Intelligent Automation
& Soft Computing

Tech Science Press

DOI: 10.32604/iasc.2022.029511

ARTICLE

News Modeling and Retrieving Information: Data-Driven Approach

Elias Hossain[1], Abdullah Alshahrani[2] and Wahidur Rahman[3,*]

[1]Electrical & Computer Engineering, North South University, Dhaka, 1229, Bangladesh

[2]Department of Computer Science and Artificial Intelligence, College of Computer Science and Engineering, University of Jeddah, Jeddah, 21493, Saudi Arabia

[3]Department of Computer Science and Engineering, Uttara University, Dhaka, 1230, Bangladesh

*Corresponding Author: Wahidur Rahman. Email: mwrahman@uttarauniversity.edu.bd

Received: 05 March 2022 Accepted: 19 April 2022 Published: 05 February 2024

ABSTRACT

This paper aims to develop Machine Learning algorithms to classify electronic articles related to this phenomenon by retrieving information and topic modelling. The Methodology of this study is categorized into three phases: the Text Classification Approach (TCA), the Proposed Algorithms Interpretation (PAI), and finally, Information Retrieval Approach (IRA). The TCA reflects the text preprocessing pipeline called a clean *corpus*. The Global Vectors for Word Representation (Glove) pre-trained model, FastText, Term Frequency-Inverse Document Frequency (TF-IDF), and Bag-of-Words (BOW) for extracting the features have been interpreted in this research. The PAI manifests the Bidirectional Long Short-Term Memory (Bi-LSTM) and Convolutional Neural Network (CNN) to classify the COVID-19 news. Again, the IRA explains the mathematical interpretation of Latent Dirichlet Allocation (LDA), obtained for modelling the topic of Information Retrieval (IR). In this study, 99% accuracy was obtained by performing K-fold cross-validation on Bi-LSTM with Glove. A comparative analysis between Deep Learning and Machine Learning based on feature extraction and computational complexity exploration has been performed in this research. Furthermore, some text analyses and the most influential aspects of each document have been explored in this study. We have utilized Bidirectional Encoder Representations from Transformers (BERT) as a Deep Learning mechanism in our model training, but the result has not been uncovered satisfactory. However, the proposed system can be adjustable in the real-time news classification of COVID-19.

KEYWORDS

COVID-19; news retrieving; data-driven; machine learning; BERT; topic modelling

1 Introduction

With the exponential growth of unstructured textual data on the Internet, it has become necessary to categorize the text to analyze and understand essential perspectives that aid decision-making. As a result, text categorization or classification refers to assigning tags or groups of unstructured data based on its textual content [1]. This is not only makes indexing the rapidly increasing data more accessible, but it also aids in retrieving desired material from a vast information space. The advancement of textual data processing strategies has sparked research interest in recent years. These techniques are essential because textual data grows in size over time, and such techniques aid in indexing and retrieving such growing text data. Data mining is a versatile technology that can forecast future patterns and behaviour [2]. Exploring data, discovering trends in data, and making forecasts are the three basic operations of data mining. Text mining is a technique for retrieving previously hidden facts from unstructured data to discover them. This unstructured data is exponentially nowadays [3].

According to the Atlantic reports of 2016 [4], the New York Times reveals about 150 articles daily (Monday-Saturday), 250 articles on Sundays and 65 blog posts per day. In March 2019, more than 4.4 million blog posts were published every day [5]. However, there are currently many news articles about COVID-19 being published on social media, some of which are verified and contain false or misleading information [6]. These days, the Internet comprises a plethora of news from diverse categories such as athletics, technology, culture, music, and politics. Such information can be found on the Internet by every person. If consumers are interested in news relating to a specific section, they must choose the option and then view the report by clicking on it [7]. This is a time-consuming procedure. It will be a safer alternative if it is possible to view the customer's news based on his preferences. While the number of websites offering information on the Internet has grown, it has become more challenging for the user to find news of personal interest. Consequently, it is essential to categorize the news sentences so that users can quickly view them.

Machine Learning (ML) models have drawn a lot of attention in recent years. ML-related methods learn to classify data based on previous observations and understand the inherent connections between pieces of text and their labels using pre-labelled examples as training data [8]. State-of-the-art techniques in NLP are now widely used because they have revolutionized text classification research. As search engines provide vast amounts of data, it is essential to filter and classify related news in COVID-19, especially in the current global pandemic [9]. This is the key idea and motivation of this study to bring about an effective solution to classify the news reports regarding COVID-19 and find the most dominant topic in each document to obtain meaningful insights through topic modelling. This research's priority is to analyse COVID-19 news articles. These sorts of news effectively are extensively being published by various news reporters, and it is often difficult to track out the authentic information [10]. For instance, whether the economy goes down or people take such a pandemic positively has to measure to make a country step forward. While these types of pandemics may come more in the future, a pipeline has been prepared so that any meaningful message can be extracted by analyzing the behavior of the users. In this research, there are three contributions found in this proposed study:

- Computational complexity analysis has been accomplished to classify important news like COVID-19, which will play a very influential role for those who desire to work in this domain. We have developed a sequential model to classify COVID-19 news and reduce the loss function to enhance the model's accuracy.
- Topic modelling has been carried out by extracting confidential information related to COVID-19 from the dataset, which will positively contribute to the research community finding the pattern and diving into more research.
- The state-of-the-art Natural Language Processing (NLP) techniques have been developed to create a robust model applicable in real life.

The paper is organised into five interconnected sections. Section 2 gives the literature search results and reviews and associated discussion. Section 3 shows the overall research methodology and proposed system analysis. Section 4 presents the result with the corresponding relevant discussions. Finally, Section 5 illustrates the conclusion of this manuscript.

