ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 9, Issue 11, November 2022

Sentiment Analysis of Covid-19 Tweets using Various Machine Learning Techniques with Hyperparameter Tuning on Twitter Database

[1] Tejaswini Zope, [2] Dr. K .Rajeswari

[1] [2] Department of Computer Engineering Pimpri Chinchwad college of Engineering, Akurdi, Pune, India.
Corresponding Author Email: [1] tejaswini.zope20@pccoepune.org, [2] Kannan.rajeswari@pccoepune.org

Abstract— People around the world are now being affected by the 2019 coronavirus disease (COVID-19) epidemic. People are using social media to express their opinions and general thoughts about the epidemic, which has affected their daily lives both in general and during lockdown periods. Twitter, one of the most used social media platforms, has seen a massive surge in tweets related to the coronavirus, including happy, sad, fear, and anger tweets, in a very short period. In the COVID-19 tweet dataset, there is a total number of tweets count is 179108 data in unstructured form, and after that preprocess that data to become semi structured form. After preprocessing feature extraction is used by applying CountVecotorizer and TF-IDF methods. Various Machine Learning Models such as Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), and Stochastic Gradient Decent (SGD) are considered machine learning methods for sentiment analysis, To improve performance (accuracy), with Hyperparameters tuning method is used.

Index Terms— Sentiment Analysis, Machine Learning, COVID-19 tweets, Twitte, emotions.

I. INTRODUCTION

Recent advances in artificial intelligence have greatly increased the importance of automatic emotion recognition, pattern recognition, and computer vision with applications in a variety of fields. Social networking sites like Twitter have recently produced vast amounts of organized, unstructured, and semi-structured data. One of the most recent examples is COVID-19 tweets, which shows that information spread via social media can be significantly natural event like a pandemic.

The continuing COVID-19 coronavirus pandemic is a major source of worry for people all around the world. spreading misinformation through social media channels like Twitter. Since the COVID19 increasing levels, social media has quickly developed into an essential instrument for communication in the generation, dissemination, and consumption of information. The sentiments can be categories in positive and negative, with other subcategories of emotions including happy, sad, angry, and fear.

In this paper we used different Machine Learning technique such as Random Forest (RF), Decision Tree(DT), Support Vector Machine (SVM), and Stochastic Gradient Decent (SGD) to predict the sentiment analysis in form of different emotions in various form like happy, sad, fear, and anger. Also used the different Machine Learning method to improve the performance of ML model using Hyperparameter Tunnig with GridsearchCv.

Ensemble technique is used to utilized with Machine Learning Classififers such as Support vector classifier and random forest classifier to measure the performance in the form of accuracy with TF-IDF. A machine learning method called ensemble classification combines the basic classifiers based on several methods to get a single best model.

II. LITERATURE SURVEY

ANAM YOUSAF et al. ML Algorithms are used Support Vector Machine, Random Forest, Logistic Regression, Decision Tree, Naive Bayes, Stochastic Gradient Descent and Proposed Voting Classifiers LR-SGD with TF-IDF Produces the most optimal result with 79% accuracy and 81% F1 score[1].

Chaudhary Jagrit Varshney et al. studied the different Machine Learning model to improve the accuracy the got highest accuracy 80.29% by uusing Logistic Regression, Naive Bayes and SGD Classifiers gives the accuracy 77%, and Ensemble (voting Classifiers) gives the accuracy 79%[2].

Three different algorithm is used to find out the best accuracy.Machine Learning Algorithms are SVM, RF, Decision Tree. Random forest gives better accuracy of 97% for the dataset. Feature extraction technique combine with machine Learning algorithm are used to produce the better accuracy[3].

There are several types of classification techniques for data mining. The author in [5] has used different classifiers such as J48, Multilayer Perceptron, NaiveBayesUpdatable, and BayesNet have compared the accuracy of soybean and weather datasets. [5]

Data mining is analyzing data from several angles and condensing it into crucial information to find hidden patterns in a group of information. The author's [6] research has

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 9, Issue 11, November 2022

discussed assessing students' performance based on various properties and applying the association rule.

