Heterogeneous Ensemble Learning Framework for Sentiment Analysis on COVID-19 Tweets

RUBUL KUMAR BANIA

Department of Computer Application
North-Eastern Hill University, Tura Campus
Meghalaya-794002, India
rubul.bania@gmail.com/rubul.bania@nehu.ac.in

Abstract. During catastrophe, detecting tweets associated to the target incident is an exigent task. Sentiment analysis is one kind of the study of sentiments shared by diverse users in social networking sites like, Twitter, Facebook, etc., on various social phenomena. In this article, analysis of sentiments on thousands of tweets collected for the period of July to August 2020 and May 2021 to June 2021 on the ongoing pandemic of COVID-19 is carried out. By adopting the majority voting idea one novel ensemble learning model is proposed to classify the tweets into *negative*, *neutral*, and *positive* groups. Data pre-processing, polarity and other various analysis techniques are applied on the COVID-19 related tweets. By applying TF-IDF with unigram and bigram approaches text features are extracted and five machine learning models such as Naïve Bayes (NB), logistic regression (LR), K nearest neighbour (KNN), decision tree (DT) and random forest (RF) are judiciously combined to build an ensemble learning model. Experimental results suggest that on both the feature extraction approaches i.e., on unigram and bigram, proposed model has performed better than the other compared models. With 70%–30% train-test set, proposed model is able to achieve an accuracy of 94.94% to classify the tweets into various categories.

Keywords: COVID-19, Ensemble learning, Sentiment analysis, Twitter.

(Received September 30, 2021 / Accepted November 20, 2021)

1 Introduction

Everyone is aware of the coronavirus disease of 2019 (COVID-19), the pandemic is taking a toll on the world's health care infrastructure as well as the social, economic, and psychological well-being of humanity. COVID-19 was declared as pandemic by World Health Organization (WHO) [3] on 13 March, 2020. As per WHO (https://covid19.who.int/), till 29 September 2021 there have been at least 233,592,026 confirmed infected cases and more than 4,779,704 confirmed death cases in the COVID-19 pandemic. As of 29 September 2021, a total of 5,924,819,985 vaccine doses have been administered. Due to such catastrophe situation different people have gone through several mental trauma, and leads to depression [17]. During this pandemic circumstance, the government is forced to framed new policies such as stay at home & stay safe, work from home, social distancing, restrictions on movement of individuals, wearing of mask as mandatory etc. Therefore, the number of active users on internet has increased dramatically. People are engaging themselves most of their time in social media platform such as Facebook, Twitter etc. But it would be better if via the social media useful information get reach to the people, as people share their thoughts and share different tweets or posts [19]. In the domain of artificial intelligence (AI) and natural language processing (NLP), sentiment analysis or opinion mining is used to detect the opinion or mood of different sentences which are in the form texts as *negative*, *neutral*, and *positive* [4, 20, 8].

However, the attitude and feelings comprise an essential part in evaluating the behavior of an individidual that is known as sentiments. These sentiments can further be analyzed towards an entity, known as sentiment analysis or opinion mining. By using sentimental analysis techniques, we can interpret the sentiments or emotions of others and classify them into different categories such as *negative*, *neutral*, and *positive* which may help an organization to know people's emotions and act accordingly [4, 20, 8]. However, this analysis depends on its expected outcomes, e.g., analyzing the text depending on its polarity and emotions. By incorporating the recent advance computational techniques from the domain of Natural language processing (NLP), and machine learning different researchers have developed computational models which can analyzed the sentiments towards the COVID-19 pandemic tweets [12].

In the machine learning and statistical analysis field of research, ensemble learning is a very powerful and robust tool [6]. This technique involves by merging of decision of multiple experts or classifiers; as it is based on the proverb "two brains are better than one". Ensemble models for supervised learning has two essential steps; the first step involves for selecting a set of different learning models known as the base models and in the second step is to judicious combination of the results. Moreover, in the literature there are two basic approaches available viz., homogeneous and heterogeneous ensemble learning model [18, 6]. In homogenous technique, the members of the base models have a single-type base learning algorithm. Some popular methods such as bagging and boosting generate diversity by sampling from the training sets but utilize a single type of base classifier to build the final ensemble model. On the other hand, heterogeneous ensemble consists of members having different base learning algorithms. Thus, this technique use different fine-tunes classification algorithms. Through using the ensemble technique, the risk of over-fitting can be reduced.

In this article, initially by leveraging a set of tools (Twitter's search application programming interface (API)), Tweepy Python library and by using a set of hashtags thousands of English language tweets are extracted. From the Kaggle dataset repository for the period of 25/07/2020 to 29/08/2020, one dataset is prepared. Then tweets for the month of May and June 2021 are also scarped and another dataset is also prepared. Thereafter, one novel heterogeneous majority voting-based ensemble classification model is proposed to classify COVID-19 tweets for the two datasets. The underlying concept based on the informative features extracted by the TF-IDF model and five base classifiers namely Naïve bayes, logistic regression, K nearest neighbour, decision tree and random forest.

The remainder of this paper is organized as follows. Section 2 describes the overall structure of the proposed methodology. The experimental setup and evaluations are highlighted in Section 3. Experimental result analysis are presented in Section 4. Thereafter, discussion and summarization of the study is highlighted in Section 5. The paper is ended with concluding remarks with future perspectives in Section 6.

2 Literature review

Study and analysis of the sentiments/emotions of individuals on Twitter or other social media are being a very useful tool to prevent or measure some health care issues as well as to prevent the circulation of miss leading information by the government agencies. Also, this kind of studies may help in predicting the outbreak and its early detection too. One recent study has focused on the psychological effect of COVID-19 on human behavior [17]. It has reported that people are tense and their depression level increased due to COVID-19 news. Also authors have highlighted several measures which might minimize the emotional impact derived from such situation. In a very recent work Kaur *et al.* have performed recurrent neural network (RNN) and support vector machine(SVM) based sentiment analysis, which classifies the tweets based on their sentiment values. Tweets for a sample of 20, 50 and 250 tweets are considered during their experiment. Alrazaq *et al.* (2020) [5] have analyzed English language tweets from February 2, 2020, to March 15, 2020. They have analyzed the collected tweets using word frequencies of single (unigrams) and double words (bigrams) models. The analysis identified 12 topics, which were grouped into four main themes: origin of the virus; its sources; its impact on people, countries, and the economy; and ways of mitigating the risk of infection. In another very recent work, public sentiment associated with the progress of COVID-19, Aslam *et al.* [7] have performed a studies on sentiments and emotions evoked by different news headlines on coronavirus disease. Authors have reported that 52% of 141,208 news headlines updated till June 3, 2020, in various platforms have negative sentiments. Jelodar *et al.* [12] have applied an automated extraction of COVID-19 related discussions from reddit. Authors have developed an unsupervised topic model, with collaborative of LSTM and recurrent neural network (RNN) model to analyse COVID-19 related reddit comments for January 20, 2020 and March 19, 2020. A total of 451,554 comments were analyzed in that study. In another work [16] related to classification of Twitter data into three categories viz., *positive* and *negative* authors have developed a RNN model. Different emotional nature of several hundreds of tweets for the time frame of April 2020 and May 2020 were examined in their study. Some other related works can be found in [22, 9].