2 Literature Review

The spreading Coronavirus brings many crises worldwide and a human-made problem that disseminates inaccurate and false information. However, the feelings of an imperative need of understanding the most accurate and authentic news of this sentimental issue, the Tweets and R statistical software is used by the authors Jim et al. in the paper [11] for textual analysis with the help of two Machine Learning (ML) alignment which measures the practicability and effectiveness of Coronavirus Tweets in terms of accuracy, where a short tweet accuracy is 91% and 74% based on the Naïve Bayes method and the logistic regression method, respectively. Still, for long tweets, both ways provide comparatively weak performance. Paper [12] have applied the Natural Language Processing (NLP) literature and Deep Learning Long-Short Term Memory (LSTM) literature as word inflicting *corpus* and Recurrent Neural Network (RNN). The texts are collected from the Customer Relationship Management (CRM) system of an online ads consultancy farm from 2009 to 2020 to use the structured text data for valuable information mining and more magnificent vivisection. From the paper [13], it can be said that for detecting each context's sentiment polarity, Long Short-Term Memory (LSTM) is usually complex and more time consuming for training.

On the other hand, though the CNN and gating mechanisms are more straightforward, efficient and need less training time, it neglects each context modelling's specific representation [14]. So, the Interactive Gated Convolutional Network (IGCN) was proposed by Avinash et al. with a bidirectional gating mechanism to represent the target and resemble review context relationship, positional Information, POS tags, and domain-specific word inflicting for the sensibility of the target. SemEval 2014 datasets depict the effectiveness of the propound IGCN model. Generally, short and noisy texts are collected from online based sources. So, achieving high performance is beyond the Natural Language Processing (NLP) method, such as the probabilistic dormant semantic Bow-of-Word (BOW) model. However, to overcome this problem, the LSTM model and a word sensitive keyword vocabulary were proposed in [15], which can provide the full text's semasiology. For completing this research successfully, the IMDB and SemEval-2016 datasets were used to examine the short-text sentiment analysis, where the result came out with an accuracy of 1∼2%. An improved variant of the LSTM model known as the Gated Recurrent Unit (GRU) is also proposed to show performance. In the paper [16], the authors try to present a new model, named BLSTM and CNN, where the deep learning model is used for achieving a better result in the classification of Chinese text. The model consists of two layers of LSTM. One layer of CNN where LSTM obtains a serialized output from past to future context and CNN is for eliciting features from a visualized image with significant performance.

In the paper [17], the authors offered a method to detect false news or rumour in social media with a deep learning model and a CNN. The study also demonstrated a comparison with other related works, which mentioned that the proposed work provides laudable performance accuracy, f-measure and recall. The authors of the paper [18] described an intelligent topics classification and extracting model known as TClustVID, which can analyze the news or posts related to COVID-19 in terms of accuracy. The Twitter dataset of COVID-19 is collected from the IEEE Dataport repository and created a word-to-index dictionary by pre-processing the dataset to clean the raw data. TClustVID model represented a high performance compared to other existing methods.

This paper [19] concatenated the sentiment analysis regarding multiple sources. The authors of this study utilized various newly adopted deep learning and tracked out the efficiency of the models. The performance score was found to be satisfactory.

Based on the review of the above paper, it can be said that these have dealt only with Natural Language Processing and Traditional Classification issues (Table 1). Topic modelling of Twitter's dataset in paper [20] has been done by utilising Deep Learning's algorithm, but no news article has been classified in COVID-19. Besides, the research limitation applied the state-of-the-art NLP technique to classify the news articles. The above study has not analysed computational complexity, so this research provides a compact solution. This research includes identifying COVID-19 related news articles and topic modelling by performing the state-of-the-art NLP techniques, which will play a crucial role. We have extracted essential insights into the data through a text analytics approach and performed computational complexity analysis.

Table 1: Comparison with the previous results in terms of their methodology, strength and results

Source	Method	Strength	Accuracy/result
[11]	Naïve Bayes and Logistic regression	measures the practicability and effectiveness of Coronavirus Tweets	91% and 74%
[12]	LSTM, RNN	Information mining	The information mining has been accomplished.
[13]	LSTM	Analyze context sentiment	Detecting each context's sentiment polarity using a computational method.
[14]	LSTM, GRU	Sentiment analysis	Short-text sentiment analysis with satisfactory performance.
[15]	BI-LSTM, CNN	Text classification	Achieved significant performance with the use of concentrated methods.
[16]	CNN	Detect false news or rumour on social media	Using the deep learning method and obtained significant results.

3 Methodology

The Proposed Research Methodology (PRM) has been separated into three sections for instance, Text Classification Approach (TCA), Proposed Algorithms Interpretation (PAI), and Information Retrieval Approach (IRA). The Text Classification Approach (TCA) is further subdivided into three segments to illustrate Research Dataset, Text Preprocessing, and Features Extraction Pipeline (FEP). The Features Extraction Pipeline (FEP) is additionally classified into two portions, namely, Pre-trained Model Structure (PTMS) and Non-Pre-trained Model Structure (NPTMS). The novel Deep Learning Algorithms has been combined in the Algorithm Selection section. Fig. 1 shows the Proposed Research Methodology (PRM). The Text Classification Approach (TCA) section describes how text preprocessing was accomplished for applying to machine learning. The explanation of the practical algorithms is shown in the Proposed Algorithms Interpretation (PAI) section. Finally, the mathematical description of how confidential information was extracted from the dataset is shown in Information Retrieval Approach (IRA) section. However, the detailed illustration and consequence are demonstrated in Fig. 1.