Garima et.al [7] extracted tweets from the Twitter dataset by applying Feature Extraction Technique Bag of Word and TF-IDF. Here the authors classified whether speech is hateful or not by Logistic Regression an obtain 94.11% Accuracy [7].

Using a hidden representation learning approach on graphs of the textual contents at the sentence-level, [8]. built a framework for in-depth sentiment analysis. The results shown that textual sentiment analysis may be successfully improved by deep learning and graph-based representation.

[9] examined a pre-processing method known as "stop words" using an English data set with 1,000 positive samples and 1,000 negative samples. The SVM method's accuracy is 94.06 percent, compared to the NB approach's 89.53 percent. This method can only be used in English[9], though.

Evaluations are done on the two datasets of tweets and goods. Semi-supervised sarcasm recognition was tested on product dataset[10]. The recall and F-score accuracy of the proposed method are both favourable.

III. PROPOSED METHODOLOGY

In this paper, various methodologies of machine learning have been employed in this work to achieve its aims. Various approaches and procedures were used to assess various experiments. Multiple classifiers were applied on the dataset, however the classifier that used an ensemble of SVM and RF provided the best accuracy.

This experiment uses Twitter data that was taken from the Kaggle repository. The dataset is first pre-processed by discarding unnecessary data. The training set is used to apply the feature engineering approaches. Various ML Classifiers are trained on training dataset, to improve the performance.

A. Dataset

Covid-19 Twitter dataset extracted from kaggle website. In this Twitter dataset total 17809 number of tweets are present. Then classify dataset tweets in different emotions like happy, joy, fear, sad, and anger.

B. Data Preprocessing

Once the data has been gathered and stored in.csv files, it is crucial to preprocess the data and eliminate any unnecessary data. There are several preprocessing procedures required, including the following:

1. Remove Stop words:

Stop words are the parts of a sentence that are not necessary for any text mining segment; thus we often eliminate these phrases to improve analysis efficiency. The forms of stop words vary in different languages and countries. A list of stop words is kept in NLTK (Natural Language Toolkit) in Python and is available in 16 languages.

For Eg: (br, the, me, my, myself, it's, being, she's,her…..etc.)

2. Remove Punctuation:

Removing punctuation is a frequent preprocessing step in many activities involving machine learning and data analysis. Sometimes we need to do certain cleaning procedures while working with textual data. Typically, one of these steps is the punctuation elimination, which might occur before tokenization.

For Eg: (. , " " : ! ? ' - _( ) {} [ ] etc)

3. Remove URL's:

An online resource's URL serves as its connection. All resources have their own unique URLs, although they all have the same basic layout. Every text will include a unique URL, and a specific text could also contain one, therefore we must first determine the URL out of its structure and get rid of it.

For eg: Text1 http://url.com/

After removing URL results is Text1

C. Feature Extraction

A dimensionality reduction technique called feature extraction divides a large amount of raw data into smaller, easier-to-process groupings. These huge data sets have the trait of having many variables that demand a lot of computational power to process. The term feature extraction refers to techniques for choosing combining variables into features, which significantly reduces the quantity of data that has to be processed while properly and fully characterizing the initial data set.

In machine learning algorithm cannot work on directly on row data so, it need to extract features, methods of feature extraction are:

1. CountVectorizer

A fantastic utility offered by the Python scikit-learn module is CountVectorizer. It is used to convert a given text into a vector based on the number of times that each word appears across the full text. When we have several of these texts and want to turn each word into a vector, this is useful.

Suppose we taken the Text1 and Text2 from dataset. Each unique word is represented by a row in the matrix that is created by CountVectorizer, and each sample of text from the document is represented by a column in the matrix. Each cell's value is just the number of words in that specific text sample. As an example, consider the following:

Text1: coronavirus disappearing in Italy show this to intellectuals who say lockdowns do not work.