2.1 Contributions of the present work

After reviewing the related literature study, motivated to answers the following research questions (RQ).
RQ1: Are there any differences of sentiments in the tweets of *negative*, *neutral* and *positive* classes between two different time periods?
RQ2: What are the most popular words that exist in the mainstream of the tweets?
RQ3: How do these tweets may have an effect on the health care service providers?
RQ4: How does the machine learning models may help to analyze the COVID-19 related text sentiments?
To answers the above highlighted research questions the key contributions of this research are as follows:

(i) By measuring the score of the polarity tweets of the two separate time frames are categorized into three classes. (ii) Proposed a novel majority voting-based heterogeneous ensemble classification model to identify COVID-19 tweets during the pandemic period. The underlying concept based on the informative features extracted by the TF-IDF model and five base classifiers namely Naïve bayes, logistic regression, K nearest neighbour, decision tree and random forest. (iii) Extensive experiments are carried out using informative features to identify the performance of the proposed model.

3 Methodology

This section of this article sums up the methodology that has been applied and implemented in this work. The pipeline of the methodology, which is followed is shown as a block diagram in Figure 1. It is divided in to three processes viz., data collection, preprocessing and data analysis. The Natural Language Processing Toolkit (NLTK) [1], which is a python based platform is extensively used. The data analysis phase is derived in two parts. The first part of the analytic includes sentiment analysis of the textual component of the Twitter data. Tweets are assigned sentiment scores of polarity by using different methods and Python libraries such as TextBlob. The second part of analytic includes the feature extraction using TF-IDF technique for further prediction with the proposed ensemble classification model.

3.1 Data Collection

In this work, two sets of datasets of two different time frames are prepared. First dataset (Dataset-I) of the year 2020 is collected from Kaggle and second dataset (Dataset-II) was prepared by scraping Twitter platform by using the Twitter standard search API in Python. Most of the Twitter feature is open source to developers in Twitter API. The API utilizes "OAuth", an open authentication mechanism that is commonly used, to authenticate a request. Some predefined key terms such as #Covid19, #Corona, #Coronavirus, #Coronavaccine, consisting of a set of the most widely used scientific and news media terms relating to the novel corona virus are used. Tweets are extracted and stored in two different comma separated version (CSV) files.

It is worthy to mention here that the total number of unprocessed tweets in Dataset-I is 179104 and in Dataset-II is 70334. After performing the data preprocessing task the number of tweets get reduced. In Figure 2 samples of the Dataset-I and Dataset-II are shown.

3.2 Data preprocessing

After collecting the data, the next step is to preprocess the data. This is an important phase in text processing as the data which are collected from Twitter are not clean, which means it contains lots of special characters, urls, hashtags and unnecessary symbols which doesn't contribute for the analysis purpose. Pre-processing refers to the transformations applied to the data before feeding it to the learning algorithms.

Before applying the data preprocessing task, non-English tweets which are presents in the tweets are removed with the help of language field. In Figure 3, wordcloud representations of the two datasets before applying the data preprocessing tasks are shown. It is obvious from the figures that data preprocessing steps are very essential. Some unnecessary text and symbols such as username-tags (e.g., @Skahn), RT symbols, hashtags, URLs, and punctuations are present in the data. By applying proper regular expression (*re* module) texts are cleaned. Moreover, the punctuation are removed with a python constant. Next, the text data needs to be converted into lower case i.e., the normalization process. For instance, "Vaccine" and "vaccine" are seen as two different words by the program. Therefore, it is important to normalize the case of the words so that every word is in the same case and the algorithm doesn't process the same word as two different tokens. Thereafter, stop word removal process is applied. Stop words are those words that do not contribute

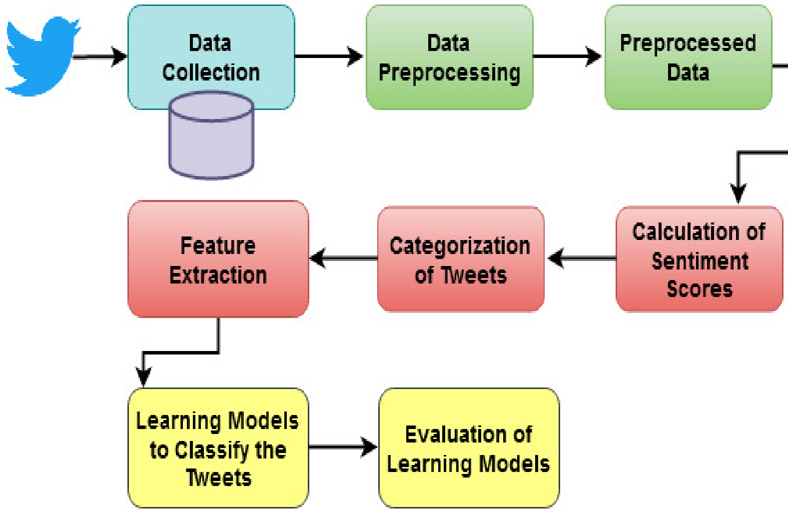


Figure 1: Block diagram of the proposed methodology.

to the deeper meaning of the phrase. They are the most common words such as: “the", “a", and “is". In the English language, there are words that are used more frequently than other words in the language but they do not necessarily add more value to a sentence, hence it is safe ignore them by removing it from tweeter text. After cleaning all the texts, duplicate types of tweets are also removed from the datasets. Thus, after applying all the necessary data preprocessing task the total number of tweets in Dataset-I become 142334 and in Dataset-II is 65668.

	user_location	date	text
0	astroworld	2020-07-25 12:27:21	If I smelled the scent of hand sanitizers toda...
1	New York, NY	2020-07-25 12:27:17	Hey @Yankees @YankeesPR and @MLB - wouldn't it...
2	Pewee Valley, KY	2020-07-25 12:27:14	@diane3443 @wdunlap @realDonaldTrump Trump nev...
3	Stuck in the Middle	2020-07-25 12:27:10	@brookbanktv The one gift #COVID19 has give me...
4	Jammu and Kashmir	2020-07-25 12:27:08	25 July : Media Bulletin on Novel #CoronaVirus...
...	...	...	...
179100	Newton, NJ	2020-08-29 19:44:27	Wallkill school nurse adds COVID-19 monitoring...
179101	Newton, NJ	2020-08-29 19:44:27	Wallkill school nurse adds COVID-19 monitoring...
179102	T.O.	2020-08-29 19:44:23	we have reached 25mil cases of #covid19, world...
179103	Ilorin, Nigeria	2020-08-29 19:44:21	Thanks @IamOhmai for nominating me for the @WH...
179104	Ontario	2020-08-29 19:44:16	2020! The year of insanity! Lol! #COVID19 http...

