Deep Learning Based Sentiment Analysis of COVID-19 Tweets via Resampling and Label Analysis

Mamoona Humayun[1,*], Danish Javed[2], Nz Jhanjhi[2], Maram Fahaad Almufareh[1] and Saleh Naif Almuayqil[1]

[1]Department of Information Systems, College of Computer and Information Sciences, Jouf University, Sakakah, 72311, Saudi Arabia
[2]School of Computer Science, SCS, Taylor's University, Subang Jaya, 47500, Selangor, Malaysia
*Corresponding Author: Mamoona Humayun. Email: mahumayun@ju.edu.sa

Abstract: Twitter has emerged as a platform that produces new data every day through its users which can be utilized for various purposes. People express their unique ideas and views on multiple topics thus providing vast knowledge. Sentiment analysis is critical from the corporate and political perspectives as it can impact decision-making. Since the proliferation of COVID-19, it has become an important challenge to detect the sentiment of COVID-19-related tweets so that people's opinions can be tracked. The purpose of this research is to detect the sentiment of people regarding this problem with limited data as it can be challenging considering the various textual characteristics that must be analyzed. Hence, this research presents a deep learning-based model that utilizes the positives of random minority oversampling combined with class label analysis to achieve the best results for sentiment analysis. This research specifically focuses on utilizing class label analysis to deal with the multiclass problem by combining the class labels with a similar overall sentiment. This can be particularly helpful when dealing with smaller datasets. Furthermore, our proposed model integrates various preprocessing steps with random minority oversampling and various deep learning algorithms including standard deep learning and bi-directional deep learning algorithms. This research explores several algorithms and their impact on sentiment analysis tasks and concludes that bidirectional neural networks do not provide any advantage over standard neural networks as standard Neural Networks provide slightly better results than their bidirectional counterparts. The experimental results validate that our model offers excellent results with a validation accuracy of 92.5% and an F1 measure of 0.92.

Keywords: Bi-directional deep learning; resampling; random minority oversampling; sentiment analysis; class label analysis

1 Introduction

The international epidemic of COVID-19 has created consequences for virtually all nations and areas. Nations have alerted their citizens to be vigilant by setting rules and including face masks, keeping physical distance from others, and not partaking in big gatherings. The lockdown was implemented by most nations as the necessary method to deal with the issue and reduce virus spread [1]. Since the proliferation of the coronavirus, the world in its entirety has been in an exceedingly terrible state. COVID-19 has been responsible for the demise of millions of humans. This sickness has extraordinarily affected people in more than just physical aspects. Coronavirus infection and fatality rates were rising rapidly in the first year of the pandemic. Even though the impact of the virus has slowed down, the problem persists. COVID-19, produced from SARS-CoV-2, created havoc on the human population in the last two years and still causing problems for the human population. The expression "pandemic" refers to a sickness that proliferates quickly and surrounds a geographical area, such as a nation [2,3].

Sentiment analysis or otherwise called opinion mining has emerged as a prominent study area with the rise of social networks [4]. The Internet's expanding prominence has elevated it to the position of a primary source of essential knowledge. Many individuals communicate their thoughts and views through numerous internet resources. We must utilize consumer data to evaluate it automatically to continually assess popular sentiment and help decision-making. As a response, opinion mining has received significant recognition in recent times. In current times, humans tend to trust social network news more than actual news channels. It is anticipated that social networks will guide users to real news in the near future. However, in most circumstances, the news prompted people to make incorrect judgments, such as the news related to COVID-19 [5]. Older people are not the only group that has been affected by this pandemic as young people such as students can also fall victim [6]. COVID-19 also caused mental health issues as the comments and tweets on this pandemic have proven to be unsettling and a source of concern that must be addressed to handle deceptive information [7,8].

Several machine learning (ML) and deep learning (DL) methods can be utilized for the sentiment analysis of COVID-19 data [9,10]. These techniques follow a base pipeline to predict multiclass classifications. Fig. 1 presents the pipeline for standard ML techniques while Fig. 2 presents a standard deep-learning pipeline that can be used for sentiment analysis for the COVID-19 data.
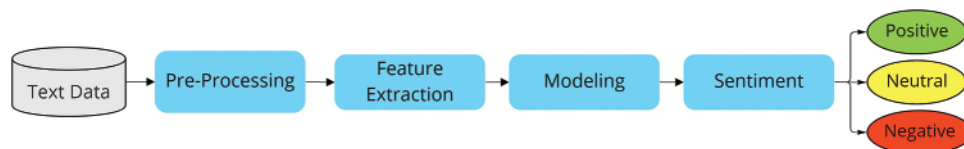

Figure 1: Standard machine learning pipeline for sentiment analysis
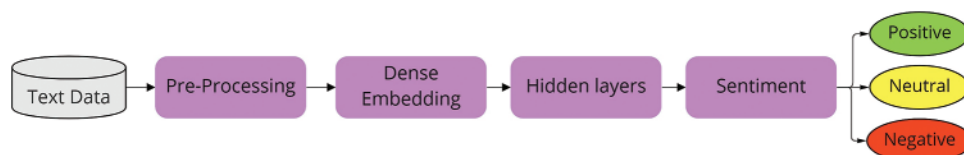

Figure 2: Standard deep learning pipeline for sentiment analysis

It is also important to notice that most datasets are imbalanced which results in poor accuracy of the model [11]. This is especially problematic for the limited-size multiclass datasets as there are not

enough samples in each class for accurate training of the ML or DL algorithms. Therefore, to address these issues, our research focuses on dealing with these problems. Our paper focuses on the supervised deep learning approach based on standard deep learning and bi-directional deep learning algorithms, as they can provide promising classification results. In this research, we have made the subsequent contributions:

- Presenting a complete model for enhanced sentiment analysis to handle class imbalance while utilizing multiple deep-learning algorithms for better predictions. As most of the available data on COVID-19 is limited therefore it is important to create a model that tackles the issue while still maintaining overall accuracy.
- Determining the optimal Twitter text preprocessing order to enhance sentiment analysis. The preprocessing steps play an important role when handling textual data. It is not only important to use the right preprocessing steps but it is equally as crucial to create a suitable order for these steps to work accurately.
- Dealing with class imbalance for multiclass classification through class label analysis and oversampling. Considering the limited data, it can be challenging when the dataset is divided into too many classes. Therefore, it is important to look at the task at hand to decide whether class label analysis must be performed. In this research, we prove that class combination can provide excellent performance while still achieving the overall objective of detecting public sentiment.
- Finding the actual impact of bi-directional algorithms against standard neural networks. Much of the previous research has shown that bi-directional algorithms can outperform their standard counterparts. Therefore, this research tests that theory by checking the performance of our model against these algorithms to find the optimal deep-learning solution for sentiment analysis.

