Explainable Text Classification Model for COVID-19 Fake News Detection

Mumtahina Ahmed[1], Mohammad Shahadat Hossain[2], Raihan Ul Islam[3], and Karl Andersson[3]*

[1]Department of Computer Science and Engineering
Port City International University, Chittagong, Bangladesh
mumtahina.ahmed.cs@gmail.com

[2]Department of Computer Science and Engineering
University of Chittagong, Chittagong, Bangladesh
hossain_ms@cu.ac.bd

[3]Department of Computer Science, Electrical and Space Engineering
Luleå University of Technology, Skellefteå, Sweden
{raihan.ul.islam, karl.andersson}@ltu.se

Received: March 28, 2022; Accepted: May 6, 2022; Published: May 31, 2022

Abstract

Artificial intelligence has achieved notable advances across many applications, and the field is recently concerned with developing novel methods to explain machine learning models. Deep neural networks deliver the best performance accuracy in different domains, such as text categorization, image classification, and speech recognition. Since the neural network models are black-box types, they lack transparency and explainability in predicting results. During the COVID-19 pandemic, Fake News Detection is a challenging research problem as it endangers the lives of many online users by providing misinformation. Therefore, the transparency and explainability of COVID-19 fake news classification are necessary for building the trustworthiness of model prediction. We proposed an integrated LIME-BiLSTM model where BiLSTM assures classification accuracy, and LIME ensures transparency and explainability. In this integrated model, since LIME behaves similarly to the original model and explains the prediction, the proposed model becomes comprehensible. The performance of this model in terms of explainability is measured by using Kendall's tau correlation coefficient. We also employ several machine learning models and provide a comparison of their performances. Therefore, we analyzed and compared the computation overhead of our proposed model with the other methods because the model takes the integrated strategy.

Keywords: fake news, COVID-19, Explainable AI, LIME, BiLSTM

1 Introduction

There are around 3.6 billion users [1] on social media like Facebook, Twitter, Instagram, and Pinterest, where users generate massive amounts of posts to express their opinion. Social media ensure faster and easier communication but have some disadvantages, like the proliferation of fake news. With the advent of Fourth Industrial Revolution(4IR), fake information spreads rapidly through various online platforms, which is known as Infodemic. The novel coronavirus pandemic had been the most newsworthy event last

Journal of Internet Services and Information Security (JISIS), volume: 12, number: 2 (May), pp. 51-69
DOI:*10.22667/JISIS.2022.05.31.051*

*Corresponding author: Forskargatan 1, Campus Skellefteå, A building, Skellefteå, Sweden, Tel.: +46-910-585364

Table 1: Example Tweet for COVID-19 related fake news classification.

Tweet	Label
Some coronavirus patients are experiencing chronic fatigue.	real
The vaccine against the new coronavirus has existed since 2001.	fake

year. Social media users actively discussed news related to the pandemic which simultaneously provokes public rumor and misinformation [2]. During this COVID-19 situation, fake news can influence people to take extreme measures that may result in a severe outcome. Therefore, both physical and mental health endangered as the public faces restless anxiety or fear due to fake-news [3]. While the quality and reliability of news are not verified, false information affects society negatively [4]. Constraint 2021 workshop organizes a shared task [5] to combat online hostile posts that spread rapidly during an emergency. Subtask-1 focuses on COVID-19 pandemic-related Fake News Detection in English posts. We utilize a dataset consisting of 10,700 online posts, manually annotated and released by the organizer [6]. Given a collection of real news and fake claims surfaced on social media, the task is to classify the post being real or fake, as presented in Table 1.

Explainable AI interprets a machine learning model in such a way where the actions can be clearly understood by humans. Goodman and Flaxman [7] stated that Explainable AI offers increased transparency and fairness along with prediction accuracy in comparison to the human counterpart. The method provides an interpretable domain, including all factors and associations to prove the prediction of an algorithm whether it is fair and ethical [8]. The term 'explanation' refers to a statement, enabling something clear as well as provide reasoning and justification in favor of an action and belief [9]. The research fields of Explainable AI can be characterized into three domains, namely, opaque system, interpretable system, and comprehensible system. The opaque system provides no details about underlying algorithmic mechanisms. The interpretable system presents a level of understanding to analyze algorithms. The comprehensible system emits symbols to enable user-driven explanation and relate input with the output [9].

Deep neural networks with multiple layers have recently been a highly successful and popular research topic in machine learning due to their remarkable performance in several benchmark problems. For example, earthquake prediction [10], facial Expression recognition [11] [12], hand gesture recognition [13], [14], image processing [15], [16], speech synthesis [17], etc. are the applications that utilized deep neural network. However, the Explanation of a machine-learning algorithm or deep neural network reveals all the mathematical operations or parameters in the process of decision making [18]. Analyzing explainable methods in deep neural architectures makes the internal system more transparent while maintaining high-performance accuracy. We employ the Bidirectional Long Short Term Memory model to classify the COVID-19 pandemic-related text as real or fake news. Then we utilize the Local Interpretable Model-Agnostic Explanations (LIME) method that creates a local interpretable model around prediction to faithfully comprehend the prediction of the BiLSTM neural network [19]. Further, we evaluate the performance of the method using Kendall's tau correlation coefficient.

The organization of the rest of the paper is as follows: In Section 2, we discuss several related works. We present our explainable methodology for fake news classification in Section 3. Section 4 demonstrates results and comparative analysis of our method with other models. A discussion on the explainability is illustrated in Section 5. Finally, in Section 6, we conclude the outcome of this research.

2 Related Work

Constraint AAAI 2021[1] organized a competition to analyze COVID-19 related social media posts during 2020. They considered two tasks, namely subtask-1 and subtask-2. Subtask-1 focuses on COVID-19 related fake news detection in the English language, while subtask-2 focuses on hostile posts detection in the Hindi language. Patwa et al. [5] released a manually annotated benchmark dataset on COVID-19 related fake news and experimented with four machine learning approaches. However, deep neural networks are also worth exploring in the experiment.

Sharif et al. [20] used Term Frequency-Inverse Document Frequency (TF-IDF) and word embedding features. They employed various models such as SVM, CNN, BiLSTM, and combined CNN-BiLSTM using TF-IDF and Word2Vec embedding features. SVM with TF-IDF features achieved the best accuracy on subtask-1 of the Constraint AAAI 2021 shared task. However, CNN and BiLSTM have not obtained satisfactory results. The reason for this it requires the use of ensemble techniques like attention mechanisms within their framework.

Felber et al. [21] utilized several linguistic features, such as n-grams, readability, and emotional tone. They used linear SVM, Random Forest, Logistic Regression, Naive Bayes, and Multilayer Perceptron. Among all the approaches, linear SVM achieved the highest performance. However, the experiment doesn't explore the capability of deep learning models. Hence the performance among traditional machine learning models and deep learning models can't be compared.

