Deakin SIT Research Project LaTeX Template

Submitted as Research Report / Honours / Master Dissertation in SIT723/SIT724

SUBMISSION DATE
T1-2021

First-Name Last-Name
STUDENT ID 1234567
COURSE - Master of Software Engineering Honours (S464)

Supervised by: Dr. Supervisor1, Prof, Supervisor2

Abstract

The Coronavirus Disease 2019 (COVID-19) pandemic, which began in late 2019, has resulted in a deluge of information regarding the epidemic. Information can be broadly disseminated using platforms such as mass media and social media. Regrettably, not all of the information is correct or reliable. Meanwhile, a lot of fake news was produced in various languages so that they could spread more easily to particular ethnic groups. Researchers are working to address this problem employing an advantage of Machine Learning (ML), Deep Learning (DL), and Natural Language Processing (NLP) approaches (NLP). Recently, the state-of-the-art deep learning methods using NLP demonstrated outstanding results in detecting COVID-19 misinformation. However, the effort to review and explain the results of complicated algorithms, on the other hand, is restricting their widespread implementation in strengthening public trust and lowering the risk of COVID-19 misinformation. As a result, this study used SHapley Additive exPlanations (SHAP) and Local interpretable model-agnostic explanation (LIME) to interpret predictions made by any complicated models.

Contents

List of Figures

List of Tables

1 Related Work

Most COVID-19 misinformation detection systems implement machine learning techniques to help the public distinguish whether the news spreading on social media is reliable or not [13]. As a subset of artificial intelligence, machine learning uses algorithms to train machines and make decisions like a human. It identifies the pattern of the data point based on mathematical relations and predicts the new data point similarly [32]. Traditional Machine Learning (ML) and Deep Learning(DL) methods are two techniques implemented to detect COVID-19 misinformation. Traditional ML includes logistic regression [17], support vector Machine [10], k-nearest neighbors [4], decision tree [25] etc[32]. While deep learning is part of the family of machine learning methods based on artificial neural networks. Some commonly used deep learning methods include Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), Bidirectional Encoder Representations from Transformers (BERT), etc. Regarding the increase of computational capability to tackle a large amount of data in recent years, DL models have better performances than traditional ML models. [5]. BERT is transformer-based deep learning and recent method that creates state-of-the-art models for a wide range of NLP tasks, such as question answering and language inference, without substantial task-specific architecture modifications, [12]. BERT trained on a large corpus of unlabeled data, including Wikipedia and Book Corpus at the pre-training stage and initialized model with parameters on pre-training data [12]. At the fine-tuning stage, all the parameters are fine-tuned with labelled data from downstream tasks, making BERT capable of specific tasks. BERTS's model architecture is a multi-layer bidirectional Transformer encoder that can represent any token based on its bidirectional property. It can capture both left and right contexts in all layers as it is deeply bidirectional [32]. Several extant studies on BERT and its variants for classifying COVID-19 misinformation have verified its model performances. For example, Muller et al. presented COVID-Twitter-BERT (CT-BERT), a transformer-based model, pretrained on a large corpus of Twitter messages on the topic of COVID-19. CT-BERT showed a 10-30% marginal improvement compared to BERT-Large. Similarly, Nguyen et al. [23] introduced BERTweet, pretrained on 850M Tweets and pre-training procedure was based on Roberta [18] for more robust performances.

In order to deploy machine learning techniques to build the COVID-19 misinformation detection system, existing studies rely on fact-checking websites to collect COVID-19 news with manually labelling, where a lot of experts and annotators classify news into different categories according to their rating systems. These fact-checkers can be social media official accounts, news websites, fact-checking websites, government or well-recognized authentic websites. The number of English-language fact-checking services increased more than 900% from January to March 2020 [8]. For example, Cui and Lee [11] built Covid-19 heAlthcare mIsinformation Dataset (CoAID) by searching across fact-checking websites (i.e. WHO, WebMD, Healthline) to collect authentic news and gathering fake news from CheckYourFact, PolitiFact, etc. Zhou et al.[37] referred to credential rating reports by *NewsGuard*[1] and *Media Bias/Fact Check*[2] to identify 22 reliable news outlets and 38 unreliable news sites, and collected data from selected news websites. A study conducted by Shahi et al. [30] also focused on multilingual COVID-19 news. They introduced a multilingual cross-domain fact check COVID-19 dataset across 40 languages and 105 countries. While [2] claimed that extant COVID-19 news dataset was limited to volume, because COVID-19 dataset require well-recognized fact-checkers to annotating, which is time-consuming due to Velocity of News.

Ribeiro et al. [27] conducted a case study and found that the algorithm with higher accuracy on

[1] https://www.newsguardtech.com/
[2] https://mediabiasf actcheck.com/

the validation set is much worse, a fact that is easy to see when explanations are provided. The public without prior knowledge about machine learning can hardly improve their decision-making abilities, even interacting with high accuracy results[5]. In addition, the complexity of most the-state-of-art DL models makes the prediction hard to interpret and explain to the public [36]. As a result, an essential criterion for ML models must be interpretable. Model interpretability is a significant challenge for applications of Explainable artificial intelligence [35]. Previous explanation methods can be divided into Model-Agnostic Methods (MAM) and Example-Based Explanation (MBE) [20]. Example-based explanation methods select particular instances of the dataset to explain the behaviour of machine learning models or to explain the underlying data distribution. While MAM aims to separate the explanations from the machine learning models [27]. For example, Ribeiro et al. [26] introduced Local interpretable model-agnostic explanations (LIME), which used interpretable models to learn variational data in order to approximate the predictions of the underlying black-box model. The learned model should be a good approximation of the machine learning model predictions locally. The other popular local interpretation method is SHAP by [19]. SHAP scores the feature importance with the help of a subset. SHAP is based on the game's theoretically optimal Shapley values [31]. The Shapley value involves fairly distributing both gains and costs to several players who cooperate in a coalition depending on players contribution to the total payout. Similarly, Shapley value considers the "game" as reproducing the outcome of the model, and the "players" are the observation features included in the model [20]. Besides, Lundberg and Lee [19] proposed KernelSHAP and TreeSHAP due to two significant advantages, namely, global and local interpretability.