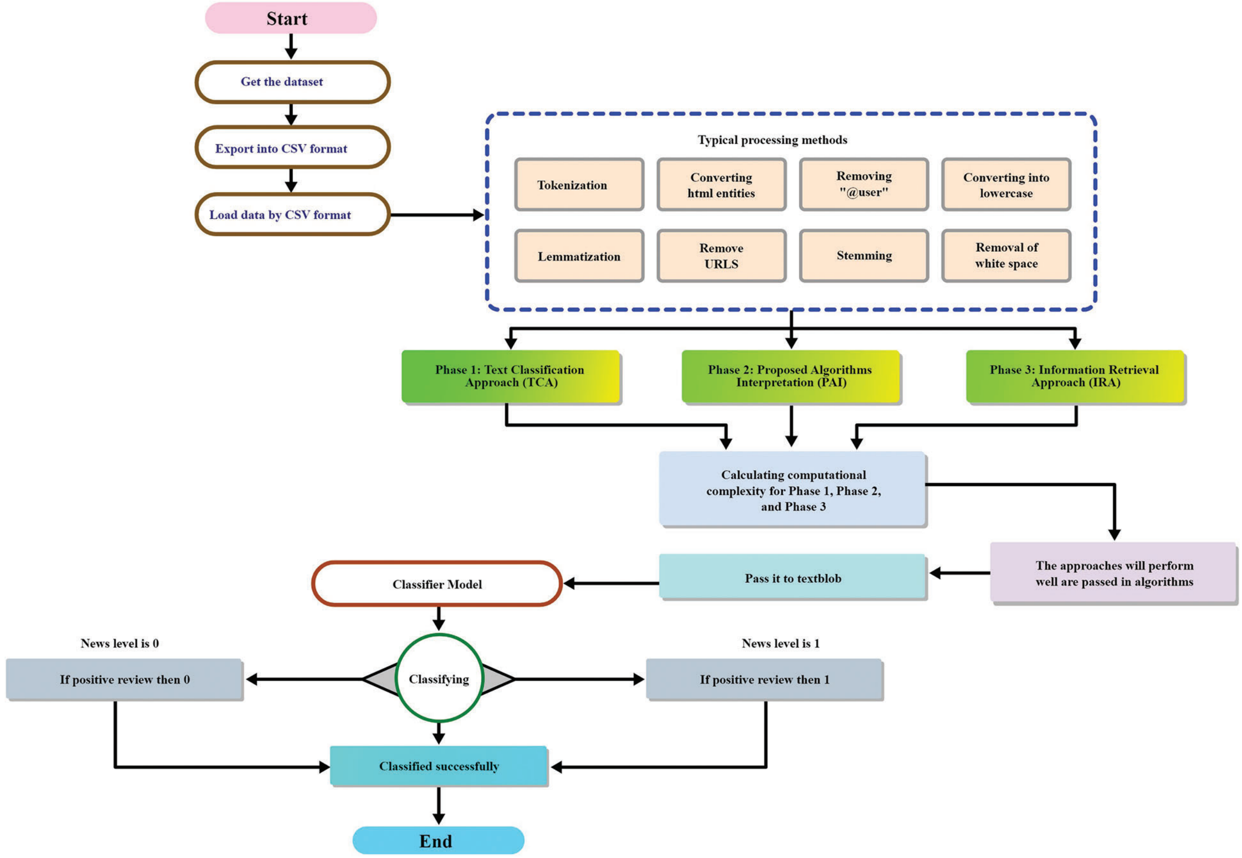
Start
Get the dataset
Export into CSV format
Load data by CSV format
Typical processing methods
Tokenization
Converting html entities
Removing "@user"
Converting into lowercase
Lemmatization
Remove URLS
Stemming
Removal of white space
Phase 1: Text Classification Approach (TCA)
Phase 2: Proposed Algorithms Interpretation (PAI)
Phase 3: Information Retrieval Approach (IRA)
Calculating computational complexity for Phase 1, Phase 2, and Phase 3
The approaches will perform well are passed in algorithms
Pass it to textblob
Classifier Model
Classifying
News level is 0
If positive review then 0
News level is 1
If positive review then 1
Classified successfully
End

Figure 1: Stage of the text classification and classifying of a news article

3.1 Dataset Preparation

The dataset that has been utilized in this research was collected from [20]. The dataset contains fake and real news regarding COVID-19. After merging all the datasets, the total sample size was 47489, of which 23948 were related to COVID-19, and 23531 were general news articles. These dataset features are classified into five parts: Label (1 or 0), headline, description, image, and source.

3.2 Corpus Preparation

NLP's biggest challenge is text pre-processing; it is suggested to implement a classification algorithm on a refined *corpus* instead of a noisy *corpus*. Otherwise, the model accuracy will be poor. The uproarious *corpus* incorporates irrelevant things inside the content. For example, incorrect spelling, numerical qualities, emoji, stop words for example ',', '!', '?', '.', '~', '||', '|' and so on. They corrupt the preparation, which gives a low exactness rate. The end of these substances from the *corpus* will upgrade the model efficiency. For example, '!', '?', etc., special characters are suitable for sentiment analysis but are not efficient for classification problems as these reduce the model's effectiveness. The following steps are followed to make a clean *corpus* in this proposed study: tokenization, stop words, capitalization, noise filtering, stemming, and lemmatization.