Glazkova et al. [2] proposed COVID-Twitter-BERT (CT-BERT) approach that is a transformer-based ensemble model. They experimented with Bidirectional Encoder Representations from Transformers, Robustly Optimized BERT Pre-training Approach, and COVID-Twitter-BERT models. They added extra data such as COVID-19 Healthcare Misinformation Dataset, and Multilingual Cross-domain Fact Check News Dataset for COVID-19 to improve the performance. However, extra data can not show any benefits to enhance CT-BERT model performance.

Bang et al. [22] proposed a robust model for fake-news classification by employing two approaches. They utilized fine-tuned transformers-based models with robust loss functions and applied influence calculation to remove harmful training instances. However, the robust loss functions do not help much in improving the F1-score on the fake news dataset.

Shushkevich et al. [23] constructed ensembles of Bidirectional LSTM, SVM, Logistic Regression, and Naive Bayes models for fake news classification. Since they didn't train the neural network model at an advanced level, their model couldn't classify fake messages effectively.

Koloski et al. [24] extracted hand-crafted features, such as n-grams of character, word-based features, and captured relevant patterns through a latent space representation. They used the Shapley Additive exPlanations (SHAP) method to obtain the most important features. They also employed multiple BERT-based models to learn the contextual information. However, the experiment can not evaluate the performance of the model regarding interpretability and explainability.

Ayoub et al. [25] introduced an explainable model using the ensemble of transformer model DistilBERT and SHAP for fake news classification. They utilized 984 fact-checking claims and performed data augmentation using back-translation. However, concerning explanation and trust, there was no significant difference between TE (Text+SHAP Explanation and TSESE (Text+SHAP) Explanation+Source and Evidence).

Yang et al. [26] proposed an explainable system XFake for fake news classification that helps users to interpret fake news. They designed the MIMIC, ATTN, and PERT framework considering both attributes

[1] https://constraint-shared-task-2021.github.io/

Table 2: Related Works on Fake News Detection

Ref. No.	Description	Model	Limitation
[5]	Released a benchmark dataset on COVID-19 related fake news and experimented with four machine learning baselines.	SVM, LR, GDBT, DT	Do not explore deep learning models.
[20]	Employed tf-idf, and word embedding features on various models including SVM, CNN, BiLSTM, and CNN+BiLSTM.	SVM	CNN and BiLSTM can not obtain a satisfactory result.
[21]	Applied classical machine learning algorithms with several linguistic features.	SVM	Do not explore deep learning models.
[2]	Proposed COVID-Twitter-BERT (CT-BERT), a transformer-based ensemble model.	COVID-Twitter-BERT	Extra data can not show any benefits to enhance CT-BERT model performance.
[22]	Utilized fine-tuned transformers-based models with robust loss functions	RoBERTa-large	The robust loss functions can not improve performance on fake news dataset.
[23]	Constructed ensembled model utilizing of Bidirectional LSTM, SVM, Logistic Regression, and Naive Bayes approaches.	SVM + LR+ NB + biLSTM	Can not explain the difference in performance.
[24]	Employed multiple BERT-based models to learn the contextual information. Extracted hand-crafted features and utilized the SHAP method to obtain most important features.	BERT-based models, SHAP	Can not evaluate the model regarding interpretability and explainability.
[25]	Proposed an explainable model using the ensemble of transformer model DistilBERT and SHAP.	DistilBERT, SHAP	Can not explain the significant difference between TE (Text+SHAP Explanation) and TSESE (Text+SHAP).
[26]	Proposed an explainable system XFake for fake news classification, that helps users to interpret fake news	MIMIC, ATTN, PERT	Human evaluation requires more time to interpret explanation.
[27]	Presented the FakeNewsTracker system that introduces a data collection approach and utilized several semantic features	LSTM	Can't be evaluated in terms of explainability.
[28]	Developed a deep hierarchical framework that captures check-worthy explainable sentences and user comments	dEFEND	Can't provide an explanation based on words or feature importance.
[29]	Proposed novel features and utilized the combination of features on the XGB model.	XGB, SHAP	Can not tackle all forms of fake news.

and statements to identify fake news effectively. The proposed frameworks generate explanations along with related examples and visualization that assist the interpretation. However, they conducted a human evaluation that requires more time to review and interpret explanation results.

Shu et al. [27] presented a system named FakeNewsTracker that introduced a data collection approach to accumulate data. They employed the LSTM model and Singular Value Decomposition (SVD) approach by utilizing several semantic features. They also developed a software interface for interactive visualization of the result. However, the model can't be evaluated in terms of explainability.

Shu et al. [28] developed a deep hierarchical framework named dEFEND that captured check-worthy explainable sentences and user comments. The framework consists of four components. Those are the sentence-comment co-attention network, news content, user comment encoder, and fake news prediction component. However, the method can't provide an explanation based on words or feature importance.

Reis et al. [29] proposed novel features for fake news classification and utilized the combination of features on the Extreme Gradient Boosting (XGB) model. They investigated the feature informativeness to generate simple models and explained the predictions for fake news stories. However, the method can not tackle all forms of fake news that require constructing a single solution through an ensemble of classifiers.

A brief description of the previous study on fake news classification is presented in Table 2. The methods mostly employed either machine learning approaches or transformer-based models. Some of them proposed explainable systems or explained the classification utilizing the SHAP explanation method. However, none of the approaches applied the LIME explanation technique with the deep neural network. Hence, they lack the capability of demonstrating transparency and comprehensibility in detecting fake news. In this work, we integrate BiLSTM neural network with the LIME method to introduce a comprehensible text classification model.

3 Methodology

COVID-19 is the most severe public health problem, and because of the virus's ability to propagate via its conveyor, it is rapidly spreading to nearly every part of the world [30]. Several researchers have used deep neural networks for Covid-19 research. For example, Ahmed et al. [30] combined data and knowledge-driven methodologies in a single framework to assess a COVID-19 patient's survival chances. While this study [31] uses neural networks to develop an integrated model for estimating the number of confirmed cases in Bangladesh. Deep neural networks based on CNN and RNN models have recently been widely explored to learn language patterns [32], [33]. CNN is capable of learning global and higher-level features through max-pooling from successive convolutional windows. However, CNN is not capable of sequential learning, while RNN deals with time-series and sequential data. RNN considers the order of sequences and long-term dependencies while learning from higher-level information. In this dissertation, We employed the LIME approach with the BiLSTM network to construct a comprehensible fake news classification model. The Bidirectional LSTM is an extended LSTM network that attains the knowledge of past and future events concurrently and learns the entire context for better results. We evaluate the LIME-Bi-LSTM model on the benchmark dataset for the fake news detection task. The overview diagram of the method is represented in Figure 1, including different phases: data preprocessing, transformation, classification, and evaluation. In the following sections, we present a detailed description of the method.