Research [5] has shown that a trustworthy prediction model demonstrated the effectiveness of the proposed method and provided good implications in detecting misinformation about COVID-19 and improving public trust. This project deploys both SHAP and LIME in ML models to better explain model prediction and inform the public about COVID-19 misinformation detection.

2 Dataset

This section demonstrates the steps of collecting the dataset. Firstly, The researcher filters out reliable new media outlets that have been cross-checked as reliable. And then perform data processing and cleaning after web crawling news from a reliable source list. Basic exploratory analysis is also present in this section.

2.1 Identifying News Site

Along with the pandemic, public is also experiencing an "infodemic"[3] of information with low credibility regarding COVID-19 [37]. Many news media and social media have contributed to combating misleading news resulting in massive misinformation spreading on social media. It is important to filter reliable source websites to collect fact-checking news about COVID-19. This project cross-checked the credentials of various news sites and thus determined a list of news sites to collect both facts and misinformation news.

2.1.1 Misinformation from Fact Check website

In order to filter fact-checking websites to collect misleading news, the researcher referred to Wikipedia and International Fact-Checking Network(IFCN), which was launched in 2015 to bring

[3]https://www.who.int/health-topics/infodemic#tab=tab_1

together the growing community of fact-checkers around the world advocates of factual information in the global fight against misinformation. We cross-checked and identified following sources: Poynter[4], Snope[5], PolitiFact[6] and BOOM[7]:

Poynter is a non-profit journalism organisation. Poynter stepped up during the COVID-19 incident to enlighten and educate the public in order to avoid spreading false information. Poynter runs an International Fact-Checking Network (IFCN), and the institute has created a hashtag campaign called CoronaVirusFacts and DatosCoronaVirus to collect misconceptions regarding COVID-19. Poynter maintains a database of over 100 fact-checkers from all around the world, covering more than 70 nations and 40 languages.

Snope is an independent publication owned by Snope Media Group. Snope confirms the accuracy of misinformation on a variety of issues. They manually evaluate the news article's veracity and perform a contextual analysis during the fact-checking process. In reaction to the COVID-19 pandemic, Snope has compiled a list of fact-checked news items divided into categories based on the topic of the article.

PolitiFact is an non-profit project operated by the Poynter Institue[8]. Despite the fact that it belongs to Poynter, this study discovered that the data in Poynter's COVID19 database is not updated in real-time. Therefore, we crawled the data on the PolitiFact website separately.

BOOM BOOM was the first Indian member of the International Fact-Checking Network (IFCN) and the first Indian organisation to collaborate with Facebook on its Third Party Fact Checking Program. BOOM is also a non-profit digital journalism venture whose purpose is to combat disinformation, explain topics, and make the internet a safer place. BOOM is a significant fact-checking website and effort in India, dedicated to presenting readers with journalistically verified information.

2.1.2 Facts from Reliable Websites

The researcher determined reliable news medium outlets to collect news data related to COVID-19 using Media Bias/Fact Check (MBFC)[9]. **MBFC** is a website that assesses the news media's factual accuracy and political bias. Dave Van Zandt, the chief editor and website owner, is part of the fact-checking team, which also includes journalists and researchers (more details can be found on its "About" page). Based on the fact-checking results of the news pieces it has published (additional details can be found on its "Methodology" page), MBFC assigns each news organisation to one of six factual-accuracy levels: I extremely high, (ii) extremely high, (iii) most accurate, (iv) mixed, (v) low, and (vi) extremely low [33]. For automatic fact-checking investigations, such news data has been employed as ground truth. For example, figure 1 is example of how MBFC evaluated news medium of *Healthine*. It generated diagnostic report, including attributes: *History*, *Funded by/Ownership*, *Analysis/Bias*. And we also provided overview of all reliable news websites to collect fact news in table 1, including: Healthline[10] World Health Organisation(WHO)[11], BBC[12], ScienceDaily[13], National

[4] https://www.poynter.org/ifcn-covid-19-misinformation/
[5] https://www.snopes.com/fact-check/
[6] https://www.politifact.com/
[7] https://www.boomlive.in/fact-check/
[8] https://en.wikipedia.org/wiki/Poynter_Institute
[9] https://mediabiasfactcheck.com/
[10] https://www.healthline.com/
[11] https://www.who.int/
[12] https://www.bbc.com/
[13] https://www.sciencedaily.com/

Institutes of Health(NIH)[14] Centers for Disease Control and Prevention(CDC)[15], Reuters[16].

