumu, M. M. Rahman, O. C. Nwokonkwo, M. C. Onuoha "*AI-Based Techniques for Online Social Media Network Sentiment Analysis: A Methodical Review*", in *International Journal of Computer and Information Engineering*, Vol.16, No.12, 2022.

[15] Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, Feng Wu, "*SKEP: Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis*", in *arXiv:2005.05635v2*, 2020.

[16] Samira Zad, Maryam Heidari, James H Jr Jones, Ozlem Uzuner, "*A Survey on Concept-Level Sentiment Analysis Techniques of Textual Data*", in *IEEE World AI IoT Congress*, 2021.

[17] Kamaran H. Manguri, Rebaz N. Ramadhan, Pshko R. Mohammed Amin, "*Twitter Sentiment Analysis on Worldwide Covid-19 Outbreaks*", in *Kurdistan Journal of Applied Research (KJAR)*, ISSN:2411-7706 2020.

[18] Koena Ronny Mabokela, Turgay Celik, Mpho Raborife "*Multilingual Sentiment Analysis for Under-Resourced Languages: A Systematic Review of the Landscape*", supported by *National Research Foundation for the Black Academics Advancement Programme.*

[19] Li Yang, Ying Li, Jin Wang, R. Simon Sherratt, "*Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning*", supported in part by the *National Natural Science Foundation of China*, 2020.

[20] Vishal A Kharde, S.S. Sonawane, "*Sentiment Analysis of Twitter Data: A Survey of Techniques*", in *International Journal of Computer Application, 2016*.