3.1 Dataset Description

The shared task (Subtask-1) of the Constraint 2021 workshop released a fake news dataset containing 10,700 manually annotated social media posts in English. The aim is to analyze news articles of real and

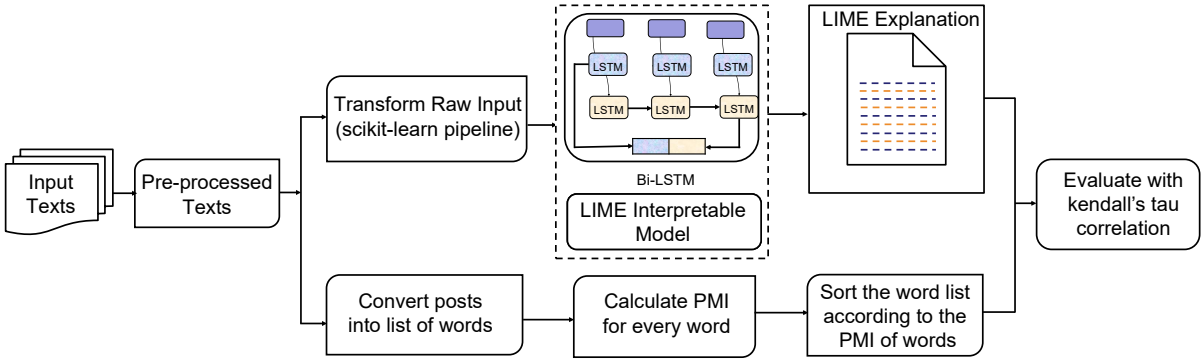

Figure 1: Overview for COVID-19 related fake news detection.

Table 3: The Constraint 2021 COVID-19 Fake News dataset [6].

Dataset Split	Real Data	Fake Data	Total Data
Training Data	3360	3060	6420
Validation Data	1120	1020	2140
Test Data	1120	1020	2140
All	5600	5100	10700

fake news across online platforms or popular press releases and combat the COVID-19 infodemic. There is a consistent balance of class label distribution with 52.34% real news and 47.66% fake news [6]. The COVID-19 fake news dataset is distributed with a ratio of 3:1:1 for training, validation, and test sets. Table 3 illustrates the distribution of the dataset into training, validation, and test sets by taking into account the ratio mentioned.

3.2 Data Preprocessing

Data preprocessing is performed to clean and preprocess the input tweet data to fit our model. We remove all non-English characters to simplify and standardize our text. URLs and links usually refer to any uniform resource on the web that doesn't provide any necessary information. Therefore, we remove these links using regular-expressions (regex), transform the text into lowercase and perform tokenization. However, some common words in English, like "the," "a," etc., don't contribute much information in a sentence. We utilize the NLTK [34] package to filter and avoid stopwords. Numbers and punctuation might not add much information while dealing with texts. Therefore, we excluded these by utilizing regex and the string module.

3.3 LIME Approach for BiLSTM Neural Network

Recurrent Neural Network is a chain-like neural structure that analyzes the current input x_s and the earlier output h_{s-1} from the hidden state to process time series and sequential data [35]. In the architecture of RNNs, the long-range gap between consequent time steps rise the vanishing gradient problem. To overcome this, Long Short Term Memory (LSTM) uses memory blocks with a gating mechanism that remembers information and enables learning long-term dependency for sequential data [36]. Recently LSTM models have achieved popularity among researchers as the hidden states extract better temporal feature representations without any loss of information from sequential input [37]. The LSTMs analyze

only the earlier context to predict the current status that may not capture valuable information. The improved bidirectional LSTM network processes the input in both forward and backward order to retrieve the past as well as subsequent context. We used a pipeline object to handle the text data conveniently by using the LIME library. The pipeline uses two transformer objects to transform a list of raw input texts into a form suitable for the BiLSTM network. The first transformer object converts the input texts into lists of indices. The second one performs padding and cropping the sequence to harmonize input text length. We have utilized the maxlen value of 100 that specifies the maximum length of each text sequence. After that, we simply chain the preprocessed input texts to our Bidirectional LSTM model that measures a word representation using both previous and subsequent information. The forward hidden vector sequence $\overrightarrow{h} = (h_1, h_2, ..., h_T)$ that processes the input in the standard order, and backward hidden vector sequence $\overleftarrow{h} = (h_T, h_{T-1}, ..., h_1)$ that processes input in reverse order [32]. The following equations iterating from time step *t = 1 to T* produces output sequence vector $d = (d_1, d_2, ..., d_T)$.

$$\overrightarrow{h_t} = \sigma(W_{f\overrightarrow{h}} f_t + W_{\overrightarrow{h}\overrightarrow{h}} \overrightarrow{h}_{t+1} + b_{\overrightarrow{h}})$$
$$\overleftarrow{h_t} = \sigma(W_{f\overleftarrow{h}} f_t + W_{\overleftarrow{h}\overleftarrow{h}} \overleftarrow{h}_{t+1} + b_{\overleftarrow{h}})$$
$$d_t = W_{\overrightarrow{h}d} \overrightarrow{h_t} + W_{\overleftarrow{h}d} \overleftarrow{h_t} + b_d$$

Next, we transfer the output feature vector d to Multilayer Perceptron, which reflects the context of the sequential input. We train an MLP network that generates a feature vector representation x and fed to the *Sigmoid* layer. A smooth S-shaped curve is known as a sigmoid. It has a value range of 0 to 1. This is simple to comprehend and apply, and it is expressed for the classification of fake news by,

$$f(x) = \frac{1}{1 + exp(-x)}$$

Given a set of training data (xi, pi), we train our model to minimize the cross-entropy loss error. The binary cross-entropy / log loss is a common loss function for binary classifications like ours. Where y denotes the label (1 for real news, 0 for false news), and p (y) denotes the anticipated probability of real news for all N points. It adds log(p(y)) to the loss for each real news (y=1), i.e. the log chance of it being real. For each fake news (y=0), it adds log(1-p(y)), that is, the log chance of it being fake.

$$Lp(q) = -\frac{1}{N}\sum_{j=1}^{N} y_i.\log(p(y_i)) + (1 - y_i).\log(1 - p(y_i))$$

Next, We utilized the LIME approach that explains the prediction of a BiLSTM neural network for fake news detection and produces an interpretable explanation for a prediction [19]. LIME locally approximates the Bi-LSTM model "l" by a linear explanation model "m" that has an interpretable domain. The model generates an identical interpretable vector $x' \in \{0,1\}^{d'}$ in the feature space for a given sample $x \in R^d$. Then it randomly cancels coordinates or x' using uniform distribution to generate N artificial points z'_i. Then a simple explanation model K-LASSO (i.e. a logistic regression) minimizes the dissimilarity function to generate the predictions of "l".

$$\mathcal{L}(l, m, \Pi_x) = \sum_{z,z'} \Pi_x(z)[l(z) - m(z')]^2$$

The objective is to minimize $\mathcal{L}$ or maximize local fidelity through investigating the Π_x proximity between random samples and original samples.