142334 rows × 3 columns

(a)

	timestamp	tweet_text	location
1	5/25/2021 23:59	b'Doctors plead for stay-at-home order and shu...	Winnipeg
3	5/25/2021 23:59	b"RT @DrJacobsRad: As of today, 50% of Canadia...	Windsor, Ontario
5	5/25/2021 23:59	b"RT @dna: 'Achieved balance between mental, p...	India
7	5/25/2021 23:59	b'I could sense another lockdown coming, Third...	Melbourne VIC, Australia
9	5/25/2021 23:59	b"RT @DrAMSinghvi: Ratan Tata Ji's decision th...	Bengaluru, India
...	...	...	...
95801	6/5/2021 11:32	b'RT @BheeshmaTalks: Megapower Star #RamCharan...	ma intlo
95805	6/5/2021 11:31	b'After missing out from 2020 calendar due to ...	Jaipur, Rajasthan
95807	6/5/2021 11:31	b'1st game back since #Covid19 and we get a wi...	Glasgow, PossilPark
95809	6/5/2021 11:31	b'@davidmweissman @SenateGOP Why won\xe2\x80\x...	Texas, USA
95811	6/5/2021 11:31	b'RT @MANJULtoons: #COVID19 #Vaccination \nMy ...	Gurgaon, India

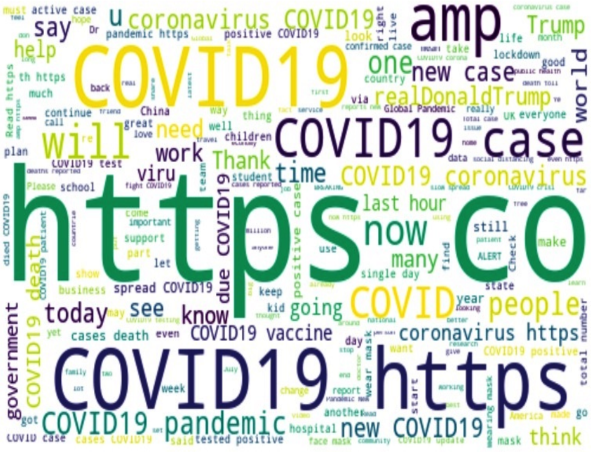
(b)

Figure 2: Sample of (a) Dataset-I (b) Dataset-II.

3.3 Calculation of Polarity and Feature extraction

After the data preprocessing task, using the NLTK and TextBlob libraries the polarity of each tweets are calculated from the CSV files. Sentiment analysis is basically the process of determining the attitude or the emotion of the writer, i.e., whether the sentence emotions inclined towards negative or neutral or positive directions. The ‘sentiment’ method of TextBlob returns two properties, polarity, and subjectivity [6, 15]. Polarity is a float value which lies in the range of [-1, 1]. The polarity 1 means *positive* sentiment and -1 means a *negative* sentiment and 0 means *neutral* sentiment. In Figure 4, one sample of subjectivity and polarity score with respect to three categories for the Dataset-I is shown.

Now, after the categorization of the tweets, to prepare any prediction model it is required to transform the Twitter text data into numeric forms. This process is known as the text vectorization. It is a fundamental step in the process of machine learning for analysing



(a)

(b)

Figure 3: Wordcloud representation without pre-processing (a) Dataset-I (b) Dataset-II.

text. It is also to be noted that different vectorization approach (bag of words model, word2vec, doc2vec etc [8]) may drastically affect the end of the results. By looking at the computational effort and by observing some of the previous successful results, in this study Term frequency-inverse document frequency (TF-IDF) technique is used [11, 13]. TF-IDF technique is one of the popular approach used in information retrieval and text mining for doing the text vectorization or extraction. It allows us to weight terms based on how important they are to a document. A higher weight is given to terms that appear frequently in a particular tweet but do not appear often in the entire data. TF is a weight representing how often a word occurs in a document. If there are several occurrences of the same word in one document, we expect the TF-IDF to rise. To put it in more formal mathematical terms, the TF-IDF score for a word t in the short document sd from a document set is calculated as follows:

$$tf_{t,sd} = \frac{n_{t,sd}}{T} \tag{1}$$

Here, in Eq. (1) in the numerator, n is the number of times the term t appears in the document sd and T is total number of terms in the document.

$$idf_t = log\frac{N}{M} \tag{2}$$

Here, in Eq. (2) in the numerator, N is the total number of short documents (tweets) and M is the number of tweets with term t.

Multiplying these two measures, results in the TF-IDF score of a word in a tweet. The higher the score, the more relevant that word appears in that particular tweet.

3.4 Preparation of the Proposed Classification Model

Once data are vectorized by using the TF-IDF process, the next task is to select the suitable base learning models to prepare a heterogeneous ensemble classification model. Here in this study, by observing various previous works [6, 21, 14] related to text classification and tweeter data classification, five different supervised learning models as the classifiers are considered viz., Naïve Bayes (NB), logistic regression (LR), K nearest neighbour (KNN), Decision tree (DT) and Random Forest (RF).

3.4.1 Naïve bayes

NB is a well known simple and effective method for text classification [6]. NB is a classification technique based

	timestamp	tweet_text	C_text	Subjectivity	Polarity	Analysis
1	5/25/2021 23:59	b'Doctors plead for stay-at-home order and shu...	doctor plead stay at home order shutdown manit...	0.000000	0.000000	Neutral
3	5/25/2021 23:59	b"RT @DrJacobsRad: As of today, 50% of Canadia...	today canadian adult half vaccin that s embarr...	0.233333	-0.033333	Negative
5	5/25/2021 23:59	b"RT @dna: 'Achieved balance between mental, p...	achiev balanc mental physic wellb rubinadilaik...	0.200000	-0.100000	Negative
7	5/25/2021 23:59	b'I could sense another lockdown coming, Third...	i could sens anoth lockdown come third time ch...	0.000000	0.000000	Neutral
9	5/25/2021 23:59	b"RT @DrAMSinghvi: Ratan Tata Ji's decision th...	ratan tata ji s decis thattata steel continu p...	0.375000	-0.125000	Negative
...	...	...	...	...	...	...