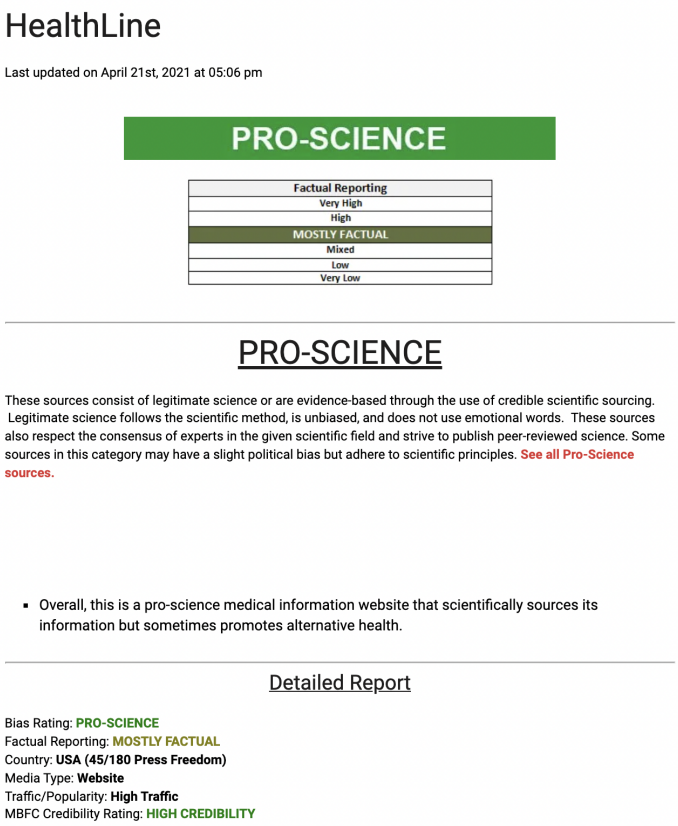


Figure 1: MBFC - Factual Rating Report of HealthLine
Source by:https://mediabiasfactcheck.com/healthline/

Table 1: Credibility Rating of News Websites

News Medium Name	MBFC's Rating	Search Query
WHO	High	*Coronavirus disease, COVID-19, delta*
ScienceDaily	High	*Covid19*
HealthLine	Mostly Factual	*Covid19*
NIH	High	*Covid19*
CDC	Very High	*SARS-CoV-2, COVID-19, Coronavirus*
BBC	High	*Covid19,*
Reuters	Very High	*Covid19*

[14] https://www.nih.gov/
[15] https://www.cdc.gov/
[16] https://www.reuters.com/

2.2 Data Collection

After we determined a list of news websites, the researcher used Beautiful-Soap[28], a python based library, to crawl news data from HTML Document Objective. In terms of COVID-19 misinformation news labels, there are different class categories that appeared, and label names vary depending on different fact-checking website rating systems discussed in section 2.2. In contrast, we consider all news collected from an identified reliable website as True. Once we obtain *News Url* linking to the news page, we used the news article scraping tool Newspaper3k[17] to crawl the news articles.

Table 2: Normalisation of News Categorisation by The Fact-Checking Websites.

Normalisation	**Original Classed**	**Definition**
True	Correct, True Correct Attribution	There are no substantial details lacking from the graded assertions, and they are all obviously true. News from selected reliable news organisations discussed in Section 2.1
False	Pants on fire, Fake, inaccurate, misleading, misinformation, conspiracy theory, exaggerated, two Pinocchios, Scam, Miscaptioned, barely-true, Labeled Satire	The fact-checker determined that the news is non-true, and that the substance of the story contains discriminatory and misleading phrase

2.3 Dataset Processing and Cleaning

Define Classes of Misinformation

Each fact-checking new site assigns a class to each piece of news and has a set of classes based on their verification system. When someone is caught in a lie, for example, "Pants on Fire," a simple rhyme popular by children all over the United States, is used. To put it another way, when someone is caught lying, [3]. PolitiFact used *"Pants on Fire"* to label news as false. And Factchek[18] utilised the term *"Manipulation"* to annotate news that contains assertions that were beyond misleading or were based on procedures that can be cheaply manipulated or structured in a manipulative manner. Furthermore, Snopes includes 98 news fact-checking websites from all around the world. For example, the word "faux" is used in a French fact-checking organisation to mean "fake," and "fasz" is a Polish word that means "false" in English. Consequently, this research manually mapped classes supplied by several fact-checking websites into binary categories (i.e. "False", "True"). A news categorisation normalisation table 2 provides overview verdict categories that original fact check websites defined misinformation. Details of the definition of fact-checking categories can be found in this study [30].

Data Cleaning

[17] https://newspaper.readthedocs.io/en/latest/
[18] www.FactCheck.kz

The news dataset was gathered from a variety of mediums and websites. The researcher thoroughly reviewed the raw data and deleted any irrelevant information. The news was filtered out if the reference link to the original page was broken or no longer existed. The researcher removed news that had been confirmed by many fact-checkers. Without losing much information on news in the processing step, this study only removed website links and special characters. Meanwhile, we kept hashtags, numbers, quote, etc., in raw news content. In addition, the researcher also used keywords in table 1 query news context to confirm that it was relevant to the COVID-19 topic.

Data Processing

Additionally, news text should always be properly digested before being fed into a model. The model is unable to handle text input, necessitating the use of numerous data processing methods. Tokenisation, stop-word elimination, emoji conversion, and text normalisation were all implemented.

2.4 Data Exploration

Data Statistics From table 3, we observe that dataset collected from multiple websites which mention in 2.1, there are totally 101 unique source websites. And both COVID19 misinformation and real news spread cross the number of 116 countries. By using python library **langdetect** [22], we detected 35 languages used in content of the articles. Furthemore, labels of COVID19 news may vary depend on original website, there are the number of 4 categories after aggregation. We found the number of 15,041 news with labels from 05 Jan 2020 to 15 Jan 2022.