3.3 Features Extraction

In the case of Text data, the features must be extracted to fit the data into classification algorithms; simply put, the computer usually understands the numeric data instead of the Text data, so it is crucial to extract the features from the text to transform the text data into the form of numeric. The Features Extraction Pipeline (FEP) is categorised into two segments: Pre-trained Model Structure (PTMS) and Non-Pre-trained Model Structure (NPTMS). The PTMS illustrates the novel word embedding's approach, namely, FastText Glove and BERT. On the other hand, the NPTMS explained the standard features extraction approach: Term Frequency-Inverse Document Frequency (TF-IDF) and Bag-of-Words (BOW) model. It is to be mentioned that, before extracting the features, text preprocessing approaches such as stemming, lemmatization, stop words and tokenization was applied because in a sentence, every word does not hold semantic information, so it is required to eliminate appropriately. After that, the features extraction techniques such as NPTMS and PTMS were utilized to transform the input data into a numeric form.

3.4 Algorithm Selection

This section describes the algorithms that have been utilised in terms of COVID-19 related news classification. In this case of text classification, we have applied the traditional Machine Learning and Deep Learning algorithms. While experimenting, it was observed to look at the performance and prediction accuracy; moreover, this study found that the Deep Learning algorithms perform well on Text data, which has been employed in this research. We have received 98% and 97% accuracy in Bi-LSTM and CNN algorithms. The following Fig. 2 demonstrates the Architecture of the proposed research.

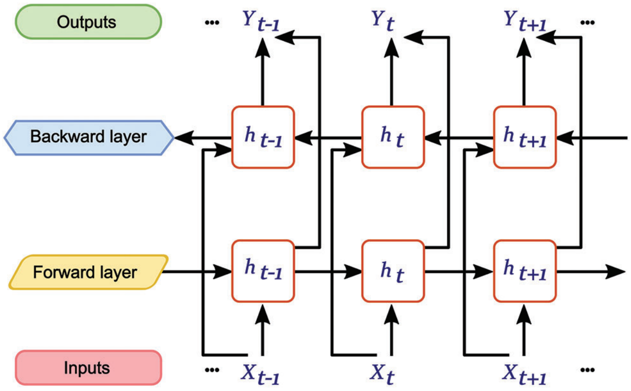

Figure 2: Architecture of the Bi-LSTM

Bidirectional Long Short-Term-Memory (Bi-LSTM):

Bi-LSTM input sequences can be in both directions with two neuron sub-layers. This orientation is to generate a complete input context. Fig. 2 shows the inputs, forward, backward and output layers. There are also backward hidden sequences, namely $\overleftarrow{h}$, $\vec{h}$. From this configuration, we can compute the output sequence *y*:

$$\vec{h}_t = \mathcal{H}\left(W_{x\vec{h}}x_t + W_{\vec{h}}\vec{h}\vec{h}_{t-1} + b_{\vec{h}}\right) \tag{1}$$

$$\overleftarrow{h_t} = \mathcal{H}(w_{x\overleftarrow{h}}x_t + w_{\leftarrow\leftarrow h}\ \overleftarrow{h}_{t+1} + b_{\overleftarrow{h}}) \tag{2}$$

$$y_t = W_{\vec{h}y}\ \vec{h}_t + \boldsymbol{w}_{\overleftarrow{h}\,y}\ \overleftarrow{h}_t + b_y \tag{3}$$

Latent Dirichlet Allocation (LDA):

We have used Latent Dirichlet Allocation (LDA) (TMA) as an Intuition Modelling subject. In Text data collection, LDA is a very interpretable and widespread architecture to evaluate topics [21]. In LDA, a complexity layer is introduced, and K is assumed, which presents a list of subjects. In a specific document, m provides probability distribution over K topics. Again, the distribution of the likelihood of each particular issue is called Vocabulary V. The formula is divided into five interconnected modules, such as LDA Total Probability, Dirichlet Distribution of Topics over Terms, Dirichlet Distribution of documents over Topics, and Prob of a Subject that appeared in a particular paper in an individual Prob of a Topic.

4 Result, Observation & Findings

The Results Analysis section is classified into four parts, for instance Empirical Consequence (EC), Model Assessment Report (MAR), Comparison of Previous Research (CPR), and Text Analytics Approach (TAA). The Empirical Consequence has illustrated the classification reports collected after building models by the traditional approach and Deep Learning (DL approach. The Model Assessment Report (MAR) interpreted and evaluated the best model that appears with satisfactory accuracy. Similarly, the loss function reports found while experimenting that explicitly narrated in the MAR section. The Comparison of Previous Research (CPR) delineated the method utilised beforehand and compared the previous study's proposed models. The Text Analytics Approach (TAA) eventually manifested the meaningful insights found and demonstrated by graphical representation.

4.1 Empirical Consequences

The accuracy of predictions from the classification algorithms is estimated by applying a classification report. The report demonstrates the precision, recall and F1-score of the key classification metrics per class. The metrics are computed by using true and false positives and true and false negatives. The metrics consist of four elements: true positive, false positive, true negative, false negative, and false negative [22]. In Table 2, the classification report of the Deep Learning algorithms and state-of-the-art NLP approach is based on the features extraction technique. Table 3 shows the traditional machine learning algorithms' classification reports on the term frequency-inverse document frequency (TF-IDF) system. Table 4 explained the classification reports applying the Bag-of-Words (BOW) model. Fig. 7 delineated the ROC curve and confusion matrix on top of the models that had excellent accuracy in the news classification of COVID-19. The following Eqs. (4)–(7) were considered for finding the precision, recall and F1-score [23].