3.4 Pointwise Mutual Information (PMI)

Next, we estimate the Pointwise Mutual Information (PMI) [38] score that estimates the association of a word for a corresponding class label as,

$$PMI(w,c) = \log \frac{P(w,c)}{P(w)P(c)}$$

where $P(w,c)$ is the number of posts having word w for class c, $P(w)$ is the number of posts with word w, and $P(c)$ is the number of posts for the class c, accordingly normalized by the total number of posts N. It generates a list of words sorted according to their score where a higher PMI score indicates a stronger correlation between the word and the corresponding class.

3.5 Evaluation of Explanation

We evaluate our comprehensible classification method considering the list of most important words for classification generated by LIME and the list of words sorted according to PMI scores described as in algorithm1. We now have a list of words determined by LIME (*lime_list*) to be the most important contributors to the document's classification, as well as a list of terms in the document arranged by PMI scores (*pmi_list*) in descending order. The next step is to assess the LIME model's performance using the two lists. While the parameter provided in the LIME method determines the number of words in the LIME list, the PMI list contains all of the words in the documents. In our classification task LIME provides the twenty most important words for each post. Next, we estimate the PMI score for each word of a post. Firstly, keeping the same order, we extract the words that are included in the LIME list by traversing the PMI List and store the words in another list (*interim_list*) [39]. Then, we estimate Kendall's tau correlation value (between *lime_list* and *interim_list*) to measure the performance.

Algorithm 1 Calculation Process of Evaluating the Explanation of LIME

```
procedure EVALUATELIME
    lime_list ← [word1, word2, word3,....,word10]
    pmi_list ← [word1, word2, word3,....,wordN]
    word ← pmi_list
    loop:
    if word in lime_list and word not in interim_list then
        interim_list ← word
        goto loop.
        close;
    correlation ← kendalltau(lime_list, interim_list)
    close;
```

3.6 Other Approaches

Several research areas regarding sequential data such as text or audio data analysis, widely adopted and utilized Recurrent Neural Networks. We employed different variants of RNN based models such as GRU, LSTM, and simple RNN in our classifier to train the COVID-19 dataset [40]. We employ tf-idf feature extraction technique and experiment with machine learning algorithms. A pipeline is created for each of the classifiers, i.e., Multinomial Naive Bayes [41], and Random Forest Classifier [42]. We define a grid of parameters that we utilize for parameter fine-tuning. After that, we select the model using GridSearchCV [43], and the model is fit for training the COVID-19 dataset.

4 Experiments and Results

We conduct the experiments on Google Colab that supports GPU with Python= 3.7.10. Besides, we implement the deep learning model using keras = 2.2.4 with Tensorflow = 2.3.1 framework that facilitates the model with a faster and parallel computation. We have used a standard Bi-LSTM network with a dropout rate of 0.1 that randomly disabled 10% of neurons. Three dense layers having respectively 500, 50, and 1 unit were also included in this network. We perform hyperparameter-tuning to acquire the parameters that best fit the network by utilizing the grid search technique. Therefore, we tune different hyperparameters such as activation function, optimizer, hidden units in BiLSTM layers, and several batch sizes to train the model according to Table 4.

In $k-$fold cross-validation, subsequent k iterations of training and validation are performed. A different fold of the data is held out for validation during each iteration and for learning purposes, the remaining $k-1$ folds are used. We experimented with up to 5 fold cross-validation using the training and validation set to avoid overfitting or selection bias [44]. The highest average accuracy of 94.01% and standard deviation of 0.49 was observed while using a batch size of 64, Adam optimizer, and Sigmoid activation function with 128 hidden units in the BiLSTM layer. Sigmoid is a differentiable function with positive derivatives mostly used in shallow neural networks [45]. Adam optimizer is mostly used for problems with huge data or parameters as it is computationally efficient and consumes little memory [46]. Moreover, we utilized Adam with the Sigmoid function to optimize the model. Finally, We evaluated the model on the fake news test dataset utilizing several evaluation metrics. We estimated accuracy, the weighted average of precision, recall, and F_1-score against the real and fake classes. Moreover, we demonstrate the comparative performance analysis of our BiLSTM method with other approaches in Table 5.

According to the results, we observe that the BiLSTM method obtained the highest accuracy of 94.25% compared to LSTM, GRU, RNN that obtained an accuracy score of 93.32%, 92.99%, and 92.71% respectively. Baseline models such as SVM and Dicision Tree (DT) [5] obtained 93.46% and 85.23%, accuracy, followed by Shushkevich et al. [23] with 94.00% accuracy. In comparison, Naive Bayes, and Random Forest classifiers reported significantly inferior performance with 90.56%, and 93.03% accuracy scores, respectively. Next, we employ the LIME to explain the predictions of the BiLSTM neural network model that obtained the best performance accuracy for the fake news classification task. We calculate the computation overhead of our proposed LIME-BiLSTM model, such as CPU usage, memory usage, and computation delay. Therefore, we analyzed and compared those of the other ones in Table 6 because the proposed method takes the integrated strategy. The percentage of CPU usage per second is 3% for our proposed method, which is the lowest compared to other models. We further estimated the memory usage of models, and the proposed model utilized 17.1% memory which occupies 2340 MB of memory. LIME-LSTM utilized the highest 2392 MB of memory compared to LIME-Naive Bayes, which used the lowest 1170 MB of memory. Finally, our proposed method demonstrated the lowest computation delay of 2.86 s, while the highest 6.53 s was observed for the

Table 4: Neural Network Parameters

Batch Size	**Units**	**Optimizer**	**Activation Function**	**Accuracy**
32	64	Nadam	Relu	83.76
32	64	Nadam	Sigmoid	93.60
32	64	Adam	Relu	84.53
32	64	Adam	Sigmoid	93.59
64	128	Nadam	Sigmoid	93.45
64	128	Nadam	Relu	92.94
64	128	Adam	Relu	92.73
64	128	Adam	Sigmoid	94.01
128	256	Nadam	Relu	76.18
128	256	Nadam	Sigmoid	92.68
128	256	Adam	Relu	92.13
128	256	Adam	Sigmoid	92.85

Table 5: Experimental results on COVID-19 Fake News detection.

Method	**Accuracy**	**Precision**	**Recall**	**F_1-Score**
BiLSTM	**94.25**	**94.27**	**94.25**	**94.25**
LSTM	93.32	93.56	93.32	93.30
GRU	92.99	92.99	92.99	92.99
CNN + BiLSTM [20]	92.01	92.01	92.01	92.01
RNN	92.71	92.92	92.71	92.71
SVM [5]	93.46	93.48	93.46	93.46
DT [5]	85.23	85.31	85.23	85.25
Shushkevich et al. [23]	-	-	-	94.00
Naive Bayes	90.56	90.84	90.56	90.52
Random Forest	93.03	93.05	93.04	93.04

LIME-GRU model.

We further analyze the performance of our classification model utilizing another benchmark dataset

Table 6: Comparison of Computation Overhead on COVID-19 Fake News detection.