Table 3: Dataset Basic information

Covid19 Misinformation Dataset Info	
Attribute	#
Source Website	101
Country	116
Languge News Used	35
Unique Label	4
Dataset Shape	15041
Dataset Collected Date Range	2020-01-05 ~2022-01-15

Label distribution From fig 3b, we observed that the proportion of false news is roughly 75% in total collected news and true news is around 20% in total collected news. Because fact-checking websites (i.e. Snope and Poynter, etc.) mainly focus on collecting false news. Although we used 101 source websites to collect news in 3, more than 95% of source websites is fact-checking websites.

Keywords and length of News Figures 4a, 4b, 4c and 4d show word clouds for true news, false news, mixed news and others news respectively. These figures indicate that there is a significant overlap of important words across among those four type of news. *COVID19*, *vaccine*, *people*, *pandemic*, *death* are most frequent words which news discussed. Specially, 2 also shows there is no significant difference of common words among false and non-false news. Furthermore, The news title is generally the first part that readers look at, and it is also the beginning of the news that is most likely to mislead readers. From figure 3a, the distribution of news title word is right-skewed. Many news's titles occur on the shorter word, and characters with a fewer number of news titles have longer lengths of words and

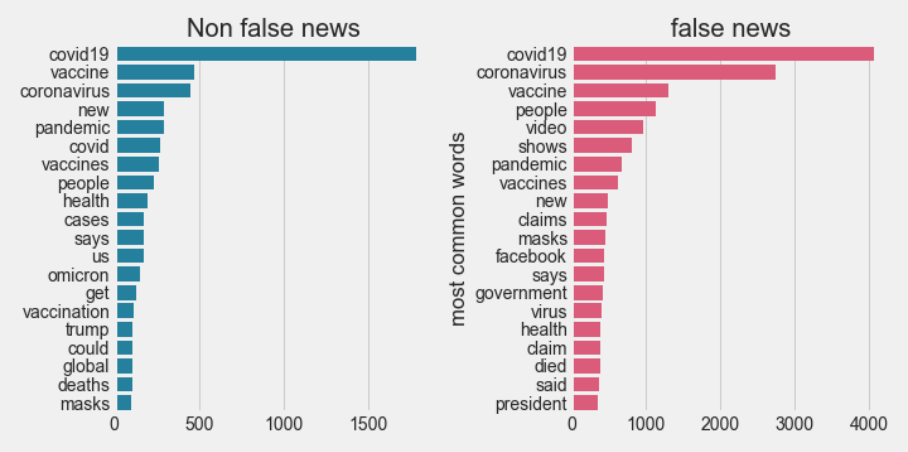
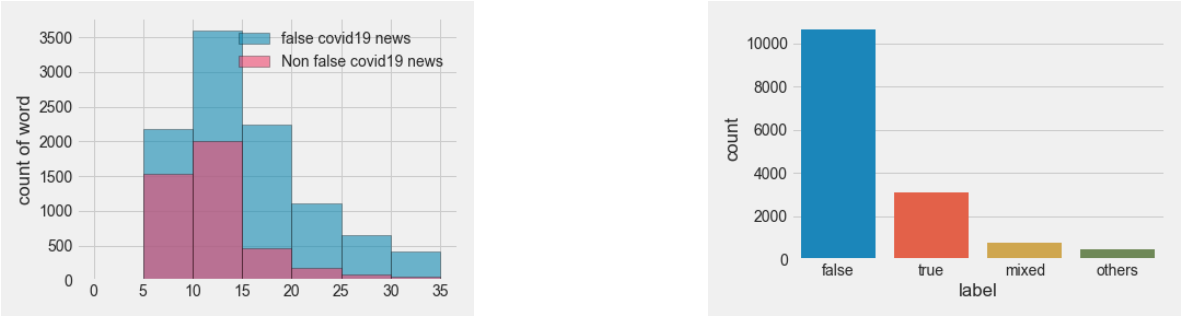


Figure 2: Most Common Words



(a) News Title Word Distribution

(b) Label Distribution

Figure 3: Word and Character Distribution for News Title

characters. In the figure 3a, no news lying on word length between 0 and 5 because we filter out that news, which mentions in section 2.3.

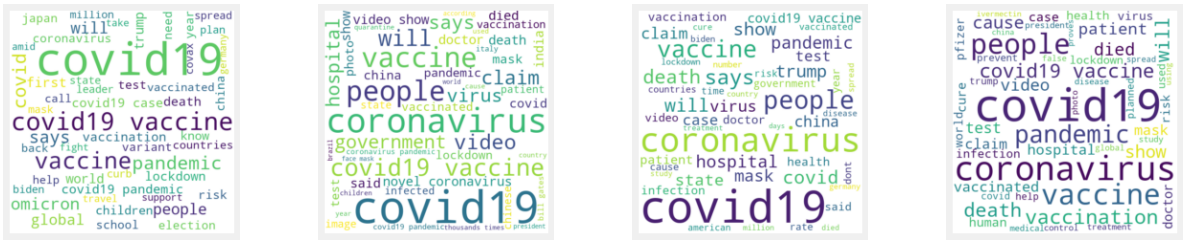
Fact Checking Website We presented data from 116 countries, from each fact-checking website operating in a specific area and language. As a result, in Figure 5a, we've shown the number of articles published by the top ten fact-checking websites. For the COVID-19, AFP2 covers the greatest amount of fact-checked news.

Across language As the fake news spread across all over the globe, it also circulated in the different national or regional language of the respective countries. We filtered the top 10 languages as per the count of the fact check articles and result is presented in figure 5b. Refer to [9], we observed that English text contribute to most proportion of COVID19 news. Followed by "es", which refer to Spanish.