The rest of the research is structured as follows: Section 2 contains the chronological literature review, Section 3 presents the methodology of our model, Section 4 provides experimentation and result details, Section 5 presents various exploratory visualizations, and Section 6 concludes this research and provides future directions.

2 Literature Review

In this section, a literature review of sentiment analysis is presented focusing on ML and DL techniques. The keywords "Sentiment analysis machine learning" and "Sentiment analysis deep learning" were used to find related research. There are several deep-learning techniques utilized for sentiment analysis. Fig. 3 provides an overall look at these techniques. Furthermore, this literature review provides a comparison of several techniques used for COVID-19 tweet sentiment analysis in Table 1 at the end of the literature review section.

Reference [12] used a conjunction of RNN and language modeling (LM) to anticipate the upcoming term in the given text. The investigations revealed a significant decrease in the word mistake rate [13]. demonstrated an RNN language model that performs better than conventional language process modeling approaches. The recommended model's training and testing phases are fifteen times faster and more precise than current language modeling approaches.

The shifting mean of streamflow temporal data was also employed as input to the ANN model by the authors in this study [14]. They examined approaches for evaluating, modeling, and improving social media in his work. In this article, the authors go through the process of handling social media

data, evaluating it, and developing accurate models for data analysis. In another research [15], authors use supervised ML algorithms such as Support vector machine (SVM), unigram, and bigram to gauge Twitter data. The findings obtained by applying these methods to Twitter data demonstrated that n-gram (uni and bi) algorithms surpassed SVM. The outcomes include distinctive tags, modifiers, and emotional phrases that are employed to improve the performance evaluation of emotions.
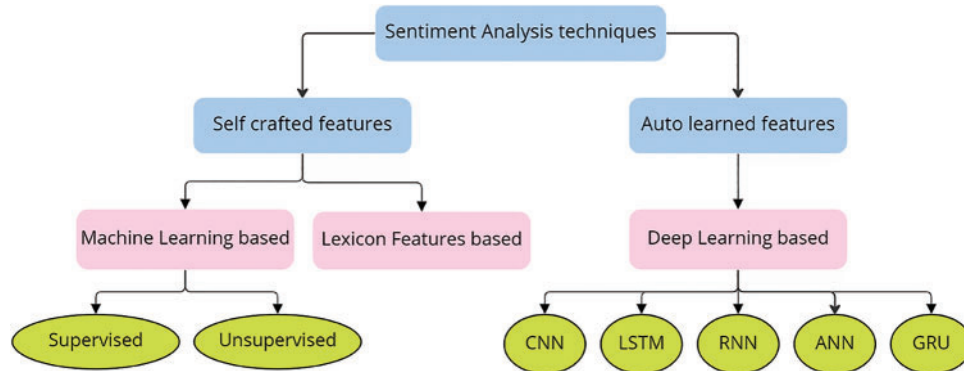

Figure 3: Sentiment analysis techniques

Reference [16] improved text classification algorithms to predict movements in picture data for sentiment analysis. Their technique demonstrated that linguistic and visual information for labeling emotions within a picture is insufficient for prediction. The authors performed the experiments on two data repositories and proved that the suggested strategy improves accuracy over state-of-the-art technologies. Reference [17] suggested a lexicon-based categorization system in which contextual grasp techniques are included via global and local perspectives. The writers have also created sentiwordnet, a hybrid technique for a universal lexicon with genre-specific terminology.

Social media sites including Facebook, Reddit, and Twitter can be an important asset in quickly creating data and disseminating material. In recent times, hate speech has risen and several topics are disseminated to promote content relating to a certain issue. Reference [18] suggested a natural language processing filtering technique to filter these sorts of talks. The results anticipated that character-level strategies outperformed token-level approaches. The technique revealed that a lexical record of resources might be valuable in grading when it is combined with different techniques. Reference [19] developed a unique metaheuristic strategy based on cuckoo and K-means searching algorithms. The suggested technique was evaluated and compared to other similar techniques by testing on distinct Twitter databases. The suggested technique surpasses modern techniques, according to experimental data.

Reference [20] proposed an effective method for extending the notion of the Layer-wise Relevance Propagation (LRP) technique to recurrent methods by employing multiplication collaborations through the use of an enhanced form of LRP. The emotive expressions inside a text were calculated by employing the Long Short-Term Memory (LSTM) approach to assess the output's significance concerning the feelings in a text. The investigational findings also indicated whether the classifier's judgment is useful to a certain class or not, in addition to how well these algorithms perform against the gradient-based breakdown.

In another study [21], authors recommended a Recurrent Neural Network (RNN) for the interpretation of neural sequences to extract terms from text. The investigational findings show that the suggested model outperforms other cutting-edge deep learning methods. This study [22] proposed

a multi-stage learning strategy based on the RNN. The authors offered three alternative information-sharing strategies to define the text utilized to execute a certain task and comprise a few layers that operate on using the standard set of rules. This strategy can progress the field of text analytics by utilizing more interrelated responsibilities.

In this research [23], the authors provide a text embedding approach that is based on unsupervised learning that uses hidden contextual semantic links and several statistical properties in tweet content. To construct a set of attributes for tweets, these text embeddings are merged with n-grams characteristics and sentiment polarity score characteristics that are then input to a Convolution Neural Network (CNN).

In another study [24], the authors presented a multimodal embedding space approach that analyzed various sets of photos and extracted text from these photos. Their model assessed the emotions in the photos using an SVM on the text characteristics. Using numerous-output support vector regression with the various input/output method.

In this research [25], the authors explore several possible applications of smart speech analysis for COVID-19 victims. They created audio-based algorithms to automatically classify patients' health conditions from four dimensions, such as the degree of sickness, sleep performance, weariness, and anxiety, by analyzing voice recordings from such patients. Their investigations reveal that assessing the severity of disease, which is determined by the length of time spent in the hospital, has an accuracy of 69%. In another research [26], the authors present a hybrid sentiment analysis paradigm that categorizes Vietnamese reviews as either positive or negative. The suggested approach has the unique property of being constructed on a mix of three separate textual representation models that concentrate on assessing social network textual features. The method achieves an accuracy score of 81.54%, which is higher than other solutions.