Method	**CPU Usage (% per sec)**	**Memory Usage (%)**	**Memory Usage (MB)**	**Computation Delay (sec)**
LIME-BiLSTM	3.0	17.1	2340	2.86
LIME-LSTM	4.5	17.7	2392	5.67
LIME-GRU	4.5	17.0	2290	6.53
LIME-RNN	4.0	16.9	2200	6.04
LIME-Naive Bayes	83.6	9.8	1170	4.13
LIME-Random Forest	5.0	12.2	1466	3.41

Table 7: Experimental results on WNUT COVID-19 tweet dataset.

Method	Accuracy	Precision	Recall	F_1-Score
BiLSTM	89.40	89.56	89.40	89.40
LSTM	89.00	89.02	89.00	88.99
GRU	87.55	87.59	87.55	87.53
RNN	88.5	88.72	88.50	88.46
DSC-IITISM [48]	87.15	83.43	91.21	87.30
CSECU-DSG [49]	81.98	81.55	82.42	82.90
Naive Bayes	73.88	74.01	73.88	73.72
Random Forest	73.09	73.42	73.10	73.11

Table 8: Comparison of Computation Overhead on WNUT COVID-19 tweet dataset.

Method	CPU Usage (% per sec)	Memory Usage (%)	Memory Usage (MB)	Computation Delay (sec)
LIME-BiLSTM	4.0	16.9	2323	5.22
LIME-LSTM	4.5	18.7	2544	8.77
LIME-GRU	3.5	17.5	2370	3.82
LIME-RNN	5.5	12.0	1627	3.71
LIME-Naive Bayes	4.5	11.2	1089	4.67
LIME-Random Forest	4.1	15.4	1891	5.70

from WNUT-2020 shared-task to identify COVID-19 informative tweets. The task-2 dataset of WNUT-2020 consists of 7000 training, 1000 validation, and 2000 test dataset. The dataset contains informative and uninformative COVID-19 tweets [47]. The experimental results are represented in Table 7. The comparison displays the significant performance of the BiLSTM model over the other methods. Therefore, we analyzed and compared the CPU usage, memory usage, and computation delay of our proposed method on the WNUT COVID-19 tweet dataset and compared it with that of the other models in Table 8. The percentage of CPU usage per second is 4% for our proposed method, while the lowest 3.5% was obtained for the LIME-GRU model. We further estimated the memory usage of models, and the proposed model utilized 16.9% memory which occupies 2323 MB of memory. LIME-LSTM utilized the highest 2544 MB of memory compared to LIME-Naive Bayes, which used the lowest 1089 MB of memory. Finally, our proposed method demonstrated the computation delay of 5.22 s, while the highest 8.77 s was observed for the LIME-LSTM model.

4.1 Kendall's Tau Correlation Coefficient

We compute Kendall's Tau correlation coefficient [50] to evaluate the performance of the comprehensible model. The coefficient considers the ranks of two non-parametric data samples to indicate correlation. Let, $\chi = x_1 ... x_n$ be a set of words to be ranked, π and σ denotes two distinct ordering of χ. Here, $T(\pi\sigma)$ is the minimum number of transpositions that are necessary to arrange π and σ, and n is the number of words (i.e.items) [51]. Then, Kendall's τ is defined as:

$$\tau = 1 - \frac{2T(\pi\sigma)}{n(n-1)/2}$$

Table 9: Comparison of correlation value and p-value

Method	Kendall's tau Correlation	p-value
LIME-BiLSTM	**0.35**	**0.16**
LIME-LSTM	0.19	0.39
LIME-GRU	0.14	0.44
LIME-RNN	0.08	0.49

The values of correlation range from -1 (strong disagreement) to 1 (strong agreement) [52]. Here, we generate two ordering of words for each test data, the PMI list, and the LIME list, sorted according to the significance score. Then we utilize Kendall's Tau to measure feature importance rank correlation. Since LIME provides the most important words ordered according to significance score, we use Kendall's tau metrics for automatic evaluation of information ordering [53].

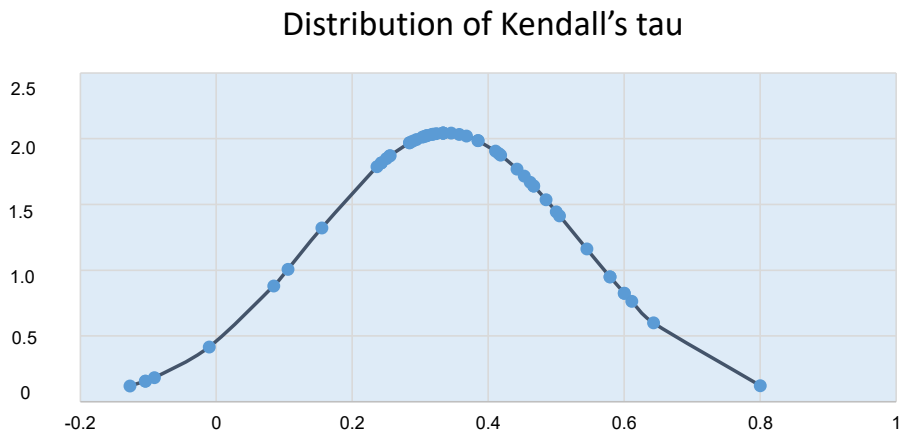

Figure 2: Distribution of Kendall's tau Correlation Coefficient.

We call Kendall's tau method from the Scipy library in python and compute the coefficient for 50 random explanations. The distribution of Kendall's tau correlation for the LIME-BiLSTM method with an average value of 0.35 and p-value of 0.16 is presented in Figure 2. The graph shows that the values of Kentall's tau are concentrated in the region of 0.3 to 0.4, indicating that there is no strong alignment of relative ordering of words in the LIME list and words in the PMI ranked word list. The length of the document may be one of the factors that influences the interpreter's performance. Another observation is that short comments contain more high-value of Kentall's tau than large ones. The alignment between LIME and PMI should be worse in longer comments because the LIME model is a document-specific interpreter while the PMI is a global assessment. However, we have employed different RNN based models, e.g. BiLSTM, LSTM, GRU, and RNN coupling with LIME as an explanatory mechanism. Then, we calculate the value of Kendall's tau correlation and the p-value for the coefficient to examine the effect of LIME for the RNN based models. The comparative correlation value and p-value for the RNN based methods are presented in Table 9.

5 Discussion

We provide a normalized confusion matrix to analyze the performance of the BiLSTM method in Figure 3. Here, 7% of real posts are wrongly classified as fake, and 4.4% of fake posts are wrongly predicted as real. Next, we explore more to interpret the explanations of individual predictions. For an individual text from the test dataset, we get an ensemble of explanation feature words that have the most impact on prediction. Therefore, the method displays the corresponding weights of the features according to Figure 4. Here, "impact", "dwarf", and "climate" accordingly have large positive weights, followed by "prince", "crisis" which have negative weights.