3 Methodology

The method can be summarized in following steps and shown in figure 6:

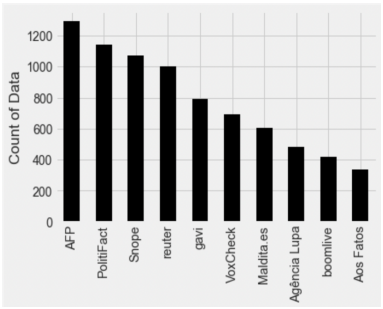
1. Firstly, we identified trustworthy, reliable news media outlets to collect news related to COVID-19 in section 2. The researcher considers news from high credential news websites (i.e., WHO, CDC,



(a) True News (b) False News (c) Mixed News (d) Others News

Figure 4: Word Cloud generated from dataset

(a) Distribution of Top 10 Fact-checking Website



(b) Distribution of Top 10 Languages Used

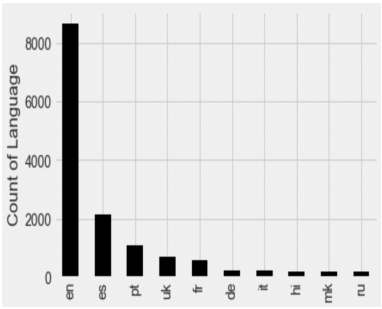


Figure 5: Top 10 language and fact_checking website distribution

etc.) as non-false claims. In contrast, false news collected from the fact-checking website have various labels, such as partially false and partially true. We keep original news labels annotated by different fact-checker. In addition, the researcher observes that more than 90% of news source websites is fact-checking websites, resulting in unbalancing the distribution of news categories. Therefore, we decided to add additional data from AAAI2021 - COVID-19 Fake News Detection Shared Task [24], CoAID, ReCOVery,

2. To better understand COVID19 misinformation themselves, we also conduct basic exploratory data analysis in 2.3. Visualization is a powerful way to improve communication between humans and data.

3. Transformer model has demonstrated outstanding performance recently. In this study, we employ BERT, RoBERTa and COVID-Twitter-BERT, etc., in our dataset. Then we select a model with the best generalization ability in the dataset and implement model-agnostic interpretation methods (i.e. SHAP and LIME) to explain what is happening within the black-box model.

4. Finally, we select a sample from the evaluation dataset to show individual explanations by using SHAP and LIME.

3.1 Model Set Up

This study conducts comparison experiments on misinformation detection tasks to highlight the major utility of models. As a baseline, the researcher uses both traditional machine learning approaches and state-of-the-art models. The following are the methods we used:

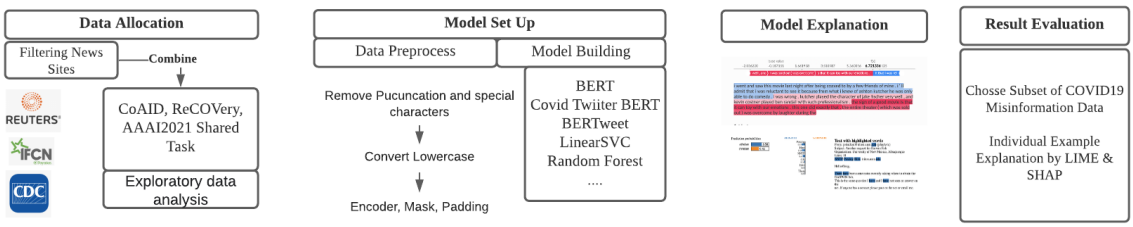


Figure 6: Summary of COVID19 Misinformation Detection Process

- Random Forest [7] is an ensemble learning method that performs classification, regression, and other tasks using a mixture of tree classifiers. Each classifier generates a random vector that is sampled independently from the input vector, and each tree votes for the most common class [7].
- LinearSVC [15]- Linear Support Vector Classifier has been widely used in classification tasks[16]. It produced a hyperplane, or a series of hyperplanes, in high-dimensional space for data point classification based on the feature set. [10].
- Logistic Regression [17] is a supervised classification model that utilizes a logistic function to describe the probability of a prediction given a set of features (i.e., a sigmoid function). It is an S-shaped curved which maps output value between 0 and 1, taking any real number as input [32].
- DistillBERT [29] is a BERT approximation approach that only needs 60% of the amount of BERT model parameters (i.e., 66 million parameters instead of 110 million). In comparison to BERT, DistillBERT retains 97 per cent of its language comprehension skills and is 60 per cent quicker.
- RoBERTa [18]: the Facebook team has created a robustly optimized BERT method. The Next Sentence Prediction task is not included in RoBERTa's pre-training phase, unlike BERT. Instead of the static masking pattern employed by BERT, RoBERTa uses larger batch sizes and dynamic masking, which allows the masked token to alter throughout training. RoBERTa-large was used in this study.
- COVID-Twitter-BERT (CT-BERT) [21] is a transformer-based model that was pre-trained on a huge corpus of COVID-19-related Twitter tweets gathered between January 12 and April 16, 2020. CT-BERT has been designed to work with COVID-19 data, particularly social media feeds like Twitter. On five separate specialized datasets, this model exhibited a marginal improvement of 10–30% over its basic model, BERT-large.
- BERTweet [23]: The first public large-scale pre-trained language model for English Tweets is BERTweet. The RoBERTa pre-training process is used to train BERTweet, which has the same architecture as BERT-base. BERTweet outperforms prior state-of-the-art models on three Tweet NLP tasks: part-of-speech tagging, named entity identification, and text categorization, according to the author's studies.