Figure 4: Sample results of the tweets with their subjectivity-score, polarity-score and categorization into three classes *negative* or *neutral* or *positive*.

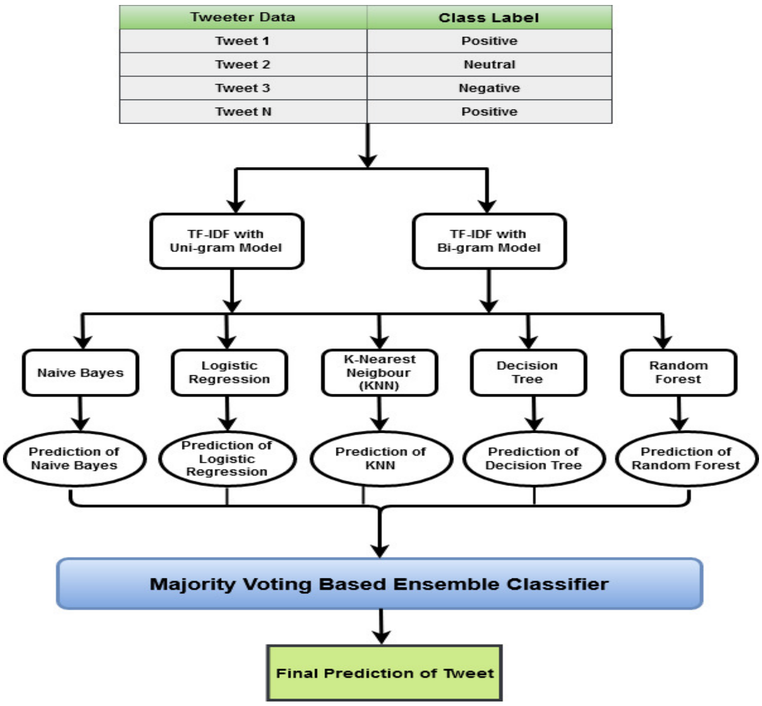


Figure 5: A block diagram of the overall heterogeneous learning process.

on Bayes' theorem with an assumption of independence among predictors. It has been used widely for document classification since 1980s. NB uses maximum posteriori estimation to find out the class (i.e., features are assigned to a class based on the highest conditional probability). There are different variations of NB, out of which Bernoulli's NB (BNB) is used here.

3.4.2 K nearest neighbour

KNN is a lazy and non-parametric learning technique used for the classification problem [10]. It is worthy to mention that KNN does not have a specialized training phase like other learning algorithms. In simple sense at the time of training phase it just load the dataset and when it gets new data point, then it classifies that data into a category based on K value which is much similar to the new data. Euclidean, Manhattan or Hamming distance are some of similarity measures used for numeric data. It achieves high performance when the number of samples is sufficiently large.

3.4.3 Decision tree

DT is one of the most basic and popular techniques for classification or inductive learning [10]. A tree like data structure is built from the training set. The most important attribute is placed at the root node. For evaluation, process start at the root node and goes down the tree by following the corresponding node that meets the condition or "decision". Two heuristic measures mainly entropy and IG of the features are the main building block behind the construction of DT.

3.4.4 Logistic regression

Logistic regression predicts the output of a categorical dependent variable [6]. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. In logistic regression, the concept of the threshold value is used, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

3.4.5 Random forest

RF classifier is basically a bagging based decision trees which falls under the ensemble learning model [10]. It is a combination of different decision trees which are considered as the base learners. Each decision tree is trained with feature-instance sampling with replacement concept with the same distribution. Thus, the trees are trained independently and an unknown samples are labeled according to the majority vote rule: i.e., it is labeled with the most popular class among those provided by the ensemble trees. The ensemble concept of different decision trees in RF classifier leads the system towards low bias and low variance.

Algorithm 1 Majority Voting based Ensemble Classifier.

Representation:
M: Number of classifiers.
N: Classifier.
D: Total number of COVID-19 related tweets.
Y_D: Predicted output of the model.
Input: M, N, D.
Output:Y_D.

```
for i=1 to D do
    for j=1 to M do
        Predict_i = Majority(Predict_ij)
    end for
end for
```

3.4.6 Majority Voting based Ensemble Classifier

The majority voting based classifier comes under the umbrella of ensemble learning. In the literature, there are different approaches available for designing the ensemble models. They are namely weighted averaging, max or majority voting and averaging. By observing several well known previous studies, in this research majority voting technique is being adapted. Majority voting aims to combine different classifiers (the above-mentioned classifiers) by using the "voting" (i.e., predictions given by each model are considered as a "vote") mechanism for improving the accuracy of a learning model. In practice, different classifiers might have performed well on various data points (tweets) in the tweeter dataset (Dataset I and Dataset II). Individual tweets related to COVID-19 domain are considered as the data points. Tweets are classified correctly by some classifiers, and the same tweets may be misclassified by some other classifiers. It is difficult to predict the correctness of the model. However, there is a chance to increase the accuracy of the model by combining the different classifiers by using a voting mechanism. Thus, significant improvement in the model can be achieved by combining the five different classifiers with the voting mechanism. Majority voting is used for predicting the class label of the tweet and the basic steps are given in Algorithm 1. Also, in the Figure 5 one block diagram of the overall learning process is shown.

4 Experimental Setup and Evaluation

To get uniform experimental results, all the methods are implemented in Python. Programs are simulated in a machine with Processor: Xenon(R) CPU- E5-1630, 3.70 GHz clock speed and random access memory of 32GB having Windows 10 environment.

The detailed experimental setup are as follows:
(i) All the methods and functions are implemented in Python 3.7 environment. Data structures like data frames, dynamic lists, collections, and dynamic arrays are used.
(ii) The partitioned of the individual datasets are performed according to a train-test (70% and 30%) spilt scheme. For RF classifier, number of estimator i.e., the number of decision trees in the forest is set to 100 and the maximum depth of the tree is set *none*. For the KNN classifier the "$n - neighbors$" value is set to 5. Then for the other three classifiers BNB, LR, and DT default settings of the 'sklearn' [2]environment are used.
Four different kinds of classifier validity measures [8, 10] namely, (i) percentage accuracy, (ii) precision, (iii) recall, (iv) F1-score are computed to evaluate the performance of the proposed model compared to other models. Also, confusion matrix is drawn for further analysis.