Text2: UK records lowest daily due to coronavirus death toll since start of lockdown govt.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 9, Issue 11, November 2022

Table 3.1 CountVectorizer Matrix

Unique Word	Text1	Text2
Coronavirus	1	1
Disappearing	1	0
Italy	1	0
UK	0	1
Lockdown	1	1
Daily	0	1
Death	0	1

The above table 3.1 shows CountVectorizer Matrix where Unique words is represented by rows and sample Text1 and Text2 represented as column.

2. Term Frequency – Inverse Document Frequency(TF-IDF)

The goal of feature extraction is to turn a text document in any format into a list of characteristics that text classification algorithms can quickly process. One important preprocessing method used in data mining and text classification to determine the value of features in documents is feature extraction. Term weighting often makes use of effective feature extraction approaches like term frequency-inverse document frequency (TF-IDF) techniques.

The frequency (TF) of a term or word is the ratio of the number of times the term appears to the overall number of words in a document.

$$Tf = (\frac{Number\ of\ time\ term\ t\ appears\ in\ a\ documnet}{total\ muner\ of\ terms\ in\ a\ document}) \ldots\ldots(1)$$

The IDF of a phrase indicates how frequently the term appears in corpus texts. Higher relevance ratings are assigned to terms that are particular to a limited subset of texts than to words that appear in all documents.

$$IDF = \log(\frac{\text{number of the document in the corpus}}{\text{numbet of the documentin corpus cantain the terms}}) \ldots\ldots(2)$$

A term's TF-IDF is determined by multiplying the TF and IDF values.

$$TF{-}IDF{=}TF{*}IDF \quad \ldots\ldots..(3)$$

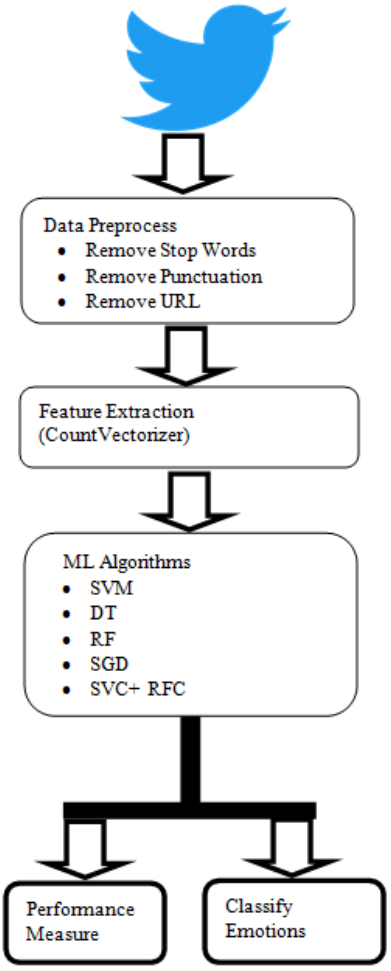

Fig.3.1 Flow of Proposed Sentiment Analysis using ML Technique

D. Machine Learning Algorithm

1. Support Vector Machine(SVM)

It is believed that the Support Vector Machine (SVM) accurately performs sentiment analysis [11]. SVM exemplifies preference, limits and employs assessment processes, and analyses data obtained inside the index area [12]. Important information are embodied in the vector arrangements for each intensity. To accomplish this goal, information has been grouped in type (presented as a vector). The boundary is then classified by strategy in two training sets.From any location in the training samples, this is far away [13]. Support-vector machines techniques used in machine learning that analyse data used for categorization as well as reverse inspection through the use of targeted learning models linked to learning evaluations [14].

2. Random Forest

Random Forest (RF) is a tree-based classifier that uses a randomly generated input vector to create trees. In order to build numerous decision trees and a forest, RF leverages

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 9, Issue 11, November 2022

random characteristics. After that, test data class labels are predicted using the combined vote of all trees. Higher weights are given to decision trees with low-value errors. By taking into account trees with low error rates, overall prediction accuracy is increased.