Authors of this research [27], present a unique model to investigate cross-correlations between pictures, words, and their social ties, which may learn complete and complimentary characteristics for efficient sentiment analysis. To understand the cross-connections between image and text at multiple levels, the authors integrate visual material with distinct semantic pieces of textual information using a 3-level LSTM. To efficiently leverage the connection information, the links between social photos are described by a weighted relationship network, with every node contained in a vector.

Another research [28] introduced a method where Twitter data from Canada and texts containing social distance keywords were generalized, then the SentiStrength program was used to extract sentiment polarity from tweet texts. Following that, the SVM technique was used to classify sentiment. The purpose was to comprehend and evaluate popular attitudes regarding social distance as expressed in Twitter textual data.

In this research [29], the authors of the study used Twitter to collect data on Filipinos' attitudes toward the Philippine government's efforts and employed Natural Language Processing (NLP) practices to comprehend the overall emotion, which will assist the authority in assessing their reaction. The emotions were labeled and trained with the Naïve Bayes (NB) algorithm to categorize English and Filipino text into three categories. The accuracy of the results was 81.77%, which outperformed prior sentiment analysis research employing data from the Philippines.

This current study [30] aims to evaluate Persian tweets to investigate the attitude of the public towards the COVID-19 vaccine and contrast Iranian attitudes about local and imported COVID-19 vaccines. The sentiments of the obtained tweets were then detected by employing a deep-learning opinion-mining approach centered on CNN-LSTM design. The authors noticed a modest disparity in

the number of favorable attitudes toward local and imported vaccinations, with the latter having the majority positive polarity.

Table 1: COVID-19 tweet sentiment analysis-A comparison of techniques and applications

Cite	Data type/volume	Techniques	Application	Future application
[31]	Tweets/20,325,929	CrystalFeel	Trends regarding rage, grief, and joy	Inclusion of more media platforms
[32]	Tweets/500,000	TextBlob	Calculating the polarity of tweets	Inclusion of more social media platforms
[33]	Tweets/57.5 million	Transformer/BERT	Sentiment regarding vaccines	Social network monitoring in real time
[34]	Tweets/3.37 million	SVM	A shift in racial sentiment	A shift in racial attitudes over time
[35]	Tweets/16 million	TClustVID	Analyze sentiment and topics	Analyze other datasets with different topics
[36]	Tweets/293,000	Binary logit model	Sentiment re-opening	Socio-economic and household data
[37]	Tweets/7528	TextBlob, DL, and ML algorithms	Sentiment analysis	Utilizing more DL approaches
[38]	Tweets/900,000	Naïve Bayes, Logistic Regression	Public's opinion of the spread of COVID-19	Include information from news articles
[39]	Tweets/293,597	N-gram, R packages	Sentiment re-opening	Apply on different social media data
[40]	Tweets/4 million	LDA/NLP	Sentiment analysis regarding COVID-19	Analyze public trust

3 Methodology

This section describes the proposed model in great depth. All of the essential preprocessing processes are also covered. Furthermore, our deep learning-based model is divided into several modules to enhance sentiment analysis. All the modules are thoroughly explained. This is followed by a discussion of how to fine-tune the model and how to accurately utilize this model.

3.1 Proposed Model

The proposed model consists of various functional modules that are divided into text-dependent and text-independent components. These functional modules and the complete model are shown in Fig. 4. Our model employs a modular method that utilizes numerous theories with attention to improvement in terms of accuracy. This model consists of a preprocessing module which cleans the text in a format that can be useful for the classification algorithm. The lemmatization module normalizes the preprocessed text to reduce the number of features. Class label analysis combines classes with similar overall sentiment, it is further explained in Section 3.1.3. The class balancing module deals with the issue of class inequality through random minority oversampling to avoid overfitting the majority class. One hot encoding transforms the text for the classification algorithm, so it can accurately make the predictions. Furthermore, the data is divided into three parts which are train, test, and validation

set. A train set is used to train our deep learning-based model while a test set is used to test the performance and tweak our model and finally the results are produced based on the validation set.
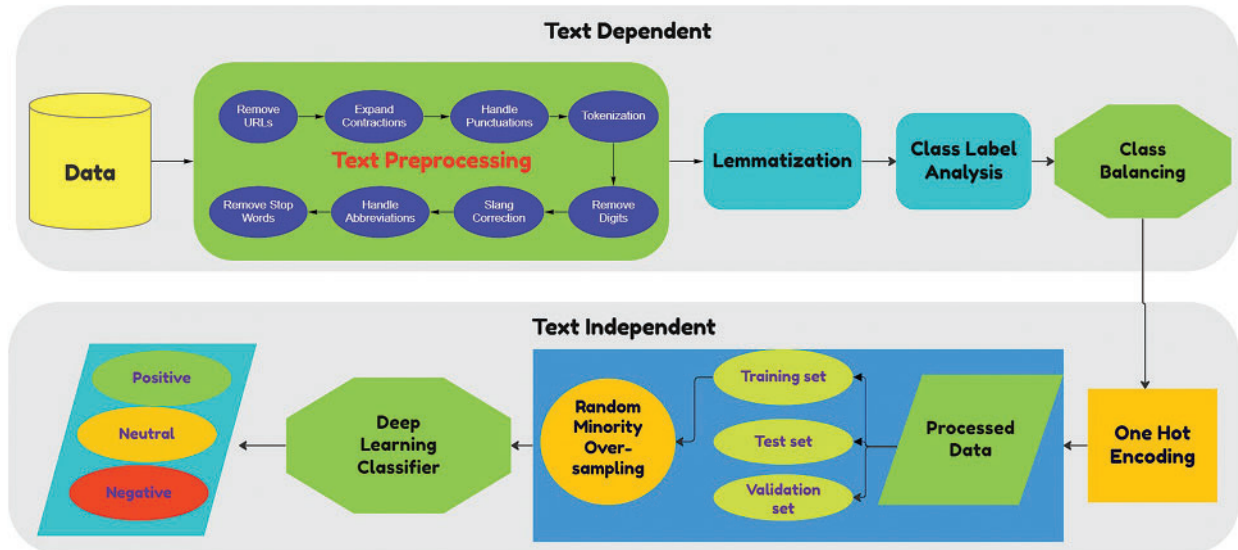

Figure 4: Deep learning-based model

3.1.1 Data Exploration and Pre-processing

Data exploration is the initial stage in any process of data analysis since it allows you to identify the themes and trends buried in the data. The first thing to do is to see whether there are any absent or null values in the data. While the data may consist of thousands of rows, if crucial fields such as 'Text' or 'sentiment' include any missing value, they must be handled by either deleting the entire row or calculating the missing value. After that, we must deal with the text of the tweets according to our problem, as the task of preprocessing is process dependent. Since the proposed model is dealing with COVID-19 data, it is important to preprocess the data accordingly. The presented model handles the preprocessing through some standard steps as well as some manual steps to create a clean text that favors COVID-19 sentiment analysis. Data cleansing is the first stage of this framework's processing pipeline. The extracted data is taken from the files and kept in memory for cleanup at this step [41]. After the initial cleaning of data, we observed the cleaned data and then adjusted our preprocessing steps until a suitable preprocessed text was available.