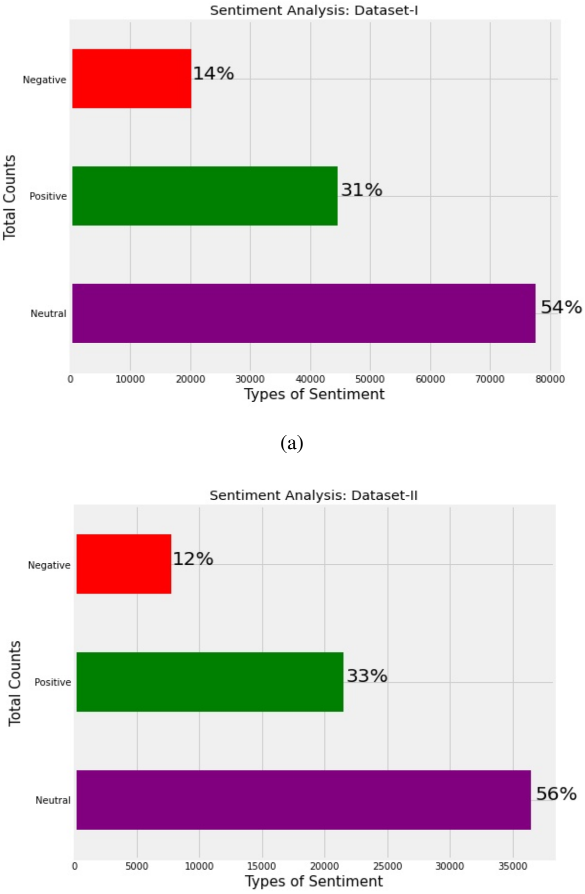
5 Experimental Results Analysis

The detailed experimental results are analysed here.

5.1 Analysis of text categorization and Word frequency

The bar chart representation for the Dataset-I (Kaggle) and Dataset-II is shown below in Figure 6 (a)-(b). In the given figure, the x axis represent the type of sentiments and y axis represents the total counts of the tweet. The percentage of *negative*, *neutral* and *positive* tweets for Dataset-I and Dataset-II can be seen in the figures. It can be observed that the sentiment inclined towards positive is more in 2021 and less in 2020. Also, the number of neutral tweets is more in Dataset II than in Dataset I. The number of negative sentiment is lesser in 2021 than in 2020. Thus, it can be confirmed that the sentiment of the people regarding the pandemic is bit positive in 2021 as compared to 2020.

In Figure 7 and Figure 8 with the utilization of bar charts, 20 most frequently used words in uni-gram bi-gram models of the datasets are shown. Some of the



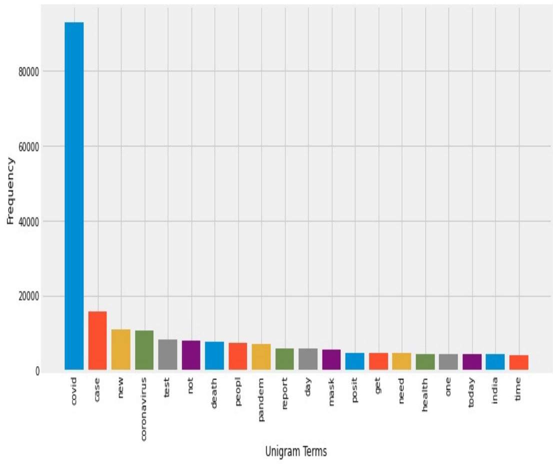
(b)

Figure 6: Graphical Bar chart for the count of Positive, Negative & Neutral Tweets (a) Dataset-I (b) Dataset-II

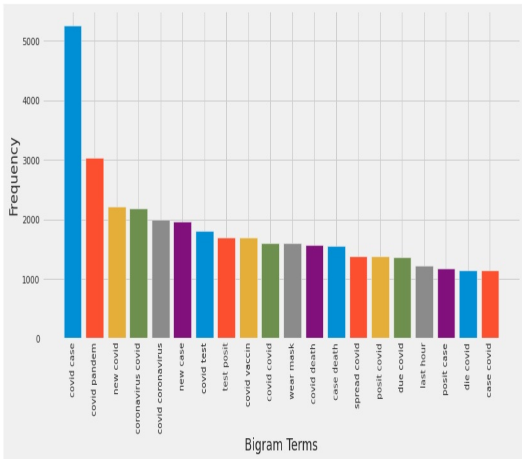
provoking words which contributes the positive and negative sentiments are as follows: negative sentiments, "pandemic", "isolate" , "break, "virus", "coughing", "death", "cry", "help", "fever" "apocalypse", "symptoms", "hospital", "infected", "crisis", "infection" etc. Further, few words which inclines towards the positive sentiment includes, "recovery", "doctor", "relief", "medicine", "food", "trust", "vaccine", "govern", "patient", "good", "support", "mask", "extend", "expert", "protect" etc.

5.2 Analysis of Classifiers Evaluation Measures

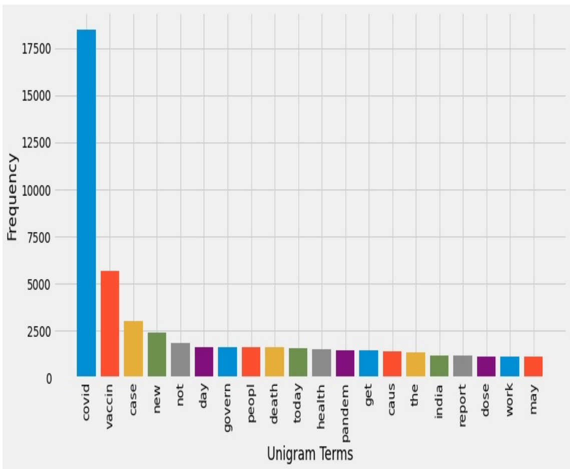
The overall summary of the experimental results achieved on the two datasets (Dataset I and Dataset-II) by using 70% training and 30% test samples are reported in Table 1 and Table 2. The total support of samples on the 30% of the test set of Dataset-I is *negative* 5999, *neutral* 23371 and *positive* 13,331. Similarly the



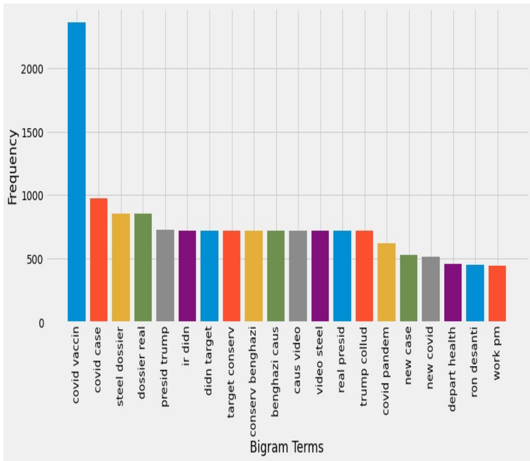
(a)



(a)



(b)



(b)

Figure 7: Bar chart for word frequency of Unigram model for (a) Dataset-I (b) Dataset-II.