3.2 Approaches of Explainability

The ability to appropriately comprehend the output of a prediction model is critical. It builds appropriate user trust, gives insight into how a model can be improved, and aids comprehension of the simulated process [19]. In this work, the researcher analyzes the model output using both SHAP and LIME.

3.2.1 SHapley Additive exPlanations (SHAP)

Lundberg and Lee proposed SHapley Additive exPlanations (SHAP) in 2017 to describe machine learning models given predictions [19]. Furthermore, the Shapley value is most relevant in scenarios when each actor's contributions are uneven, yet each participant works together to achieve the benefit or payout [31]. The SHAP value for the $i-$th feature-value set is calculated as follows:

$$\theta_i = \beta_i \cdot (x_i - E[X_i]) \tag{1}$$

where $E[x_i]$ is the mean effect estimate for feature i, β_i is the weight corresponding to feature i, x_i is a feature value, and $E[x_i]$ is the weight corresponding to feature i. And a feature value's Shapley value is its weighted and summed contribution to the payment over all feasible feature value combinations [19]:

$$\theta_i(f) = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{i\}} \frac{|S|!\,(p - |S| - 1)!}{p!} \left(f_x\left(S \cup \{i\}\right) - f_x(S)\right) \tag{2}$$

Where f is value function of players in subset of feature S, x is the vector of feature values of the instance to be explained and p the number of features. $f_x(S)$ is the prediction for feature values in set S that are marginalized over features that are not included in set S:

$$f_x(S) = \int \hat{f}(x_1, \ldots, x_p) dP_{x \notin S} - E_X(\hat{f}(X)) \tag{3}$$

In NLP tasks, we consider each token as a feature. For example, if we want to detect COVID19 news is false or not, each word will be assigned corresponding β_i and Shapley value will push the base rate to either left or right see figure 7. For any observation and its prediction, Shapley value try to explain the difference between current prediction (i.e. $f(x)$) and base rate (i.e. $E[f(X)]$). SHAP started from background prior expectation of $E[f(X)]$ and add one features one time until we reach the current model output $f(x)$ see figure 7.

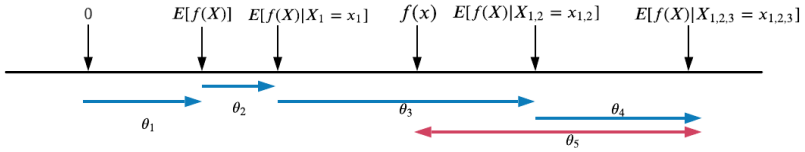


Figure 7: This diagram shows SHAP values arise from averaging the θ_i values across all possible ordering. Where $E[fX]$ is base value if we do not know any feature about to current model output.

3.2.2 Local Interpretable Model-agnostic Explanations (LIME)

Local Interpretable Model-agnostic Explanations (LIME) is the model-agnostic method that represents the model behaviour in the neighbourhood of the predicted sample [27]. The idea behind LIME is that calculate change of model predictions by perturbing features. And on this new dataset, LIME trains an local interpretable model, which assigns weight to interesting features. The general expression of LIME derived as follow :

$$\xi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g) \tag{4}$$

Where G stands for the family of plausible explanations, such as all linear regression models, $g \in G$ is a approximating model which provide explanation to the user and $\Omega(g)$ is a measure of complexity. $\xi(x)$ is the approximating model g that minimized loss L, which measure how near the explanation is to the prediction of the original model f while keeping the model complexity $\Omega(g)$ low [20] .
Although LIME is model-agnostic, the authors would like to minimise loss $L(f, g, pi_x)$ without making any assumptions about the original model f. Since the sample is weighted by pi_x, LIME is resilient to sampling noise [27].

4 Experiment and Result

To improve the quality of the model to classify misleading news of COVID-19. This study prepared two separate datasets. Since the researcher owns collected news mainly from fact-checking websites and a majority of news is labelled of false, this study increased data variety by merging data from ReCOVery, CoAID, Constraint@AAAI2021 and describe as follows:

- **ReCOVery - A Multimodal Repository for COVID-19 News Credibility Research**: Dataset analyses 2,000 news providers, identifying 60 with varying levels of trustworthiness. COVID19 gathered 2,029 news stories and 140,820 tweets from January to May 2020.
- **CoAID:COVID-19 Healthcare Misinformation Dataset**: Both facts and misinformation related to COVID19 topics, from December 2019 to September 2020. CoAID includes 4,254 news, 296,000 related to user engagements, 926 social platform posts about COVID19.
- **Constraint@AAAI2021 - COVID19 Fake News Detection in English**: Both real and false information claims, which could be in the form of news, events, social phenomenon, etc. are collected from various fact-checking websites and tools like Google fact-check-explorer[19] and IFCN chatbot [20]. There are 5,600 real news and 5100 false news related to COVID19 news.

There were three stages to this experiment process. The first is data preparation, which can take the form of word embedding or TF-IDF. In the second phase, the classifiers were trained to distinguish between fake and truthful news. Finally, SHAP and LIME were implemented to explain model prediction.

4.1 Model Experiment Setting

Experiment conducts on Google Colab Pro (CPU: Intel(R) Xeon(R) CPU @ 2.20GHz; RAM: 25.51 GB; GPU: Tesla P100-PCIE-16GB with CUDA 10.1). Details can be found below:

[19] http://toolbox.google.com/factcheck/explorer
[20] http://poy.nu/ifcnbot

Parameters This experiment implemented model by Tensorflow [1]. For the study own collected dataset, the transformer was trained on the training set for three epochs and evaluated on the validation set. The max sequence length is set as 128 and 8 for batch size on both two datasets. Furthermore, each transformer model was optimised by using Adam with a learning rate of 3e-6. Other configurations of the model are set as default as Hugging Face library [34].