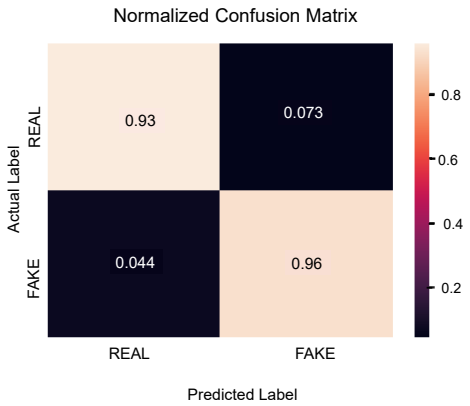

Figure 3: Normalized confusion matrix.

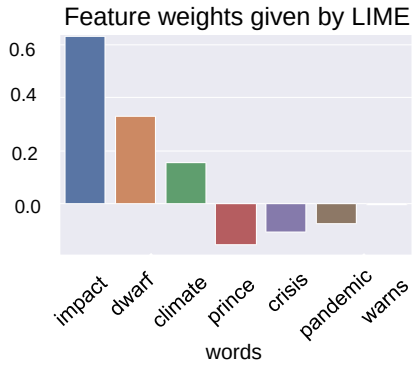

Figure 4: Feature weights given by LIME.

Figure 5 is the explanation for a test data. Here, Negative (blue) words indicate fake, while positive (orange) words indicate real. LIME interprets the weights by applying them to the prediction probabilities. Therefore, by removing the words impact from the post, we expect that the classifier will predict the post as real with a probability of 0.92 - 0.63 = 0.29.

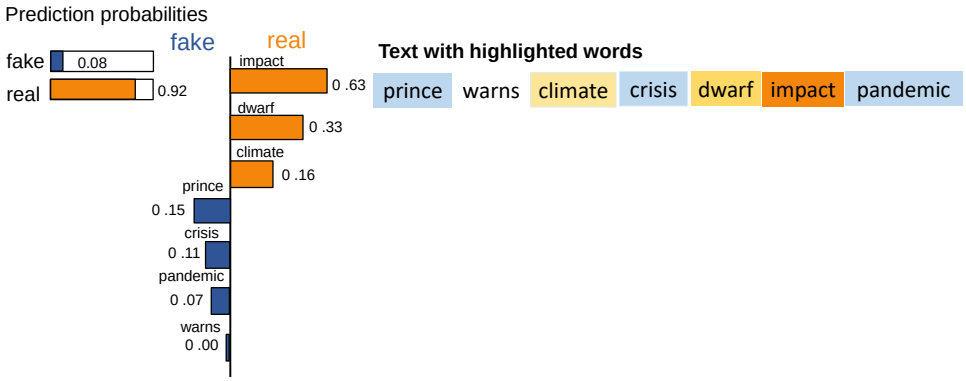

Figure 5: LIME Explanation on COVID-19 Fake News Prediction.

6 Conclusion

We utilize the BiLSTM neural network for COVID-19 fake news detection in this research. After that, we employ the LIME approach that creates a local interpretable model to explain the prediction of the neural network model [19]. Therefore, LIME has an interpretable domain that comprehend the BiLSTM model prediction for individual text data. Our proposed model achieved 94.25% accuracy and outperformed several other machine learning approaches. However, we measure the Pointwise Mutual Information (PMI) [38] score that estimates the association of a word for a corresponding class label. Finally, we use Kendall's tau correlation to measure the performance of LIME compared to PMI. There is more scope for future work to find an accurate baseline to evaluate the performance. In the future, more investigation is necessay to calculate the document-specific metric, that is, how much a word contributes to a document's classification. Measuring every word's contribution in a global scope for the classification is another way to evaluate performance. Standard evaluation metrics may be proposed to evaluate the interpretability of several explanation methods.

Acknowledgments

We thank M.S.H., R.U.I., and K.A. for their contributions to conception and methodology. M.S.H. and R.U.I. for their insightful comments on the paper. Our fellow researcher, M.A., contributed valuable insight and knowledge that substantially assisted the research. We also want to thank Sustainable Computing Lab members for their helpful recommendations on domain-specific challenges.

References

[1] S. Panke. Social media and fake news - aace, April 2019. https://www.aace.org/review/social-media-and-fake-news/ [Online; Accessed on May 15, 2022].

[2] A. Glazkova, M. Glazkov, and T. Trifonov. g2tmn at constraint@aaai2021: Exploiting ct-bert and ensembling learning for covid-19 fake news detection. arXiv:2012.11967, January 2021. https://arxiv.org/abs/2012.11967.

[3] J. Xiong, O. Lipsitz, F. Nasri, L. M. W. Lui, H. Gill, L. Phan, D. Chen-Li, M. Iacobucci, R. Ho, A. Majeed, and R. S. McIntyre. Impact of covid-19 pandemic on mental health in the general population: A systematic review. journal of affective disorders, December 2020. https://www.sciencedirect.com/science/article/pii/S0165032720325891?via%3Dihub [Online; Accessed on May 15, 2022].

[4] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu. Fake news detection on social media: A data mining perspective. *SIGKDD Explorations Newsletter*, 19(15):22–36, September 2017.

[5] P. Patwa, M. Bhardwaj, V. Guptha, G. Kumari, S. Sharma, S. PYKL, A. Das, A. Ekbal, M.S. Akhtar, and T. Chakraborty. Overview of constraint 2021 shared tasks: Detecting english covid-19 fake news and hindi hostile posts. In *Proc. of the 2021 Combating Online Hostile Posts in Regional Languages during Emergency Situation(CONSTRAINT'21), Dublin, Ireland*, volume 1402 of *Communications in Computer and Information Science*, pages 42–53. Springer, Cham, April 2021.

[6] P. Patwa, S. Sharma, S. Pykl, V. Guptha, G. Kumari, M. S. Akhtar, A. Ekbal, A. Das, and T. Chakraborty. Fighting an infodemic: Covid-19 fake news dataset. arXiv:2011.03327, November 2021. https://arxiv.org/abs/2011.03327.

[7] B. Goodman and S. Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57, October 2017.

[8] R. Chimatapu, H. Hagras, A. Starkey, and G. Owusu. Explainable ai and fuzzy logic systems. In *Proc. of the 7th International Conference on Theory and Practice of Natural Computing (TPNC'18), Dublin, Ireland*, volume 11324 of *Lecture Notes in Computer Science*, pages 3–20. Springer, Cham, November 2018.

[9] D. Doran, S. Schulz, and T. R. Besold. What does explainable ai really mean? a new conceptualization of perspectives. arXiv:1710.00794, October 2017.

[10] M. H. A. Banna, T. Ghosh, M. Jaber M. J. A. Nahian, k. A. Taher, M. Shamim M.S. Kaiser, M. Mahmud, M. S. Hossain, and K. Andersson. Attention-based bi-directional long-short term memory network for earthquake prediction. *IEEE Access*, 9:56589–56603, April 2021.

[11] T. U. Ahmed, S. Hossain, M. S. Hossain, R. U. Islam, and K. Andersson. Facial expression recognition using convolutional neural network with data augmentation. In *Proc. of the 8th Joint International Conference on Informatics, Electronics & Vision (ICIEV'19) and 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR'19), Spokane, Washington, USA*, pages 336–341. IEEE, October 2019.