Figure 8: Bar chart for word frequency of Bigram model for (a) Dataset-I (b) Dataset-II.

total support of samples on the 30% of the test set of Dataset-II is *negative* 2295, *neutral* 10943 and *positive* 6463.

From the Table 1 it can be observed that for the unigram and bigram feature extraction techniques, proposed ensemble classification model has achieved better accuracy then the base models. The improvements in accuracy achieved by proposed model are (2.67%, 0.29%, 9.15%, 5.9%, 0.61%) and (12.32%, 5.01%, 10.67%, 1.81%, 1.44%) with respect to BNB, LR, KNN, DT, RF methods respectively. Similarly from the Table 2 also it can be observed that for the unigram and bigram feature extraction techniques, proposed ensemble classification model has achieved better accuracy then the base models. The improvements in accuracy achieved by proposed model are (8.36%, 7.22%, 11.33%, 8.94%, 3.33%) and (4.61%, 2.57%, 5.33%, 1.05%, 0.98%) with respect to B-NB, LR, KNN, DT, RF methods respectively.

In Figure 9 and Figure 10 depicts the confusion matrices of the test phase of the proposed model and the base models for COVID-19 Twitter® data classification. A confusion matrix is a technique for evaluating the performance of a classification algorithm. It is a summary of prediction results on a classification problem. The accuracy metric alone can be misleading if we are dealing with more than two classes in the dataset. So, in order to avoid this problem, implementing the confusion matrix can give a better idea of whether the classification model is giving the accurate result or what type of error it is making. The scikit-learn library for machine learning in Python is used to calculate the confusion matrix. In Figure 9 (f) among the 42701 testing tweets samples (30% of Dataset-I), 5494 tweets were misclassified by the proposed ensemble model. Similarly in Figure 10 (f) among the 19701 testing tweets samples (30% of Dataset-II), 997 tweets were misclassified by the proposed ensemble model.

Now, it can be concluded from the summarized experimental results shown in Table 1, Table 2 and from the confusion matrices that proposed ensemble classifier model has achieved better classification evaluation results with both the unigram and bigram models compared to the B-NB, LR, KNN, DT, and RF classifiers.

6 Concluding Remarks and Future Work

Twitter is a massive collection of data which make it more convenient to perform sentiment analysis. In this article, analysis were performed on a total of 142,334 tweets of the Dataset-I for year 2020 and 65,668 tweets of Dataset-II for the year of 2021. Tweets are analysed to get some computational results, so that an in-depth understanding about the opinions about the ongoing pandemic COVID-19 can be understood. One novel heterogeneous majority voting based ensemble learning framework is proposed to predict the tweets into three classes viz., *negative*, *neutral*, and *positive*. Extensive experiments were conducted by applying the feature extraction schemes with the unigram and bigram with TF-IDF technique. Experimental results suggest that on both the feature extraction models i.e., on unigram and bigram techniques, proposed model has performed better than the other compared models. With 70%–30% train-test set, proposed model is able to achieve an accuracy of 94.94% to classify the tweets into various classes. Moreover, it is worthy to mention that this kind of study may help the government to make use of relevant information in policymaking as they will be able to know how people are reacting to this new strain, what all challenges they are facing such as food scarcity, panic attacks, etc. Also, negative sentiments tweets may have some misleading information which also may be prevented from further circulations. However, presently in this study English language tweets are only considered for performing the sentiment analysis, but as a future work author is planning to incorporate multilingual feature in the proposed model.

References

[1] Natural language toolkit (NLTK). `http://www.nltk.org/`.

[2] scikit-learn machine learning in python. `https://scikit-learn.org/stable/`.

[3] World health organization corrona virus dashboard. `https://covid19.who.int/`.

[4] Alamoodi, A. H., Zaidan, B., Albahri, O., Mohammed, K., Malik, R., and Musaab, A. Sentiment analysis and its applications in fighting covid-19 and infectious diseases: A systematic review. *Expert Systems With Applications*, 167:114–155, 2021.

[5] Alrazaq, A. A., Alhuwail, D., Househ, M., and Hamdi, M. Top concerns of tweeters during the covid-19 pandemic: Infoveillance study. *Journal of Medical Internet Research*, 22(4):1–9, 2020.

[6] Ankit and Saleena, N. An ensemble classification system for twitter sentiment analysis. *Procedia Computer Science*, 132(2):937–946, 2018.

[7] Aslam, F., Awan, T. M., Syed, J. H., Kashif, A., and Parveen, M. Sentiments and emotions evoked

Table 1: Summary of experimental results of different classifiers evaluation measures on 70% training and 30% test set obtained by BNB, LR, KNN, DT, RF, and Proposed model classifiers on Dataset-I.

Methods	Unigram					Bigram				
	Class	**Pre**	**Rec**	**F1-SC**	**Acc**	**Class**	**Pre**	**Rec**	**F1-SC**	**Acc**
BNB	Neg	0.80	0.55	0.65	84.53	Neg	0.64	0.13	0.21	63.19
	Neu	0.83	0.96	0.89		Nue	0.61	0.96	0.75	
	Pos	0.90	0.78	0.83		Pos	0.78	0.28	0.42	
LR	Neg	0.93	0.55	0.69	86.91	Neg	0.80	0.12	0.21	64.33
	Neu	0.83	0.99	0.90		Nue	0.61	0.99	0.76	
	Pos	0.95	0.79	0.86		Pos	0.89	0.27	0.42	
KNN	Neg	0.92	0.29	0.44	78.05	Neg	0.45	0.33	0.49	60.22
	Neu	0.73	0.99	0.84		Neu	0.65	0.49	0.31	
	Pos	0.95	0.62	0.75		Pos	0.47	0.68	0.57	
DT	Neg	0.60	0.61	0.60	81.30	Neg	0.80	0.18	0.29	62.61
	Neu	0.85	0.87	0.86		Neu	0.58	0.97	0.72	
	Pos	0.84	0.81	0.82		Pos	0.85	0.34	0.49	
RF	Neg	0.86	0.57	0.69	86.59	Neg	0.76	0.56	0.55	68.22
	Neu	0.84	0.97	0.90		Nue	0.68	0.55	0.70	
	Pos	0.93	0.81	0.86		Pos	0.80	0.53	0.71	
Proposed Model	Neg	0.87	0.66	0.75	**87.20**	Neg	0.81	0.44	0.57	**71.55**
	Neu	0.89	0.99	0.94		Nue	0.63	0.91	0.75	
	Pos	0.97	0.87	0.92		Pos	0.89	0.60	0.72	

Table 2: Summary of experimental results of different classifiers evaluation measures on 70% training and 30% test set obtained by BNB, LR, KNN, DT, RF, and Proposed model classifiers on Dataset-II.