Model Evaluation Criteria Since the categorisation of a document into false or non-false is a binary classification problem, the classification results can be evaluated using the confusion matrix. [6] [13]. Five metrics are derived from the confusion matrix, as shown in table 4, to assess the classifier's performance from multiple angles: accuracy, precision, recall, F1-score, and area under the curve. All of these evaluation metrics are derived from the values in the confusion matrix, which is a tabular representation of a classification model's performance on the test set and includes four parameters: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) based on the calculated predicted class versus actual class (ground truth) [14].

Evaluation Metric	Formula	Focus
Accuracy (ACC)	$\frac{TP+TN}{TP+TN+FP+FN}$	Calculate the proportion of accurately predicted occurrences to the total number of examples evaluated.
Precision	$\frac{TP}{TP+FP}$	It's used to divide the number of successfully predicted positive instances in a positive class by the total number of projected positive instances.
Recall (TPR)	$\frac{TP}{TP+FN}$	It is used to calculate the percentage of correctly classified positive events.
F1-Score (F1)	$\frac{2\times Precision\times Recall}{Precision+Recall}$	Harmonic mean of precision and recall. It is used to compare models when low precision and high recall or vice versa
Area Under The Curve (AUC)	$\frac{1-FPR+TPR}{2}$	It is used to construct an optimised learning model and for comparing learning algorithms. Its value reflects the overall ranking performance of a classifier.
Fall-Out Rate (FPR)	$\frac{FP}{TP+FP}$	It is proportion of false negatives that are incorrectly rejected.

Table 4: Evaluation metrics for our misleading-information detection system

4.2 Model Performance

From table 5, tested models includes traditional ML algorithms (i.e. LinearSVC, RandomForest, Logistic Regression), transformer model(i.e. BERT, RoBERTa and DistilBERT) and transformer (i.e. CT-BERT, BERTweet), which pretrained on the corpus of news related COVID19. Firstly, the researcher started with our own collected dataset. Three traditional ML models generally perform well; the ACC can reach 0.86 to 0.89. However, looking at F1-score and Recall, there is a significant difference between correctly predicting false news and real news. Traditional ML methods performed better on detecting false news but did not classify real news. For example, RandomForest correctly classified 91.09% false news among all correctly classified false and real news, but only 68.05% of real news has been classified among corrected classified instances. Similarly, LinearSVC and Logistic Regression have the same results. Furthermore, compared to traditional ML methods with the Transformer model, multiple variants of the BERT models can achieve outstanding results in classifying false and true news. Moreover, CT-BERT performs slight better cross all transformers with the highest ACC at 0.9677, 0.9791 F1 scores on false and 0.9291 F1 score on real news. Interestingly, RoBERTa with a large

version has the highest AUC and 0.9932 of precision on classifying false news. This study continued to test those models on additional datasets with balanced real and false news labels to demonstrate more comparison results.

Table 5 also showed that state-of-the-art NLP methods (i.e. BERT with its family) performed much better than traditional ML methods on an additional dataset of classifying COVID19 misinformation (i.e. real news and false news). Firstly, among multiple variants of the BERT models, the DistillBert - BERT model, which reduced the size of a BERT model by 40 during the pre-training phase, achieved only 0.9234 of ACC, 0.9219 of AUC, 0.9288 of F1-score on false news and 0.9172 pf F1-score on real news. While Bert-large, RoBERTa-large and CT-BERT can reach above 0.96 of ACC and F1-Score on both real and false. Significantly, it observed that RoBERTa-large perform best on additional datasets among other BERT versions. CT-BERT and BERTweet also show the out-performance result with a 3% increase on F1-score, ACC and AUC. Overall, the researcher decides to use RoBERTa-large to improve explainability when conducting prediction explanation because it has a slightly better performance on the BERT family and a significant improvement on LinearSVC, Random Forest and Logistic Regression.

Table 5: Model Performance on COVID-19 Misinformation Datasets

		Metrics				
	Model Name	F1-Score (False/True)	Recall (False/True)	Precision (False/True)	ACC	AUC
No Additional Data	LinearSVC	0.9188/0.7086	0.9125/0.7266	0.9252/0.6914	0.8730	0.8196
	LogisticRegression	0.9317/0.7295	0.9071/0.8172	0.9576/0.6588	0.8909	0.8621
	RandomForest	0.9109/0.6805	0.9049/0.6971	0.9170/0.6647	0.8606	0.8010
	CT-BERT-v2	0.9791/0.9291	0.9780/0.9325	0.9802/0.9256	0.9677	0.9553
	BERTweet	0.9740/0.9049	0.9846/0.8704	0.9636/0.9422	0.9591	0.9275
	Bert-large	0.9668/0.8918	0.9573/0.9218	0.9766/0.8636	0.9492	0.9400
	RoBERTa-large	0.9718/0.9106	0.9514/0.9773	0.9932/0.8525	0.9572	0.9643
	DistilBERT	0.9661/0.8787	0.9721/0.8595	0.9601/0.8987	0.9470	0.9158
Added Extra Data	LinearSVC	0.8416/0.8302	0.8409/0.8310	0.8423/0.8295	0.8361	0.8359
	LogisticRegression	0.8604/0.8498	0.8277/0.8152	0.8580/0.8523	0.8553	0.8552
	RandomForest	0.8327/0.8189	0.8628/0.8472	0.8282/0.8238	0.8261	0.8260
	CT-BERT-v2	0.9642/0.9613	0.9696/0.9555	0.9589/0.9671	0.9628	0.9625
	BERTweet	0.9379/0.9300	0.9621/0.9044	0.9150/0.9570	0.9342	0.9332
	Bert-large	0.9676/0.9645	0.9786/0.9528	0.9569/0.9766	0.9661	0.9657
	RoBERTa-large	0.9693/0.9676	0.9615/0.9759	0.9771/0.9595	0.9685	0.9687
	DistilBERT	0.9288/0.9172	0.9657/0.8782	0.8946/0.9599	0.9234.	0.9219