3. Decision Tree

The DT algorithm is the category of supervised ML and is commonly used in regression and classification tasks. Choosing the root node of a tree of each level is the main challenge, called attribute selection. Gini Index and Information Gain are the most commonly used methods for selecting attributes. In this study, the Gini index is used to find the probability of the root node by calculating the sum of the squares of the attribute values and then subtracting by 1.

Selection of root node of a tree of each level is main challenge which is called as attribute selection

Calculate the Gini Index formula is given:

$G = \sum_{i=1}^{C} p(i) * (1 - p(i))$ .....(4)

Where as C= total number of class

P(i)= probability of data point with class

4. Stochastic Gradie nt Descent (SGD)

A quick and effective method for fitting linear classifiers and regressors under convex loss functions, such Support Vector Machines and Logistic Regression, is stochastic gradient descent (SGD). SGD has been present in the machine learning field for a while, but in the context of large-scale learning, it has just lately attracted a lot of interest.

E. Ensemble Machine Learning Model (SVC+RF)

To quantify performance in the form of accuracy using TF-IDF, ensemble approach is employed with Machine Learning Classifiers like Support Vector Classifier and Random Forest Classifier. Ensemble classification is a machine learning technique that combines the fundamental classifiers based on many techniques to get the single best model.

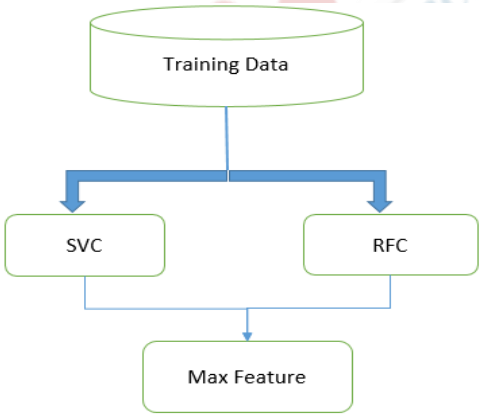

Fig. 3.2 Ensemble of SVC+ RFC

1. Support Vector Classifier (SVC)

SVC is a pattern-based classifier that maps input vectors xi into a high dimensional vector space using a linear kernel function [25]. Then, using the optimum margin between the target classes yi, it builds a linear hyper-plane. The following mathematical procedure is integrated for pattern recognition by the linear kernel function.

$K(xi, yi) = xi'yi$..... (5)

2. Random Forest Classifiers(RFC)

Machine learning algorithms known as random forests are utilised for classification and regression problems. Using input data, a classifier model categorises the information. An ensemble of different decision trees may be thought of as random forest. The goal is to combine the predictions from various decision trees to provide a final result based on the Max feature.

3. Max Feature

The most attributes that Random Forest can test in a single tree is as many as these. The model performs better overall when max features is increased since there are more alternatives to consider at each node.

IV. RESULTS

Accuracy evaluates the accuracy of a forecast and is expressed as:

Accuracy=$\frac{Total\ Num\ of\ Corrected\ prediction}{All\ Prediction}$.....(6)

while for binary classication, accuracy is calculated as:

Accuracy=$\frac{TP+TN}{TP+TN+FP+FN}$ ..... (7)

Table 4.1 Performance Measure

ML Algorithm	Accuracy
Random Forest	0.99%
Decision Tree	0.96%
SVM	0.26%
SGD	0.99%
Ensemble SVC+RFC	0.97%

The above table 4.1 represent the measurement of performance.

V. CONCLUSION

In this paper different ML algorithm are utilized to predict sentiments in the form happy, sad, fear, and anger. Measure the performance of ML Algorithm. To Iimprove the performance of ML model, ensemble technique is used Hyperparameter tuning using GridserachCv. Random Forest and SGD gives the Highest Accuracy that is 0.99%, where as decision tree gives the accuracy 0.96% , and SVM gives less accuracy 0.26% as compared to other ML Algorithm. The

main reason behind the ensemble the ML Algorithm is SVM alone give less accuracy, this paper work on ensemble technique (SVC+RFC) to give better performance.