3.1.2 Lemmatization

The process of converting a word to its original form is called lemmatization. For text normalization, lemmatization is utilized as it generates better results when compared to other techniques such as stemming. Stemming only cuts certain parts of the words which can lead to ambiguity. Although stemming is faster than lemmatization and therefore requires less processing time, the difference in quality must be given priority when we are dealing with sentiment analysis. The proposed model utilizes part-of-speech tagging and *WordNetlemmatizer* to perform lemmatization.

3.1.3 Class Label Analysis

The dataset originally contained 5 classes, but due to limited samples in the dataset, it can be challenging to train the classifier to produce accurate results. Therefore, the model utilizes class

combination in which it combines multiple classes with the same overall sentiment. For example, 'extremely positive' and 'positive' fall into the same overall sentiment category. Similarly, 'extremely negative' and 'negative' fall into the same sentiment category. Thus, it combines these classes which results in overall class labels reduced from five to three. This makes sense as we are dealing with COVID-19-related sentiment analysis and therefore, the severity of the sentiment is of little value, but the overall accuracy should be given higher importance.

3.1.4 Class Balancing

When employing machine learning or deep learning methods, it is vital to train the classifier on a dataset that has a similar amount of training samples to avoid class bias. This is referred to as class balancing. To accurately train a model, it can be useful to have balanced class labels; but, if the class labels are not balanced, we can utilize a class balancing approach before utilizing an ML or DL algorithm. For our research, avoiding Class Imbalance is critical when employing a deep-learning-based technique since we aim to train our model that generalizes well across all conceivable classes. Therefore, random minority oversampling is applied to solve the issue of class imbalance. In this technique, random samples from minority classes are duplicated to increase their size. Since the samples are chosen randomly, it can provide useful results.

3.1.5 Word Representation

This research uses a one-hot encoding which is a process through which we transform categorical variables into a shape that can be given to the classification algorithm so that it can do its job of making predictions. The categorical value symbolizes the numerical value of a sample in the dataset. We have opted to utilize *sklearn.preprocessing.OneHotEncoder* to encode categorical values as a one-hot numerical array. The input to this technique must be integers or strings, representing the values shown by the categorical attributes. Table 2 below provides an example of One Hot Encoding.

Table 2: One hot encoding example

Company name	Categorical value	Price	OHE transformation	Honda	Toyota	Suzuki	Price
Honda	1	100,000	→→→ 1		0	0	100,000
Toyota	2	150,000	0		1	0	150,000
Suzuki	3	70,000	0		0	1	70,000
Suzuki	3	50,000	0		0	1	50,000

3.1.6 Random Minority Oversampling (RMO)

After class analysis, where we combine classes of similar sentiment, it can create the issue of class imbalance against the neutral tweets, therefore, to solve this issue, we apply random minority oversampling. Class imbalance can be dealt with in two ways, either we utilize under-sampling or oversampling. But as the dataset has limited samples, therefore, it is better to use oversampling [42]. RMO will take random samples from the minority class and duplicate these samples to create a balanced dataset for the classifier. But we must be careful as oversampling can lead to overfitting, therefore, we will only apply the oversampling process to the training data while the test data will not be touched. This method will create accurate training data for the learning of our deep learning classifiers for accurate predictions.

3.1.7 Training via Deep Learning

For this module, we train our model through two separate measures. First is standard deep learning algorithms and second is bi-directional deep learning. RNNs are one kind of neural network that can be used to analyze sequential input. An RNN, unlike a standard neural network, does not limit its input or output to a set of fixed-sized vectors. It also does not limit the number of computing steps needed to train a model. Instead, it enables us to train the model with a series of vectors.

A typical RNN state (basic RNN, GRU, or LSTM) is based on past and current occurrences. A state at time t is determined by the states. However, there are times when a forecast is dependent on past, present, and future occurrences. Bidirectional RNNs are used to permit straight (past) and backward traversal of input (future). A Bi-RNN is a combination of two RNNs, one of which goes forward from the beginning of the data sequence, and the other moves backward from the end of the data sequence. A Bi-RNN's standard blocks can be simple RNNs, GRUs, or LSTMs. Fig. 5 shows the standard architecture for deep neural networks.
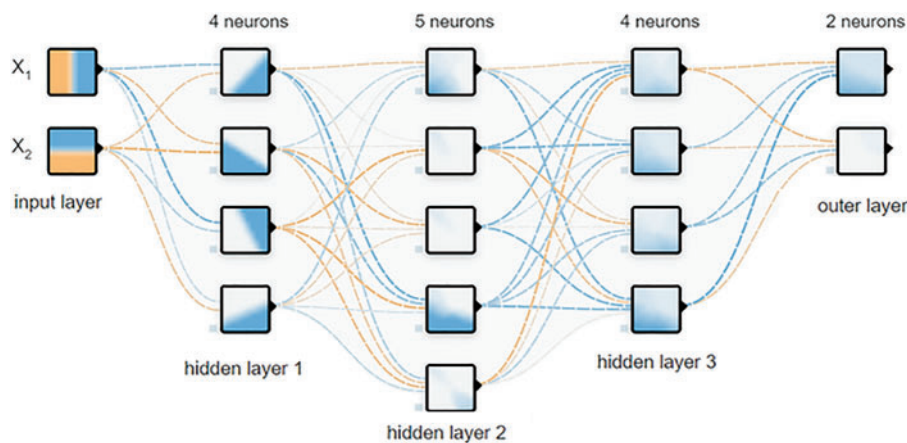

Figure 5: Standard deep learning architecture

Fig. 6 presents the pipeline used for the implementation of the deep learning algorithms through *keras* which is a deep learning API for python. *Keras* provides an embedding layer that may be utilized for neural networks on textual data. Then, the DL layer is added which represents a specific DL algorithm such as LSTM or Bi-LSTM. Pooling layers decreases the dimensions of the feature maps. As a consequence, the amount of network computation and characteristics that need to be learned is lowered. We chose max pooling for our task. Keras dense layer applies element-wise activation while drop out value is set to 0.2. Finally, the softmax activation function is used which will allocate probabilities to every class.