[12] T. U. Ahmed, M. N. Jamil, M. S. Hossain, K. Andersson, and M. Hossain. An integrated real-time deep learning and belief rule base intelligent system to assess facial expression under uncertainty. In *Proc. of the 9th Joint International Conference on Informatics, Electronics & Vision (ICIEV'20) and 4th International Conference on Imaging, Vision & Pattern Recognition (icIVPR'20), Kitakyushu, Japan*, pages 1–6. IEEE, August 2020.

[13] N. Basnin, L. Nahar, and M. S. Hossain. An integrated cnn-lstm model for micro hand gesture recognition. In *Proc. of the 2021 Intelligent Computing and Optimization (ICO'21), Pattaya, Thailand*, volume 1324 of *Advances in Intelligent Systems and Computing*, pages 379–392. Springer, Cham, February 2021.

[14] M. Z. Islam, M. S. Hossain, R. U. Islam, and K. Andersson. Static hand gesture recognition using convolutional neural network with data augmentation. In *Proc. of the 8th Joint International Conference on Informatics, Electronics & Vision (ICIEV'19) and 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR'19), Spokane, Washington, USA*, pages 324–329. IEEE, October 2019.

[15] N. Basnin, L. Nahar, and M. S. Hossain. An integrated cnn-lstm model for bangla lexical sign language recognition. In *Proc. of the 2021 International Conference on Trends in Computational and Cognitive Engineering (TCCE'21), UTHM, Malaysia*, volume 1309 of *Advances in Intelligent Systems and Computing*, pages 695–707. Springer Singapore, October 2021.

[16] T. U. Ahmed, M. S. Hossain, M. J. Alam, and K. Andersson. An integrated cnn-rnn framework to assess road crack. In *Proc. of the 22nd International Conference on Computer and Information Technology (ICCIT'19), Dhaka, Bangladesh*, pages 1–6. IEEE, December 2019.

[17] S. N. Zisad, M. S. Hossain, and K. Andersson. Speech emotion recognition in neurological disorders using convolutional neural network. In *Proc. of the 13th International Conference, Brain Informatics(BI'20), Padua, Italy*, volume 12241 of *Lecture Notes in Computer Science*, pages 287–296. Springer-Verlag, September 2020.

[18] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. Explaining explanations: An approach to evaluating interpretability of machine learning. arXiv:1806.00069, May 2018. https://doi.org/10.48550/arXiv.1806.00069.

[19] M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you?": Explaining the predictions of any classifier. In *Proc. of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (KDD'16), San Francisco, California, USA*, pages 1135–1144. ACM, August 2016.

[20] O. Sharif, E. Hossain, and M. M. Hoque. Combating hostility: Covid-19 fake news and hostile post detection in social media. arXiv:2101.03291, January 2021. https://arxiv.org/abs/2101.03291.

[21] T. Felber. Constraint 2021: Machine learning models for covid-19 fake news detection shared task. arXiv:2101.03717, January 2021. https://arxiv.org/abs/2101.03717.

[22] Y. Bang, E. Ishii, S. Cahyawijaya, Z. Ji, and P. Fung. Model generalization on covid-19 fake news detection. arXiv:2101.03841, January 2021. https://arxiv.org/abs/2101.03841.

[23] E. Shushkevich and J. Cardiff. Tudublin team at constraint@aaai2021 – covid19 fake news detection. arXiv:2101.05701, January 2021. https://arxiv.org/abs/2101.05701.

[24] B. Koloski, T. Stepi˘snik-Perdih, S. Pollak, and B. ˘Skrlj. Identification of covid-19 related fake news via neural stacking. In *Proc of the 2021 Combating Online Hostile Posts in Regional Languages during Emergency Situation (CONSTRAINT'21), Online*, volume 1402) of *Communications in Computer and Information Science*, pages 177–188. Springer International Publishing, February 2021.

[25] J. Ayoub, X. J. Yang, and F. Zhou. Combat covid-19 infodemic using explainable natural language processing models. arXiv:2103.00747, March 2021. https://arxiv.org/abs/2103.00747.

[26] F. Yang, S. K. Pentyala, S. Mohseni, M. Du, H. Yuan, R. Linder, E. D. Ragan, S. Ji, and X. Hu. Xfake: Explainable fake news detector with visualizations. In *Proc. of the 2019 World Wide Web Conference (WWW'19), San Francisco, California, USA*, pages 1–4. ACM, May 2019.

[27] K. Shu, D. Mahudeswaran, and H. Liu. Fakenewstracker: A tool for fake news collection, detection, and visualization. *Computational and Mathematical Organization Theory*, 25(1):60–71, March 2019.

[28] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu. Defend: Explainable fake news detection. In *Proc. of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'19), Anchorage, Alaska, USA*, pages 395—-405. ACM, July 2019.

[29] J. C. S. Reis, A. Correia, F. Murai, A. Veloso, and F. Benevenuto. Explainable machine learning for fake news detection. In *Proc. of the 10th ACM conference on web science (WebSci'19), Boston, Massachusetts, USA*, pages 17—-26. ACM, June 2019.

[30] T. U. Ahmed, M. N. Jamil, M. S. Hossain, R. U. Islam, and K. Andersson. An integrated deep learning and belief rule base intelligent system to predict survival of covid-19 patient under uncertainty. *Cognitive computation*, 14(2):660–676, December 2021.

[31] S. N. Zisad, M. S. Hossain, M. S. Hossain, and K. Andersson. An integrated neural network and seir model to predict covid-19. *Algorithms*, 14(3):94, March 2021.

[32] M. Ahmed, A. N. Chy, and N. K. Chowdhury. Incorporating hand-crafted features in a neural network model for stance detection on microblog. In *Proc. of the 6th International Conference on Communication and Information Processing (ICCIP'20), Tokyo, Japan*. ACM, May 2020.

[33] R. R. Chowdhury, M. S. Hossain, S. Hossain, and K. Andersson. Analyzing sentiment of movie reviews in bangla by applying machine learning techniques. *Proc. of the 2019 International Conference on Bangla Speech and Language Processing (ICBSLP'19), Sylhet, Bangladesh*, pages 1–6, September 2019.

[34] S. Bird and E. Loper. NLTK: The natural language toolkit. In *Proc. of the 1st ACL Interactive Poster and Demonstration Sessions (ACL'04), Barcelona, Spain*, pages 214–217. ACL, July 2004.

[35] C. Zhou, C. Sun, Z. Liu, and F. C. M. Lau. A C-LSTM neural network for text classification. *arXiv*, 1511.08630:1–10, November 2015.

[36] G. Keren and B. W. Schuller. Convolutional rnn: An enhanced model for extracting features from sequential data. In *Proc. of the 2016 International Joint Conference on Neural Networks, (IJCNN'16), Vancouver, British Columbia Interior, Canada*, pages 3412–3419. IEEE, July 2016.