Table 2: Classification report of the deep learning algorithms

Algorithm	For the case of "0."				For the case of "1."			
	Features extraction technique (FET)	P	R	F1	P	R	F1	Accuracy score
CNN	Text to sequence	0.98	0.98	0.98	0.98	0.98	0.98	0.98
LSTM	Text to sequence	0.98	0.97	0.99	0.97	0.98	0.99	0.98
BI-LSTM	Text to sequence	0.99	0.98	0.99	0.98	0.99	0.99	0.99
BI-LSTM	Fast Text	0.98	0.99	0.98	0.97	0.98	0.97	0.98
BI-LSTM	Glove	0.99	0.99	0.99	0.99	0.99	0.99	0.99
BERT	Bert (FET)	0.98	0.95	0.97	0.95	0.99	0.97	0.97

Table 3: Classification report of the traditional machine learning algorithms (TF-IDF-Uni-Gram approach)

Algorithm	For the case of "0."				For the case of "1."			
	Features extraction technique (FET)	P	R	F1	P	R	F1	Accuracy score
DTC	TF-IDF	0.98	0.98	0.98	0.98	0.98	0.98	0.98
RFC	TF-IDF	1.0	0.97	0.98	0.97	1.0	0.96	0.98
KNN	TF-IDF	0.85	0.99	0.92	0.98	0.83	0.90	0.91
MNB	TF-IDF	0.95	0.98	0.96	0.98	0.95	0.96	0.96
GB	TF-IDF	0.99	0.96	0.98	0.96	0.99	0.98	0.98
LR	TF-IDF	0.99	0.97	0.98	0.97	0.99	0.98	0.98
SVM	TF-IDF	1.0	0.94	0.97	0.95	1.0	0.97	0.97

Table 4: Classification report of the traditional machine learning algorithms (BOW approach)

Algorithm	For the case of "0."				For the case of "1."			
	Features extraction technique (FET)	P	R	F1	P	R	F1	Accuracy score
DTC	BOW	0.98	0.98	0.98	0.98	0.98	0.98	0.98
RFC	BOW	1.0	0.97	0.98	0.97	1.0	0.98	0.98
KNN	BOW	0.68	0.98	0.80	0.96	0.55	0.70	0.76
MNB	BOW	0.85	0.97	0.91	0.97	0.84	0.90	0.90
GB	BOW	0.99	0.95	0.97	0.95	0.99	0.97	0.97
LR	BOW	0.99	0.98	0.99	0.98	0.99	0.99	0.99
SVM	BOW	1.0	0.85	0.92	0.87	1.0	0.93	0.93

Precision: The ratio of the model's true positive estimate to the total (correct and incorrect) positive estimate. It is articulated as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

Recall/Sensitivity: The ratio of being able to predict as positive. It is given in mathematical form as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

F1-score: This is the harmonic mean for Precision and Recall and provides a better estimate than the Accuracy Metric of the wrongly classified instances. It is given, mathematically, as:

$$F1 = 2 \cdot \frac{\text{Precision.Recall}}{\text{Precision} \ + \ \text{Recall}} \tag{6}$$

Accuracy: It is the measure of all the instances correctly predicted. It is given as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

4.2 Model Evaluation

The Model Assessment Report (MAR) is organised by several approaches, such as the ROC (Receiver Operating Characteristics)- AUC (Area Under Curve), Test Accuracy of Models, Test Loss of Models, and Confusion Matrix. It should be noted here that the model assessment of this section has been performed on top of the algorithm that gives the best accuracy. We found a satisfactory score by using the Bi-LSTM with the Glove approach. The MAR in this section shows how well our proposed model performs in this particular work.

Fig. 3 illustrates the ROC-AUC curve on Bi-LSTM with Glove pre-trained model. Another method of determining how good the performance of different classification models is is the ROC-AUC curve. ROC stands for Receiver Operating Curve, and AUC stands for Area Under Curve. The higher the ROC value, i.e., closer to 1, the better our model. This performance evaluation indicator tends to plot based on TPR against the FPR based on the various threshold values. By looking at Fig. 3, it can be clearly described that the value of AUC is close to almost 1 (0.99), so our model can admirably distinguish between all the Positive and the Negative class points correctly.