Figure 6: Pipeline for Bi-directional algorithms

3.2 Dataset

In our research, we have opted to use the Coronavirus tweets NLP-Text Classification dataset which is separated into two separate documents training document and a test document. The dataset

contains various tweets from multiple users about COVID-19 and their sentiments. This Twitter data is comprised of six features, but the username and screen names of users have been coded to avoid any privacy concerns. Table 3 lists the features of the dataset.

Table 3: Attribute description of the selected dataset

Dataset attributes	Details
Username	Ordered number that is assigned as a username.
Screen name	Ordered number that is assigned as a screen name.
Locations	Location attribute mentions the location of the person who made the tweet.
TweetAt	The date when the tweet was posted.
Original tweet	The original text of the tweet as posted by the user.
Sentiment	Class labels of the tweet.

The original dataset contains five class labels and it is divided into two files training data and test data. Fig. 7a provides the class distribution for the training data while Fig. 7b shows the class distribution of the test data. For our task, the 'extremely positive' and 'positive' class is combined, similarly, the 'extremely negative' and 'negative' class is combined.
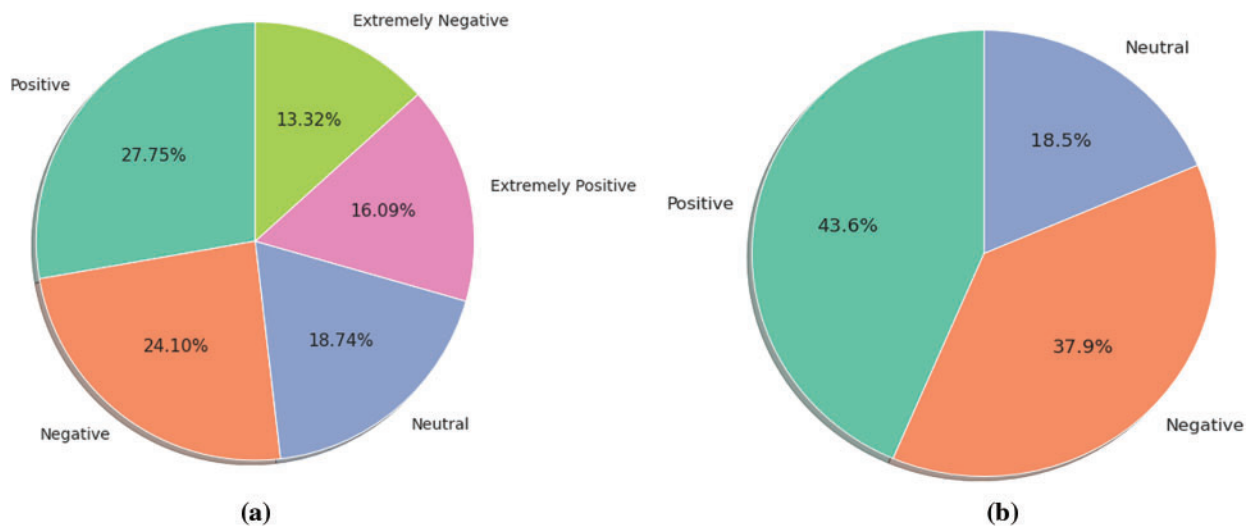

Figure 7: (a) Class distribution before class combination (b) Class distribution after class combination

4 Results and Discussion

In this section, the article will provide specifics on the results obtained during our experimentation. It begins by discussing the computer hardware used for experiments and various tools used for testing. Next, we also examine various evaluation techniques and our model's performance. We also discuss the range of performance measurements, consisting of precision, recall, and F1-measure. Then we provide results using the bi-directional deep learning algorithms. Sentiment analysis for a specific task (such as COVID-19) can be impacted by a variety of reasons such as pre-processing of the data. The choice of a classification algorithm is another important aspect that must be considered. In our research, we

explore the impact of various famous deep learning-based models and their impact on the COVID-19 tweet classification. The results are examined with several bi-directional deep learning algorithms.

4.1 Experimental Setup

Experimentation was performed on a computer with a 3.1 GHz Intel core i7 11th gen CPU, 16 GB of DDR 4 RAM, and an SSD of 512 GB. A python programming tool Spyder was utilized for the design and implementation of the model. Spyder was developed by the company named "Spyder project contributors" and it is an open-source development ecosystem for the python programming language.

4.2 Evaluation Metrics

Accuracy, recall, and precision criteria, as well as the F1 measure, were used to evaluate our model in this study. These metrics are comparable to those used in previous studies. The following formulae can be used to determine these values in binary classification tasks through the use of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

$$Precision = \frac{TP}{FP + TP} \tag{1}$$

$$Recall = \frac{TP}{FN + TP} \tag{2}$$

Nevertheless, to extend the previously mentioned evaluation matrix to multi-class issues, we offer alternate equations for precision and recall while keeping the F1 equation the same. For the following equations, 's' points to the value in the confusion matrix (for example; true positives (TP), true neutrals (TNt), and true negatives (TNg)), 'i' denotes rows, 'j' denotes columns, and 'c' denotes the class number.

$$Precision_c = S_{ii}/S_{ii} + \Sigma_{j=1 ton; i \neq j} S_{ij} \tag{3}$$

$$Recall_c = S_{ii}/S_{ii} + \Sigma_{j=1 ton; i \neq j} S_{ji} \tag{4}$$

The total precision and recall values for our model can be obtained by combining the precision and recall values for individual classes in a variety of ways. The three fundamental approaches to calculating average precision and recall are a weighted average, micro average, and macro average. We use a weighted average for our research experimentations.

F1 Measure: Recall and precision are the two most used measures for addressing the issue of class imbalance. They also serve as the basis for the F1 metric. We write it mathematically as:

$$F1 = 2 * \frac{precision * recall}{precision + recall} \tag{5}$$

4.3 Setting the Baseline

In this research, the utilized dataset has not been opted for deep learning before, as per our knowledge. Therefore, we will set a baseline through standard machine learning algorithms. To successfully compare our deep learning model's performance, we check the classification results against several ML algorithms such as RF, SVM, LR, XGB, DB, and NB, and the best-performing algorithm is used as a baseline for comparison. In our experiments, LR performed the best for COVID-19 sentiment analysis therefore it will be used as a baseline.