[37] U. A. Siddiqua, A. N. Chy, and M. Aono. Tweet stance detection using an attention based neural ensemble model. In *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL'19), Minneapolis, Minnesota*, pages 1868–1873. ACL, June 2019.

[38] G. Bouma. Normalized (pointwise) mutual information in collocation extraction. In *Proc. of the Biennial GSCL Conference 2009 (GSCL'09), Potsdam, Germany*, pages 31–40. Computer Science, September 2009.

[39] H. Jiesong. Understanding and evaluating a text classification model using interpretable machine learning methods, August 2020. `https://cdr.lib.unc.edu/concern/masters_papers/kp78gn73n?locale=en` [Online; Accessed on May 15, 2022].

[40] Y. Yu, X. Si, C. Hu, and J. Zhang. A review of recurrent neural networks: Lstm cells and network architectures. *Neural Computation*, 31(7):1235–1270, July 2019.

[41] S. Xu. Bayesian naïve bayes classifiers to text classification. *Journal of Information Science*, 44(1):48–59, November 2018.

[42] S. Kanish, P. Henil, S. Devanshi, and S. Manan. A comparative analysis of logistic regression, random forest and knn models for the text classification. *Springer Nature Singapore*, 5:1–16, March 2020.

[43] Y. Shuai, Y. Zheng, and H. Huang. Hybrid software obsolescence evaluation model based on pca-svm-gridsearchcv. In *Proc. of the 9th IEEE International Conference on Software Engineering and Service Science (ICSESS'18), Beijing, China*, pages 449–453. IEEE, November 2018.

[44] P. Refaeilzadeh, L. Tang, and H. Liu. *Cross-Validation*, pages 532–538. Springer US, 2009.

[45] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall. Activation functions: Comparison of trends in practice and research for deep learning. arXiv:1811.03378, November 2018. `https://doi.org/10.48550/arXiv.1811.03378`.

[46] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proc. of the 3rd International Conference on Learning Representations (ICLR'15), San Diego, California, USA*, page arXiv:1412.6980. arXiv, July 2015.

[47] D. Q. Nguyen, T. Vu, A. Rahimi, M. H. Dao, L. T. Nguyen, and L. Doan. Wnut-2020 task 2: Identification of informative covid-19 english tweets. In *Proc. of the 6th Workshop on Noisy User-generated Text (W-NUT'20), Online*, pages 314–318. ACL, November 2020.

[48] S.D. Laxmi, R. Agarwal, and A. Sinha. DSC-IIT ISM at WNUT-2020 task 2: Detection of COVID-19 informative tweets using RoBERTa. In *Proc. of the 6th Workshop on Noisy User-generated Text (W-NUT'20), Online*, pages 409–413. ACL, November 2020.

[49] F. Tasneem, J. Naim, R. Tasnia, T. Hossain, and A. N. Chy. CSECU-DSG at WNUT-2020 task 2: Exploiting ensemble of transfer learning and hand-crafted features for identification of informative COVID-19 English tweets. In *Proc. of the 6th Workshop on Noisy User-generated Text (W-NUT'20), Online*, pages 394–398. ACL, November 2020.

[50] M. G. Kendall. Rank correlation methods (book, 1975), May 1975. `https://www.worldcat.org/title/rank-correlation-methods/oclc/3827024` [Online; Accessed on May 15, 2022].

[51] M. Lapata. Automatic evaluation of information ordering: Kendall's tau. *Computational Linguistics*, 32(4):471–484, December 2006.

[52] D. Diepgrond. Can prediction explanations be trusted? on the evaluation of interpretable machine learning methods, June 2020. `https://fse.studenttheses.ub.rug.nl/21985/` [Online; Accessed on May 15, 2022].

[53] K. N. Ramamurthy, B. Vinzamuri, Y. Zhang, and A. Dhurandhar. Model agnostic multilevel explanations. *Clinical Orthopaedics and Related Research*, 2003.06005:1–21, March 2020.

Author Biography

Mumtahina Ahmed is a Lecturer in the Department of Computer Science and Engineering, Port City International University, Bangladesh. She earned her B.Sc. Engg. degree from the University of Chittagong. She has been serving as a lecturer for about a year and presented tremendous dedication to this profession. She taught several courses, e.g., Artificial Intelligence, Computer Programming, and Database Management systems. Therefore, she is currently working on several scholarly articles under the supervision of Dr. Mohammad Shahadat Hossain. Her current research area includes Explainable AI, Natural Language Processing, and Image processing.

Dr. Mohammad Shahadat Hossain is a Professor of Computer Science and Engineering of Chittagong University, Bangladesh. He earned his M.Phil and Ph.D. from the University of Manchester Institute of Science and Technology (UMIST). He has published several scholarly articles in learned refereed journals. He awarded prestigious Commonwealth Academic Staff Fellowship and European Commission sponsored Erasmus Mundus Fellowship in 2009 and 2011 respectively. He successfully completed a number of research projects as a co-investigator. His current research area includes the modeling of risk and uncertainty using evolutionary computing techniques. He is continuing his research at the intersection of computing and real world issues like economics, business, engineering and environment. He is the innovator of SDA (Spatial Domain Analysis) approach used to facilitate socio-economic research. In addition, he earned reputation as a Tawhidi scientist, who uses this method to develop pragmatic computer model of reality. His jointly authored book entitled "Computing Reality", published by Aoishima Research Institute (blue ocean press) in Tokyo, Japan, contributed significantly to enrich the knowledge of computer science.

Dr. Raihan Ul Islam is a postdoctoral researcher at Luleå University of Technology, Sweden. He did both his M.Sc. and Ph. D. degree in Computer Science from Luleå University of Technology, Sweden. Previously he worked as a software engineer at NEC Laboratories Europe, in the Context-aware Services (CAS) and Smart Environments Technologies Group. His research interests also include Mobile Edge Computing, 5G, Machine Learning, M2M Communication, Smart Homes and Cities, Mobile Systems, and Pervasive and Ubiquitous Computing.

Dr. Karl Andersson received his master degree in Computer Science and Technology from the Royal Institute of Technology, Stockholm, Sweden, and started his professional career as a consultant, project manager, business developer, and branch manager within the Capgemini Group. Returning to academia as a PhD Student he obtained his PhD degree after defending his thesis "On Access Network Selection Models and Mobility Support in Heterogeneous Wireless Networks". After visiting Columbia University in the City of New York as a postdoctoral researcher and the National Institute of Information and Communications Technology (NICT), Tokyo, Japan as a JSPS Fellow, Karl is now Professor of Pervasive and Mobile Computing at Luleå University of Technology (LTU), Skellefteå,

Sweden. His research interests include expert systems, wireless networks, blockchain, datacenters, Internet of Things, and information security. Since February 2017 Karl is leading Centre for Distance-spanning Technology (CDT) as executive director. CDT is a joint research centre at LTU where the IT industry together with LTU maintain a project portfolio in the areas of Internet of Things, Datacenters, and Communication Networks.