Methods	Unigram					Bigram				
	Class	**Pre**	**Rec**	**F1-SC**	**Acc**	**Class**	**Pre**	**Rec**	**F1-SC**	**Acc**
BNB	Neg	0.57	0.57	0.58	82.61	Neg	0.65	0.27	0.38	70.27
	Neu	0.86	0.90	0.88		Nue	0.67	0.95	0.79	
	Pos	0.86	0.79	0.82		Pos	0.89	0.43	0.58	
LR	Neg	0.85	0.62	0.72	89.92	Neg	0.89	0.26	0.40	72.31
	Neu	0.88	0.98	0.93		Nue	0.68	0.97	0.80	
	Pos	0.95	0.87	0.91		Pos	0.89	0.47	0.61	
KNN	Neg	0.79	0.62	0.70	84.26	Neg	0.77	0.33	0.45	69.55
	Neu	0.82	0.96	0.88		Neu	0.59	0.91	0.84	
	Pos	0.93	0.73	0.82		Pos	0.88	0.42	0.63	
DT	Neg	0.60	0.61	0.60	93.12	Neg	0.85	0.31	0.44	73.83
	Neu	0.85	0.87	0.86		Neu	0.69	0.97	0.81	
	Pos	0.84	0.81	0.82		Pos	0.91	0.50	0.64	
RF	Neg	0.93	0.76	0.83	93.49	Neg	0.80	0.30	0.44	73.90
	Neu	0.91	0.99	0.95		Nue	0.69	0.97	0.81	
	Pos	0.98	0.90	0.94		Pos	0.90	0.50	0.64	
Proposed Model	Neg	0.95	0.80	0.87	**94.94**	Neg	0.80	0.44	0.63	**74.88**
	Neu	0.94	0.99	0.96		Nue	0.77	0.95	0.81	
	Pos	0.97	0.94	0.95		Pos	0.86	0.66	0.71	

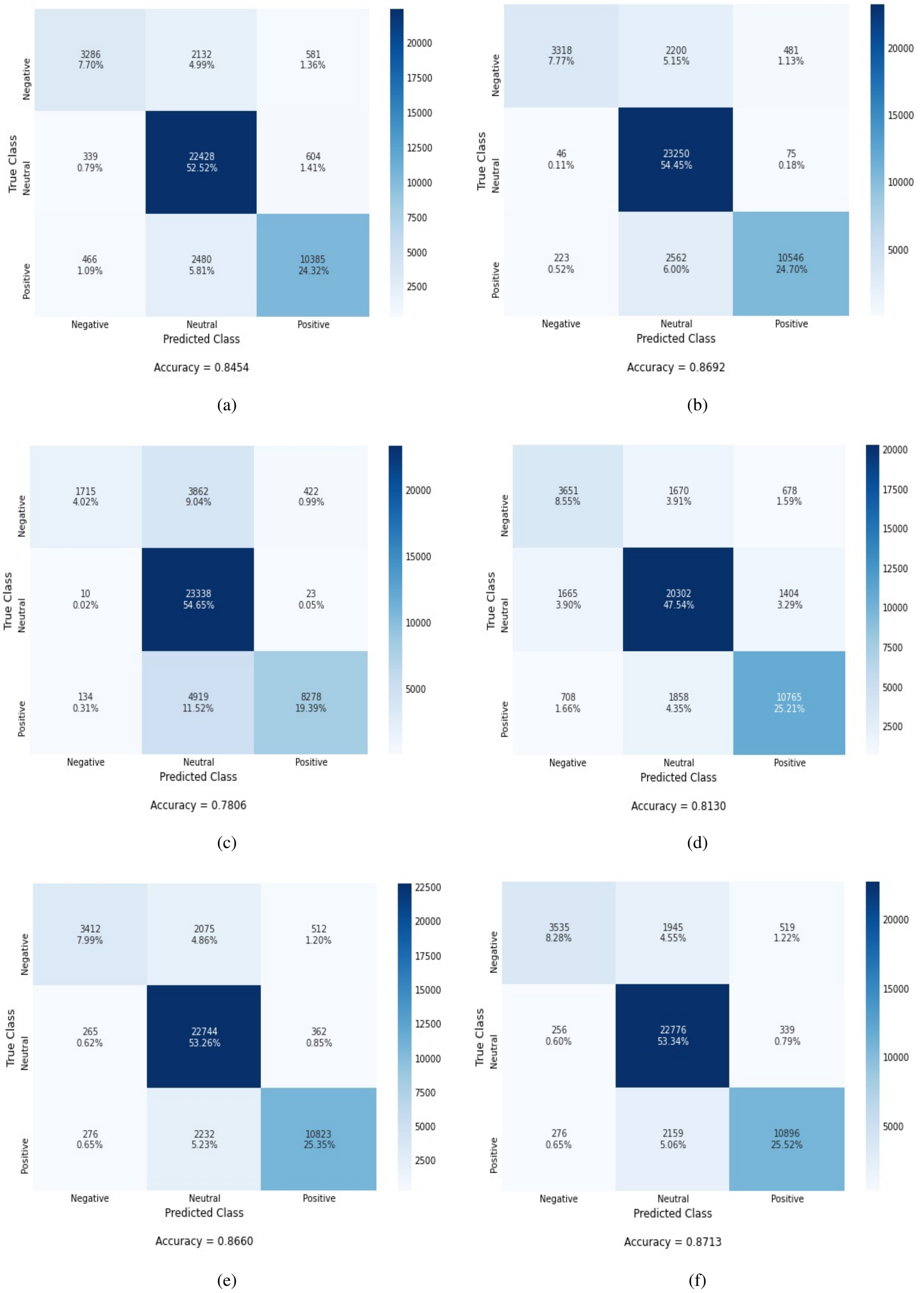


Figure 9: Confusion matrix for the unigram model by various classifiers (a) Bernoulli NB (b) Logistic regression (c) K nearest neighbor (d) decision tree (e) Random forest (f) proposed model on Dataset-I