4.4 Classification Results

GRU and LSTM are the two algorithms that provide the best accuracy at 92.5% and 91.2% respectively and weighted F1 scores of 0.90 and 0.89 respectively. Bi-LSTM and Bi-GRU provide almost similar accuracy as their standard counterparts. Both LSTM and GRU provide the best accuracy with epoch = 5. Although the final results are provided based on epoch = 5, it is valuable to point out that the model starts to overfit the training data after epoch = 3 as the training accuracy is reaching close to 1 which is visible from Figs. 8 and 9. The experimental findings showed that applying deep learning techniques improves sentiment analysis in the COVID-19 dataset. In contrast to other techniques, the suggested model does not need any feature engineering to retrieve specific characteristics like n-gram, POS, and idea extraction. Furthermore, even with the complexity of the given data, there are notable improvements by using DL algorithms in terms of the F1-measure and the accuracy in comparison to conventional ML classifiers.
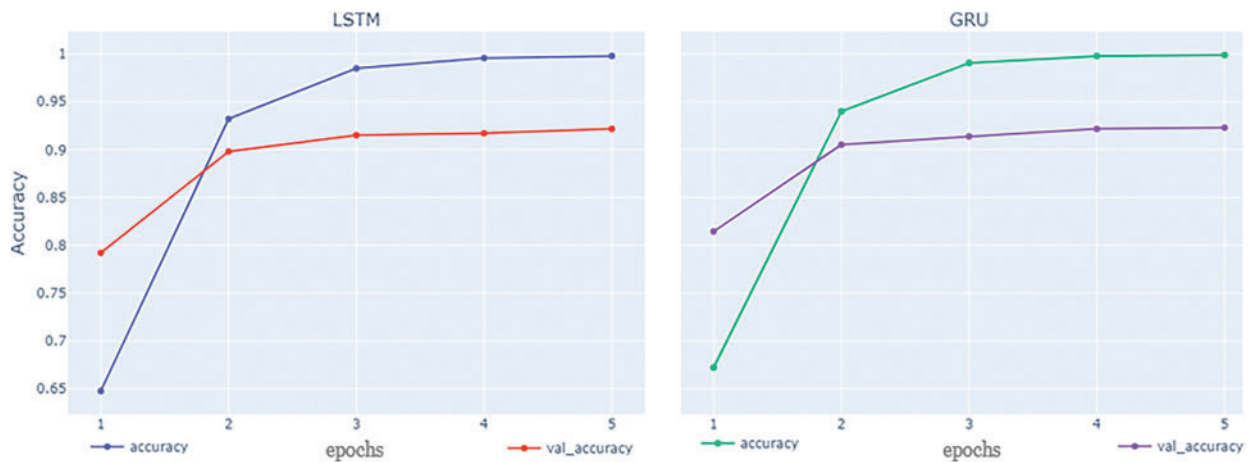
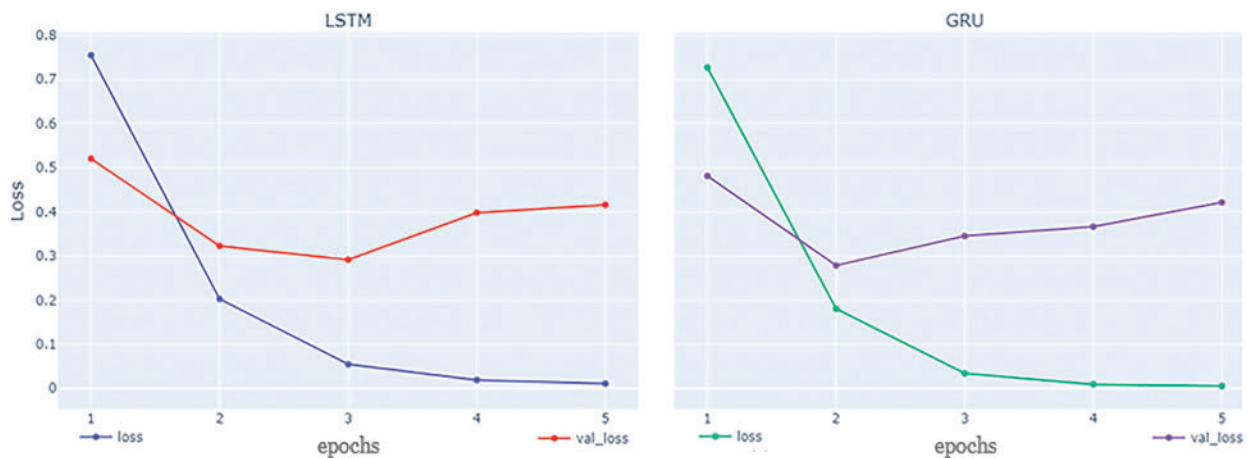**Figure 8:** LSTM and GRU accuracy

Figure 9: LSTM and GRU loss

4.5 Results Comparison: Bidirectional NN vs Standard NN vs ML

The baseline was set by measuring the performance against the LR classifier. Both standard and bi-Directional deep learning algorithms provide better results than the standard ML algorithms for the chosen dataset. The research analyzes whether bi-directional deep learning can provide better results than standard deep learning algorithms. As the results in Table 4 provide a comparison of all the algorithms used in our study. The standard deep learning algorithms are performing slightly better than bi-directional models. Standard neural networks LSTM and GRU are providing slightly better accuracy and F1 score than their bi-directional counterparts. We utilize the same parameters for all classifiers used in this study for accurate comparison.

Table 4: Classification report

	Sentiment class	Precision	Recall	F1-score
Bi-LSTM Validation accuracy = 90.5%	Negative	0.88	0.93	0.91
	Neutral	0.88	0.90	0.89
	Positive	0.95	0.87	0.91
Bi-GRU Validation accuracy = 90.3%	Negative	0.89	0.91	0.90
	Neutral	0.90	0.90	0.90
	Positive	0.90	0.87	0.89
LSTM Validation Accuracy = 91.2%	Negative	0.90	0.93	0.91
	Neutral	0.92	0.92	0.92
	Positive	0.92	0.89	0.91
GRU Validation Accuracy = 92.5%	Negative	0.90	0.95	0.93
	Neutral	0.90	0.92	0.91
	Positive	0.97	0.89	0.93
Logistic regression Validation Accuracy = 83%	Negative	0.81	0.79	0.80
	Neutral	0.85	0.86	0.86
	Positive	0.84	0.82	0.83

4.6 Data Visualization

To further visualize the impact of text cleaning, we utilize the word cloud which reveals the terms which have the highest influence in classifying a tweet text as a positive class label. Figs. 10a and 10b shows the results of preprocessing steps for positive and negative class respectively. It reveals the disparity and the outcomes of text cleaning by comparing the top 250 terms that were available in positive and negative classes after text cleaning. Normal textual data contains a lot of random words that are not useful for the task of sentiment analysis. Words such as 'https', 'co', and 'will' etc., are removed from the text. It makes sense since these words serve no purpose for the classification of a tweet. After preprocessing is performed, we can see that these words are no longer part of the text, thus increasing the quality of the text for the classification of the tweet.