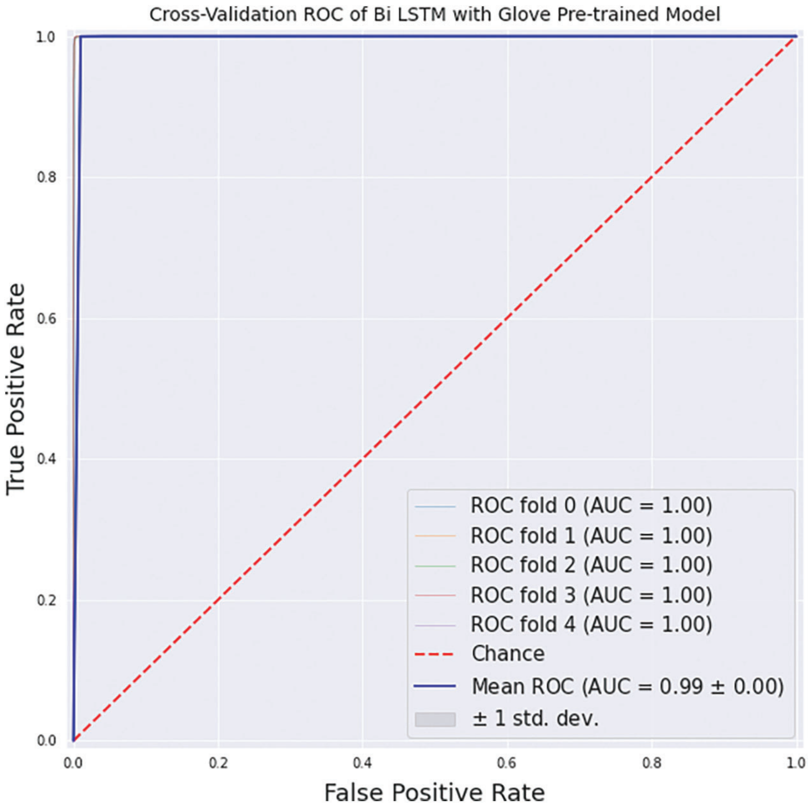

Figure 3: ROC-AUC curve on Bi-LSTM with Glove

On the other hand, Fig. 4 exhibits the Confusion Matrix on Bi-LSTM with Glove. The Confusion Matrix determines the accuracy of all types of classification algorithms. The Confusion Matrix consists of four types of values: TP, FP, TN, and FN.

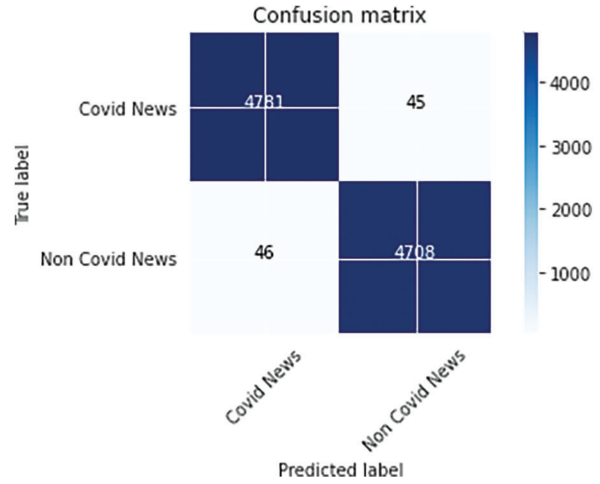

Figure 4: Confusion Matrix on Bi-LSTM with Glove

Figs. 5 and 6 demonstrate the test and loss accuracy of the models that have been considered for classifying the COVID-19 news articles. The number of epochs is similar to the number of iterations if the batch size is the entire training dataset. Figs. 5 and 6 show some variation in accuracy at the par epoch, and the loss function fluctuates.

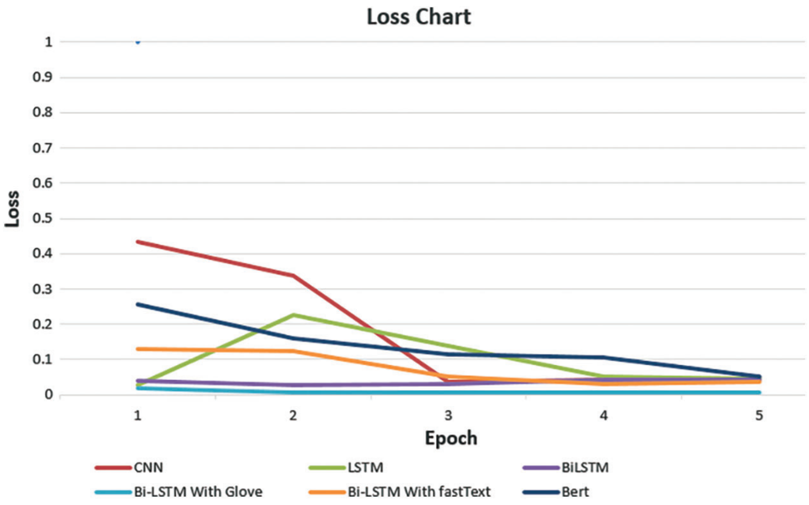

Figure 5: Test accuracy of the models

4.3 *Measuring the Complexity*

The computational complexity of machine learning is a mathematical analysis of the possibilities for productive learning by computer. It works within recently applied machine inference models centring on computational complexity theory and has a clear focus on successful and general learning algorithms. Table 5 shows the bound of the algorithms for dense data. We have the following approximations by naming n is considered to be the experimental instances, p is recognized as a features parameter, $\boldsymbol{n}_{\text{trees}}$ is the number of characteristics (for methods based on various trees), $\boldsymbol{n}_{sv}$ is the ratio of support vectors and $\boldsymbol{n}_{li}$ is the total percentage of neurons at layer $\boldsymbol{i}$ in the neural network.

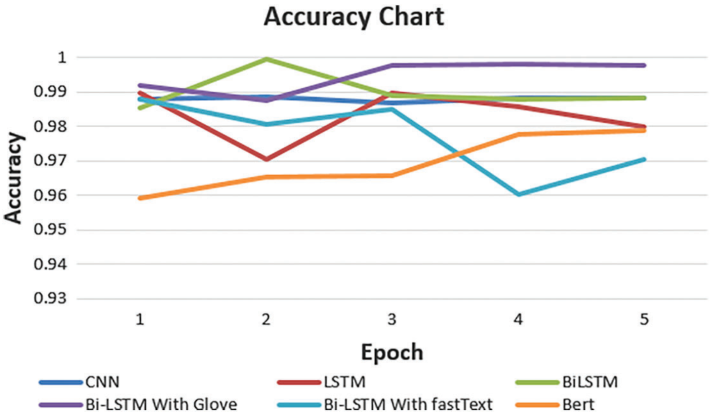

Figure 6: Test loss of the models

Table 5: Measure the computational complexity of machine learning algorithms

Algorithm	Classification/Regression	Training	Prediction
Decision tree	C + R	$O(n^2p)$	$O(p)$
Random forest	C + R	$O(n^2pn_{\text{trees}})$	$O(pn_{\text{trees}})$
Gradient boosting ($\boldsymbol{n}_{\text{trees}}$)	C + R	$O(npn_{\text{trees}})$	$O(pn_{\text{trees}})$
SVM (Kernel)	C + R	$O(n^2p + n^3)$	$O(n_{sv}p)$
k-nearest neighbours	C + R	–	$O(np)$
Neural network	C + R	–	$O(pn_{l_1} + n_{l_1}n_{l_2} + \ldots)$
Naïve Bayes	C	$O(np)$	$O(p)$