REFERENCES

[1] ANAM YOUSAF,MUHAMMAD UMER,SAIMA SADIQ " Emotion Recognition by Textual Tweets Classification Using Voting Classifier (LR-SGD)" Digital Object Identifier 10.1109/ACCESS.2020.3047831

[2] Chaudhary Jagrit Varshney,Dr. Ashish Sharma, Dhirendra Prasad Yadav" Sentiment Analysis using Ensemble Classification Technique" 2020 IEEE Students' Conference on Engineering & Systems (SCES) July 10-12, 2020

[3] Nilufa Yeasmin, Nosin Ibna Mahbub, Mrinal Kanti Baowaly, Bikash Chandra Singh" Analysis and Prediction of User Sentiment on COVID-19 Pandemic Using Tweets" Big Data Cogn. Comput. 2022, 6, 65. https://doi.org/10.3390/ bdcc6020065

[4] Babacar Gaye, Dezheng Zhang and Aziguli Wulamu "Sentiment classification for employees reviews using regression vectorstochastic gradient descent classifier (RVSGDC)" DOI 10.7717/peerj-cs.712

[5] V. Vaithiyanathan, K. Rajeswari, Kapil Tajane, Rahul Pitale, "COMPARISON OF DIFFERENT CLASSIFICATION TECHNIQUES USING DIFFERENT DATASETS",in 2003 May IJEAT, Vol. 6, Issue 2, pp. 764-768

[6] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt and B. Scholkopf, "Support vector machines," IEEE Intell. Syst. Appl., vol. 13, no. 4, pp. 18-28, July- Aug. 1998, DOI: 10.1109/5254.708428.

[7] Garima Koushik, Dr. K. Rajeswari, Mr. Suresh Kannan Muthusamy, "Automated Hate Speech Detection on Twitter", in 2019 IEEE conference, 978-1-7281-4042-1/19/$31.00

[8] Bijari, Kayvan et al. "Leveraging Deep Graph-Based Text Representation for Sentiment Polarity Applications." Expert Systems with Applications 144 (2020): 113090. Crossref. Web.

[9] Tripathy, Abinash, A. Agrawal, and S. K. Rath. "Classification of Sentimental Reviews Using Machine Learning Techniques." Procedia Computer Science 57(2015):821-829.

[10] Davidov, Dmitry, O. Tsur, and A. Rappoport. "Semi-supervised recognition of sarcastic sentences in Twitter and Amazon." Conll(2010):107-116

[11] P. Routray, C. K. Swain, and S. P. Mishra, ``A survey on sentiment analysis," Int. J. Comput. Appl., vol. 76, no. 10, pp. 18, Aug. 2013.

[12] A. Harb, M. Plantié, G. Dray, M. Roche, F. Trousset, and P. Poncelet, ``Web opinion mining: How to extract opinions from blogs?" in Proc. 5th Int.Conf. Soft Comput. Transdisciplinary Sci. Technol. (CSTST). New York,NY, USA: Association for Computing Machinery, 2008. pp. 211217.

[13] B. Pang, L. Lee, and S. Vaithyanathan, ``Thumbs up? Sentiment classification using machine learning techniques," EMNLP, vol. 10, pp. 1-9,Jun. 2002.

[14] K. P. Bennett and C. Campbell, ``Support vector machines: hype or hallelujah?"Acm Sigkdd Explor. Newslett., vol. 2, no. 2, pp. 1-13, 2000.

[15] Rezaeinia SM, Rahmani R, Ghodsi A, Veisi H. 2019. Sentiment analysis based on improved pre-trained word embeddings. Expert Systems with Applications 117:139_147 DOI 10.1016/j.eswa.2018.08.044.