Figure 10: (a) Positive class word cloud (b) Negative class word cloud

5 Conclusion

This research article discusses the design, execution, and assessment aspects of our detailed deep learning-based model. When utilizing the dataset mentioned in Section 3.2, the GRU and LSTM provide the highest accuracy of 92.5% and 91.2% respectively. In most cases, the original text is not insufficient to produce accurate results concerning classification problems due to unimportant details in the text. Therefore, it's important to include various ordered preprocessing methods in the model for better results. Class inequality is a major problem in most of the datasets, therefore it is important to deal with class inequality before classification. Resampling techniques should be a part of our model if there is a considerable difference between the classes. The proposed model utilizes random minority oversampling as it is a small-scale dataset, however, majority under-sampling can be used but it would require a large-scale dataset which can provide reduced time complexity. We also deal with the issue of limited data for multi-class classification problems by class label analysis which results in better predictions from the model while providing the same overall functionality for COVID-19-related sentiment analysis. This research also provided a comprehensive analysis through various deep learning algorithms by comparing the performance of standard and bi-directional algorithms to find the optimal algorithm for COVID-19 sentiment analysis. The research concludes that standard deep learning algorithms are faster than their bi-directional counterparts while providing the same performance. Finally, we conclude that we have contributed expressively to the area of sentiment analysis through our modular deep-based model.

In the future, we plan to explore the impact of transformers and graph methods to design and implement new approaches for sentiment analysis for imbalanced datasets while simultaneously dealing with multiclass classification problems. Transformers have been shown to perform better than standard deep learning algorithms, therefore, their performance impact must be explored.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] M. Attaullah, M. Ali, M. F. Almufareh, M. Ahmad, L. Hussain *et al.,* "Initial stage COVID-19 detection system based on patients' symptoms and chest X-ray images," *Applied Artificial Intelligence*, vol. 2022, no. 1, pp. 1–20, 2022.

[2] T. Singhal, "A review of coronavirus disease-2019 (COVID-19)," *Indian Journal of Pediatrics*, vol. 87, no. 4, pp. 281–286, 2020.

[3] H. Nishiura, N. M. Linton and A. R. Akhmetzhanov, "Serial interval of novel coronavirus (COVID-19) infections," *International Journal of Infectious Diseases*, vol. 93, no. 1, pp. 284–286, 2020.

[4] C. Villavicencio, J. J. Macrohon, X. A. Inbaraj, J. H. Jeng and J. G. Hsieh, "Twitter sentiment analysis towards covid-19 vaccines in the Philippines using naïve bayes," *Information*, vol. 12, no. 5, pp. 204–224, 2021.

[5] G. Alwakid, T. Osman, M. E. Haj, S. Alanazi, M. Humayun *et al.,* "MULDASA: Multifactor lexical sentiment analysis of social-media content in nonstandard Arabic social media," *Applied Sciences*, vol. 12, no. 8, pp. 1–18, 2022.

[6] T. K. Tran, H. M. Dinh, H. Nguyen, D. -N. Le, D. -K. Nguyen *et al.,* "The impact of the COVID-19 pandemic on college students: An online survey," *Sustainability*, vol. 13, no. 19, pp. 10762, 2021. https://doi.org/10.3390/su131910762

[7] N. Chintalapudi, G. Battineni and F. Amenta, "Sentimental analysis of COVID-19 tweets using deep learning models," *Infectious Disease Reports*, vol. 13, no. 2, pp. 329–339, 2021.

[8] L. Nemes and A. Kiss, "Social media sentiment analysis based on COVID-19," *Journal of Information and Telecommunication*, vol. 5, no. 1, pp. 1–15, 2021.

[9] R. Chandra and A. Krishna, "COVID-19 sentiment analysis via deep learning during the rise of novel cases," *Public Library of Science (PLoS) One*, vol. 16, no. 8, pp. e025561, 2021.

[10] M. E. Basiri, S. Nemati, M. Abdar, S. Asadi and U. R. Acharrya, "A novel fusion-based deep learning model for sentiment analysis of COVID-19 tweets," *Knowledge Based Systems*, vol. 228, no. 1, pp. 10742–10782, 2021.

[11] S. N. Almuayqil, M. Humayun, N. Z. Jhanjhi, M. F. Almufareh and D. Javed, "Framework for improved sentiment analysis via random minority oversampling for user tweet review classification," *Electronics*, vol. 11, no. 19, pp. 3058–3078, 2022.

[12] T. Mikolov, M. Karafiat, L. Burget, J. Honza, C. ycernocky *et al.,* "Recurrent neural network based language model," *Interspeech*, vol. 11, no. 1, pp. 2877–2880, 2011.

[13] T. Mikolov, S. Kombrink, L. Burget, J. Černocký and S. Khudanpur, "Extensions of recurrent neural network language model," in *2011 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, pp. 5528–5531, 2011.

[14] J. Leskovec, "Social media analytics: Tracking, modeling and predicting the flow of information through networks," in *Proc. of the 20th Int. Conf. Companion on World Wide Web (WWW '11)*, New York, NY, USA, Association for Computing Machinery, pp. 277–278, 2011. https://doi.org/10.1145/1963192.1963309

[15] A. Balahur, "Sentiment analysis in social media texts," in *Proc. 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Atlanta, Georgia, 120–128, 2013.

[16] Y. Wang and B. Li, "Sentiment analysis for social media images," in *Proc. of 5th IEEE Int. Conf. on Data Mining Workshop*, Atlantic, NJ, USA, pp. 1584–1591, 2016.

[17] A. Muhammad, N. Wiratunga and R. Lothian, "Contextual sentiment analysis for social media genres," *Knowledge Based Systems*, vol. 108, no. 1, pp. 92–101, 2016.

[18] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in *Proc. Fifth Int. Workshop on Natural Language Processing for Social Media*. Boston, MA, USA, 1–10, 2017.