4.4 Comparative Analysis

The Comparison of the Previous Research (CPR) section illustrated the previous study's performance and compared it with the proposed model's accuracy. Here should mention that we focused on the algorithms, features extraction method, and accuracy found beforehand. Our primary focus is on Deep Learning based algorithms that have been utilised over the years in the context of text classification; however, in this case, we did not find the enormous study carried out for classifying the news article in terms of COVID-19. The previous research was concentrated on Fake News classification to mitigate the misleading information on social media. Table 6 shows the comparative analysis of the previous study. We have compared our work with the top seven recent research related to text classification and showed accuracy. The paper whose accuracy is less than our model has been written "No", the details description is shown in Table 6.

Table 6: Comparative analysis of previous research

References	Algorithm	Feature extraction method	Dataset size	Accuracy	Proposed model's accuracy (Yes/No/Equal)
[10]	NB	–	N/A	91%	No
[11]	CNN	Text to Sequence	10000	87%	No
[13]	DECR-Bi-GRU-CNN	Text to Sequence	329242	89.67%	No
[14]	LSTM	Text to Sequence	800000	93%	No
[15]	MLSM LSTM	One Hot Encoding	31000	88%	No
[16]	Bi-LSTM	Text to Sequence	25935	96%	No
[17]	IGCN	–	6000	81%	No
Our proposed model	BI-LSTM with Glove	Text to Sequence, Glove	14012	99%	

4.5 Information Retrieval

Topic Modeling Approach (TMA) is a systemic way to classify objects existent in a text document and extract secret patterns displayed by a text *corpus* [24]. It can be applied for several purposes, for instance, document clustering, feature selection, information retrieval from unstructured data, etc. The Latent Dirichlet Allocation (LDA) illustrates a topic model and is applied to categorise text into a specific topic in a document. Fig. 7 shows the coherence score based on the topic number for Latent Dirichlet Allocation (LDA). The details visualization and meaningful insight are demonstrated in Fig. 7.

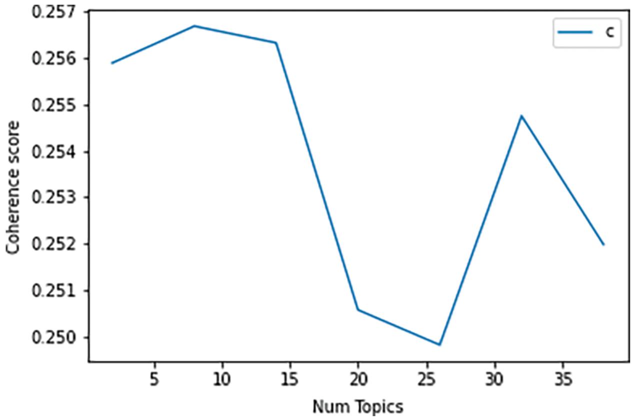

Figure 7: Coherence score for Latent Dirichlet Allocation (LDA)

Table 7 shows the most representative document for each topic, and by looking at Table 6, topic number and topic per contribution can be discovered.

Table 7: Most representative document for each topic

Document No.	Topic number	Topic per contribution	Keywords
0	0.0	0.9858	', ,, , i, s, a, o, n, e, l
1	1.0	0.9898	', a, , ,, s, e, r, t, n, i
2	2.0	0.9896	', ,, , e, a, s, p, i, d, l
3	4.0	0.9548	', ,, e, t, , i, c, a, n, l
4	5.0	0.9896	', , ,, e, a, c, s, i, o, r
5	10.0	0.9877	', s, a, , ,, t, e, r, o, h
6	12.0	0.9842	', a, , ,, n, i, s, r, t, e
7	14.0	0.9889	', , ,, a, r, e, s, t, o, i
8	15.0	0.9886	', i, , ,, r, s, o, t, a, n
9	16.0	0.9898	', , ,, e, d, i, n, o, a, l
10	17.0	0.9886	', , ,, e, a, l, r, i, s, n
11	18.0	0.9881	', ,, , o, e, s, a, d, I, r

5 Conclusion

The World Health Organization (WHO) has repeatedly stated that the virus may last for years. Therefore, effort must be given to ensure only accurate information is placed online to ensure people's safety in COVID-19. This proposed study used a predictive model using the Bidirectional Long Short-Term Memory (Bi-LSTM) with Glove pre-trained model, which can classify the news reports regarding COVID-19 with 98% accuracy. This research extracts hidden patterns by modelling topics using the Latent Dirichlet Allocation (LDA) algorithm. Computational complexity analysis has also been shown in this beneficial research. In future work, we will analyze the sentiment to find out what kind of difficulties people are facing with COVID-19 globally and identify the related factors responsible for many reasons.

Acknowledgement: None.

Funding Statement: This research received no external funding.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Elias Hossain, Wahidur Rahman; data collection: Abdullah Alshahrani; dataset analysis and interpretation of results: Elias Hossain, Abdullah Alshahrani; draft manuscript preparation: Elias Hossain, Wahidur Rahman. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Data will be made available upon request.