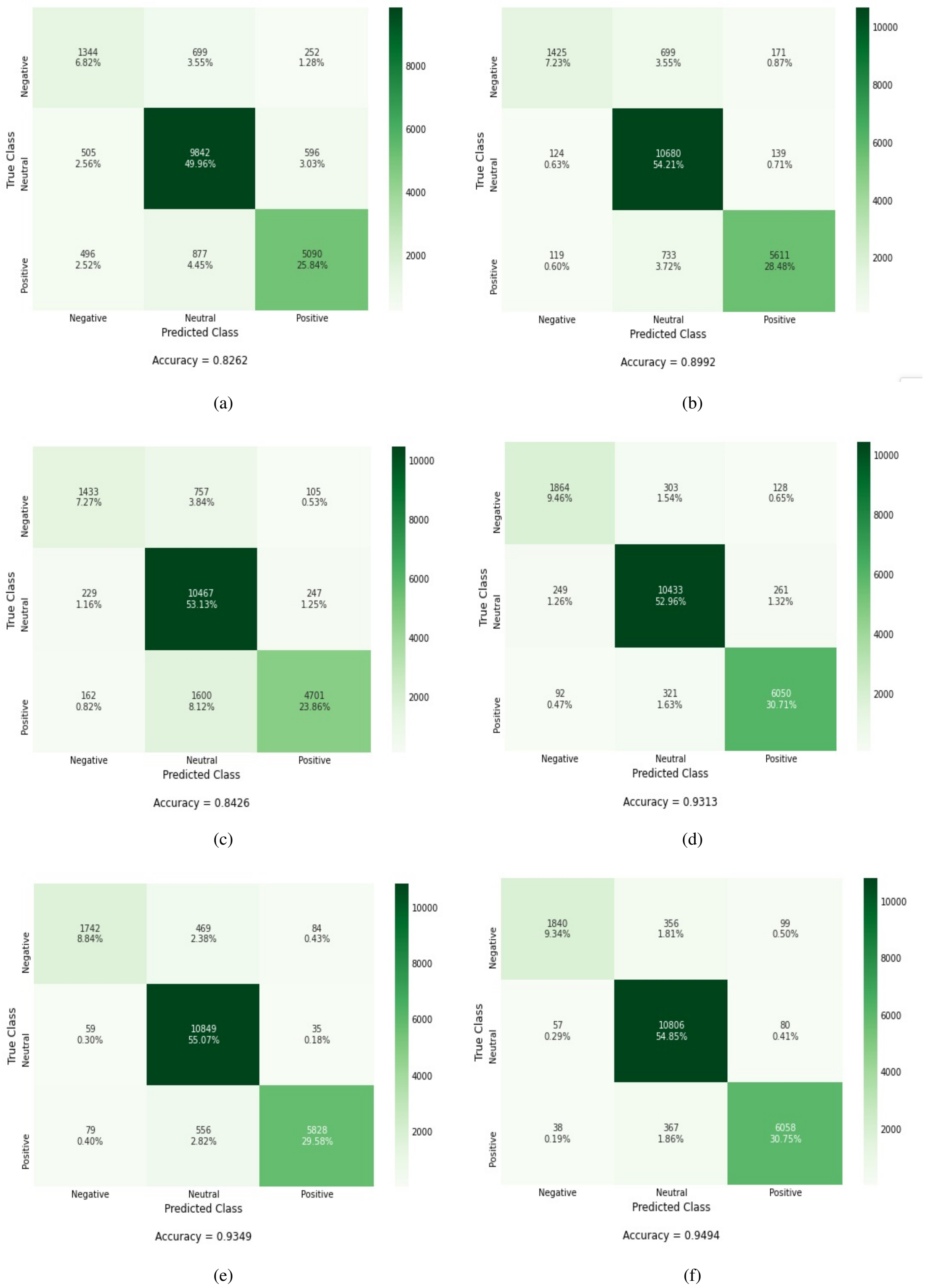


Figure 10: Confusion matrix for the unigram model by various classifiers (a) Bernoulli NB (b) Logistic regression (c) K nearest neighbor (d) decision tree (e) Random forest (f) proposed model on Dataset-II

by news headlines of coronavirus disease (covid-19) outbreak. *Humanities and social sciences communications*, 7(23):1–9, 2020.

[8] Bania, R. K. Covid-19 public tweets sentiment analysis using tf-idf and inductive learning models. *INFOCOMP Journal of Computer Science*, 19(2):23–41, 2020.

[9] Chandra, R. and Krishna, A. Covid-19 sentiment analysis via deep learning during the rise of novel cases. *PLOS ONE*, 16(8):e0255615, 2021.

[10] Han, J. and Kamber, M. *Data Mining Concepts and Techniques*. Morgan Kaufmann Publishers, New York, fourth edition, 2011.

[11] Havrlant, L. and M., G. J. A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation). *International Journal of General Systems*, 46(1):1–21, 2017.

[12] Jelodar, H., Wang, Y., Orji, R., and Huang, H. Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: Nlp using lstm recurrent neural network approach. *IEEE Journal of Biomedical and Health Informatics*, 24(10):2733–2742, 2020.

[13] Kim, S. W. and Gil, J. M. Research paper classifcation systems based onÂ tf-idf andÂ lda schemes. *Human Centric Computation and Information Science*, 9(30):1–21, 2019.

[14] Kowsari, K., Jafari, K. M., Heidarysafa, M., and Brown, D. Text classification algorithms: A survey. *Information*, 10(150):6–10, 2019.

[15] Liu, B. *Handbook of Natural Language Processing*. CRC Press, New York, second edition, 2010.

[16] Nemes, L. and Kiss, A. Social media sentiment analysis based on covid-19. *Journal of information and telecommunication*, 5(1):1–15, 2021.

[17] Pedrosa, A. L., Bitencourt, L., Cazumba, M. L., and Silva, A. C. S. Emotional, behavioral, and psychological impact of the covid-19 pandemic. *Frontiers in Psychology*, 11:1–18, 2020.

[18] Ren, Y., Zhang, L., and Suganthan, P. Ensemble classification and regression-recent developments, applications and future directions. *IEEE Computational Intelligence Magazine*, 11(1):41–53, 2016.

[19] Rosa, R. L. R., Silva, M. J., and Silva, D. H. e. a. Event detection system based on user behavior changes in online social networks: case of the covid-19 pandemic. *IEEE Access*, 8:158806–158825, 2020.

[20] Sailunaz, K. and Alhajj, R. Emotion and sentiment analysis from twitter text. *Journal of Computational Science*, 36:1–18, 2018.

[21] Samuel, J., Ali, G. G., Rahman, M., Esawi, E., and Samuel, Y. Covid-19 public sentiment insights and machine learning for tweets classification. *Information*, 22(4):1–21, 2020.

[22] Singh, M., Jakhar, A. K., and Pandey, S. Sentiment analysis on the impact of coronavirus in social life using the bert model. *Social Network Analysis and Mining*, 11(33):https://doi.org/10.1007/s13278–021–00737–z, 2021.