[19] A. C. Pandey, D. S. Rajpoot and M. Saraswat, "Twitter sentiment analysis using hybrid cuckoo search method," *Information Process Management*, vol. 53, no. 4, pp. 764–779, 2017.

[20] T. Ito, K. Tsubouchi, H. Sakaji, T. Yamashita and K. Izumi, "Contextual sentiment neural network for document sentiment analysis," *Data Science and Engineering*, vol. 5, no. 2, pp. 180–192, 2020.

[21] R. Nallapati, F. Zhai and B. Zhou, "SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents," in *Proc. of Thirty-First AAAI Conf. on Artificial Intelligence (AAAI'17)*, San Francisco, California USA, pp. 1–6, 2017.

[22] G. Lu, J. Gan, J. Yin, Z. Luo, B. Li *et al.,* "Multi-task learning using a hybrid representation for text classification," *Neural Computing and Application*, vol. 32, no. 11, pp. 6467–6480, 2020.

[23] Z. Jianqiang, G. Xiaolin and Z. Xuejun, "Deep convolution neural networks for Twitter sentiment analysis," *IEEE Access*, vol. 6, no. 1, pp. 23253–23260, 2018.

[24] A. Ortis, G. M. Farinella, G. Torrisi and S. Battiato, "Visual sentiment analysis based on objective text description of images," in *Proc. of Int. Workshop on Content-Based Multimedia Indexing*, La Rochelle, France, pp. 1–7, 2018.

[25] J. Han, K. Qian, M. Song, Z. Yang, Z. Ren *et al.,* "An early study on intelligent analysis of speech under COVID-19: Severity, sleep quality, fatigue, and anxiety," 2020. [Online]. Available: http://arxiv.org/abs/2005.00096

[26] T. K. Tran, H. M. Dinh, T. T. Phan and V. E. Balas, "Building an enhanced sentiment classification framework based on natural language processing," *Journal of Intelligent & Fuzzy Systems*, vol. 43, no. 2, pp. 1771–1777, 2022. https://doi.org/10.3233/JIFS-219278

[27] J. Xu, F. Huang, X. Zhang, S. Wang, C. Li *et al.,* "Sentiment analysis of social images via hierarchical deep fusion of content and links," *Applied Soft Computing Journal*, vol. 80, no. 1, pp. 387–399, 2019.

[28] C. Shofiya and S. Abidi, "Sentiment analysis on covid-19-related social distancing in Canada using Twitter data," *International Journal of Environmental Research on Public Health*, vol. 18, no. 11, pp. 1–18, 2021.

[29] C. Villavicencio, J. J. Macrohon, X. A. Inbaraj, J. H. Jeng and J. G. Hsieh, "Twitter sentiment analysis towards covid-19 vaccines in the Philippines using naïve bayes," *Information*, vol. 12, no. 5, pp. 20–38, 2021.

[30] Z. B. Nezhad and M. A. Deihimi, "Twitter sentiment analysis from Iran about COVID 19 vaccine," *Diabetes and Metabolic Syndrome: Clinical Research and Reviews*, vol. 16, no. 1, pp. 1–20, 2022.

[31] M. O. Lwin, J. Lu, A. Sheldenkar, P. J. Schulz, W. Shin *et al.,* "Global sentiments surrounding the COVID-19 pandemic on twitter: Analysis of twitter trends," *Journal of Medical Internet Research (JMIR) Public Health Surveillance*, vol. 6, no. 2, pp. e19447, 2020.

[32] K. H. Manguri, R. N. Ramadhan and P. R. M. Amin, "Twitter sentiment analysis on worldwide covid-19 outbreaks," *Kurdistan Journal of Applied Research*, vol. 20, no. 1, pp. 54–65, 2020.

[33] M. Muller and M. Salathe, "Addressing machine learning concept drift reveals declining vaccine sentiment during the COVID-19 pandemic," 2020. [Online]. Available: http://arxiv.org/abs/2012.02197

[34] T. T. Nguyen, S. Criss, P. Dwivedi, D. Huang, J. Keralis *et al.,* "Exploring U.S. shifts in anti-Asian sentiment with the emergence of COVID-19," *International Journal on Environmental Research in Public Health*, vol. 17, no. 19, pp. 1–13, 2020.

[35] M. S. Satu, S. Uddin, M. A. Summers, J. M. W. Quinn and M. A. Moni, "TClustVID: A novel machine learning classification model to investigate topics and sentiment in covid-19 tweets," *Knowledge-Based Systems*, vol. 226, no. 1, pp. 107–126, 2021.

[36] H. Seale, A. E. Heywood, J. Leask, M. Sheel, D. N. Durrheim *et al.,* "Examining Australian public perceptions and behaviors towards a future covid-19 vaccine," *BioMed Central (BMC) Infectious Diseases*, vol. 21, no. 1, pp. 1–9, 2021.

[37] F. Rustam, M. Khalid, W. Aslam, V. Rupapara, A. Mehmood *et al.,* "A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis," *Public Library of Science (PLoS) One*, vol. 16, no. 2, pp. e0245909, 2021.

[38] J. Samuel, M. M. Rahman, G. M. Ali, Y. Samuel, A. Pelaez *et al.,* "Feeling positive about reopening? new normal scenarios from covid-19 us reopen sentiment analytics," *IEEE Access*, vol. 8, no. 1, pp. 142173–142190, 2020.

[39] J. Samuel, G. M. Ali, M. M. Rahman, E. Esawi and Y. Samuel, "COVID-19 public sentiment insights and machine learning for tweets classification," *Information*, vol. 11, no. 6, pp. 314–331, 2020.

[40] J. Xue, J. Chen, C. Chen, C. Zheng, S. Li *et al.,* "Public discourse and sentiment during the COVID 19 pandemic: Using latent dirichlet allocation for topic modeling on twitter," *Public Library of Science (PLoS) One*, vol. 15, no. 9, pp. e0239441, 2020.

[41] C. Kuldeep, M. Yadav, R. K. Rout, K. S. Sahoo, N. Z. Jhanjhi *et al.,* "Sentiment analysis with tweets behaviour in twitter streaming API," *Computer Systems Science and Engineering*, vol. 45, no. 2, pp. 1113–1128, 2023.

[42] S. N. Almuayqil, M. Humayun, N. Z. Jhanjhi, M. F. Almufareh and N. A. Khan, "Enhancing sentiment analysis via random majority under-sampling with reduced time complexity for classifying tweet reviews," *Electronics*, vol. 11, no. 21, pp. 3624–3644, 2022.