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

References

[1] K. Kowsari, K. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes *et al.,* "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, pp. 150, 2019.

[2] M. Robinson, “How many stories do newspapers publish per day?” *The Atlantic*, 2016. [Online]. Available: https://www.theatlantic.com/technology/archive/2016/05/how-many-stories-do-newspapers-publish-per-day/483845/ (accessed on 06/02/2022)

[3] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu *et al.,* “Deep learning–based text classification: A comprehensive review,” *ACM Computing Surveys (CSUR)*, vol. 54, no. 9, pp. 1–40, 2020.

[4] N. Al Mudawi, N. Beloff and M. White, “Developing a framework of critical factors affecting the adoption of cloud computing in government systems (ACCE-GOV),” in *Intelligent Computing*, London, UK: Springer, pp. 520–538, 2022.

[5] H. P. Wu, Y. L. Liu and J. W. Wang, “Review of text classification methods on deep learning,” *Computers, Materials & Continua*, vol. 63, no. 3, pp. 1309–1321, 2020. https://doi.org/10.32604/cmc.2020.010172

[6] J. Samuel, G. Ali, M. Rahman, E. Esawi and Y. Samuel, “COVID-19 public sentiment insights and machine learning for tweets classification,” *Information*, vol. 11, pp. 314–330, 2020.

[7] Ş. Ozan, “Case studies on using natural language processing techniques in customer relationship management software,” *Journal of Intelligent Information Systems*, vol. 56, no. 2, pp. 233–253, 2021.

[8] A. Kumar, V. Narapareddy, V. Srikanth, L. Neti and A. Malapati, “Aspect-based sentiment classification using interactive gated convolutional network,” *IEEE Access*, vol. 8, pp. 22445–22453, 2020.

[9] F. Hu, L. Li, Z. Zhang, J. Wang and X. Xu, “Emphasizing essential words for sentiment classification based on recurrent neural networks,” *Journal of Computer Science and Technology*, vol. 32, no. 4, pp. 785–795, 2017.

[10] Y. Li, X. Wang and P. Xu, “Chinese text classification model based on deep learning,” *Future Internet*, vol. 10, no. 11, pp. 113, 2018.

[11] A. Alsaeedi and M. Al-Sarem, “Detecting rumors on social media based on a CNN deep learning technique,” *Arabian Journal for Science and Engineering*, vol. 45, no. 12, pp. 10813–10844, 2020.

[12] M. Satu, M. Khan, M. Mahmud, M. Uddin, S. Summers *et al.,* “TClustVID: A novel machine learning classification model to investigate topics and sentiment in COVID-19 tweets,” *Knowledge-Based Systems*, vol. 226, pp. 107126, 2021.

[13] F. Abid, C. Li and M. Alam, “Multi-source social media data sentiment analysis using bidirectional recurrent convolutional neural networks,” *Computer Communications*, vol. 157, no. 1, pp. 102–115, 2020.

[14] H. Kapoor, “COVID-19 Indian news headlines kaggle,” 2020. [Online]. Available: https://www.kaggle.com/hkapoor/covid19-india-news-headlines-for-nlp?fbclid=IwAR0i-OKgSmP9rmIpTKhZQJD3ZQ8opJa4YUybmD8q9pJ2UCAWfR39iXU0vM (accessed on 06/02/2022)

[15] A. M. Morgan, “COVID-19 Public Media Dataset by Anacode,” 2022. [Online]. Available: https://www.kaggle.com/code/ahmedmohammedmorgan/covid-19-public-media-dataset-by-anacode (accessed on 06/02/2022)

[16] K. Spirovski, E. Stevanoska, A. Kulakov, Z. Popeska and G. Velinov, “Comparison of different model’s performances in task of document classification,” in *Proc. of the 8th Int. Conf. on Web Intelligence, Mining and Semantics*, Novi Sad, Serbia, 2018.

[17] J. Pennington, R. Socher and C. D. Manning, “GloVe: Global vectors for word representation,” in *Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, pp. 1532–1543, 2014.

[18] K. S. Jones, “A statistical interpretation of term specificity and its application in retrieval,” *Journal of Documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[19] T. Tokunaga, T. Tokunaga and I. Makoto, “Text categorization based on weighted inverse document frequency,” *Egorization Based On Weighted Inverse Document Frequency*, Information Processing Society of Japan, SIGNL, vol. 94, no. 100, pp. 33–40, 1994.

[20] P. Neogi, A. Das, S. Goswami and J. Mustafi, “Topic modeling for text classification,” *Advances in Intelligent Systems and Computing*, vol. 937, pp. 395–407, 2020.

[21] J. Mandal and D. Bhattacharya, *Emerging Technology in Modelling and Graphics,* vol. 937. Singapore: Springer, 2020.

[22] S. Visa, B. Ramsay, A. Ralescu and E. VanDerKnaap, “Confusion matrix-based feature selection,” *22nd Midwest Artificial Intelligence and Cognitive Science*, vol. 710, pp. 120–127, 2011.

[23] R. Yacouby and D. Axman, “Probabilistic extension of precision, recall, and F1 score for more thorough evaluation of classification models,” in *Proc. of the First Workshop on Evaluation and Comparison of NLP Systems*, pp. 79–91, 2020. https://aclanthology.org/2020.eval4nlp-1.9/ (accessed on 06/02/2022)

[24] B. Grün and K. Hornik, “Topicmodels: An R package for fitting topic models,” *Journal of Statistical Software*, vol. 40, no. 13, pp. 1–30, 2011.