4.3 Model Prediction Explanation

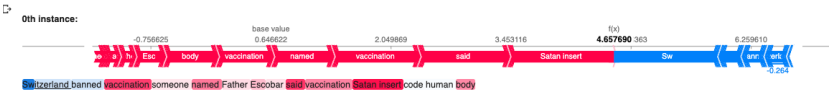
In this step, the study presenting textual or visual artefacts that establish a qualitative knowledge of the link between the instance's components (for example, words in the text) and the model's prediction [27]. The researcher selected RoBERTa-large as misleading news detector since it outperformed than other candidate model in section 4.2. And explanation news example list in table 6

4.3.1 SHAP Analysis

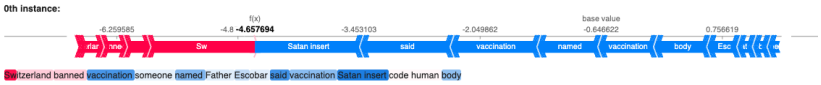
Table 6 shows two selected news examples from the explanation dataset, where only the top 7 word are presented. SHAP value of each feature can either contribute to prediction (in red) or an evidence against it (in blue). Table 6 takes absolute value and show feature importance on each prediction. And it observed that the top-ranked words in COVID-19 example 1 are: *Switzerland*, *said*, *vaccination*, *Satan*, *Escobar*, *body*, *banned*. The difference between base value and function output (DBBM) is 6.9051. SHAP value takes into account to interpret DBBM. For example, *Switzerland*, *said* has significant effect on moving prediction away from base value compare to others words. COVID-19 news example 2 is true news and the top rank words of contributing to model final output are: *together*, *learn*, *help*, *stop*, *coron*,*irus*, *spread*. Word *together* has been allocated SHAP value of 5.2064, which has a high contribution in classifying news categories. It also observed that token of *Together*, *learn*, *help* are general positive sentiment. In the figure 8a and 8b, we can see how the prediction started from base value (i.e. model outputs when the entire input text is masked) to final prediction in COVID19 false news example in table 6. Each SHAP value is an arrow that push to increase (red) or decrease (blue) the prediction. According to the definition of SHAP, it explained in an additive way how the impact of unmasking each word changes model output from the base value to the final prediction value. From 8a, we can see most of the group of red tokens push base value to the right and increase the probability of belonging to false news. While in 8b, a large number of blue group-tokens push the base value to the left and decrease the probability of belonging to real news. According to Lundberg and Lee, small groups of tokens with strong non-linear effects among them will be auto-merged together [19]. Details of feature importance on both example 1 and example 2 also show in figure 8. Figure 10a explain to us example is mostly to be false, because *someone*, *said*,*vaccination* has high blue SHAP value, which push final model output tend to be not true. In contrast, figure 10b indicated that *together*, *learn*, *help*, etc. has high red positive value, which push final model output tend to be true. Furthermore, figure 9 is the global feature SHAP bar plot, which shows global feature importance where the global importance of each feature is taken to be the mean absolute value for that feature over all the given samples [19]. However, it can hardly interpret separate words without sequential meaning.

Table 6: Example of SHAP Explanation

COVID-19 News Example 1 (Target: False)		COVID-19 News Example 2 (Target: True)	
"Switzerland banned vaccination and someone named Father Escobar said that through vaccination Satan inserts a code into a human body."		"Together help stop spread coronavirus Learn way protect others URL"	
Impact of Each Word (Top 7)		Impact of Each Word (Top 7)	
Token	SHAP	Token	SHAP
Switzerland	1.5480	together	5.2064
said	1.0480	learn	3.8318
vaccination	1.0062	help	1.4107
Satan	0.7864	stop	1.0176
Escobar	0.5015	coron	0.8441
body	0.3043	irus	0.7261
banned	0.2685	spread	0.6989
DBBM	6.9051	DBBM	10.7832



(a) SHAP Visualization of COVID-19 News Example 1 On false class



(b) SHAP Visualization of COVID-19 News Example 1 On True class

Saliency Plot
x-axis: Output Text
y-axis: Input Text

		Sw	itzerland	banned	vaccination	someone	named	Father	Esc	obar	said	vaccination	Satan insert	code human	body
LABEL_0	-0.272	-1.241	-0.264	-0.268	1.006	0.136	0.705	0.181	0.347	0.169	1.048	0.645	1.246	-0.036	0.609
LABEL_1	0.272	1.241	0.264	0.268	-1.006	-0.136	-0.705	-0.181	-0.347	-0.169	-1.048	-0.645	-1.246	0.036	-0.609

(c) Summary SHAP Plot of COVID-19 News Example 1

Figure 8: SHAP visualization

4.3.2 LIME Analysis

We also implemented LIME to interpret model prediction on both example_1 and example_2. LIME learning an interpretable model locally around our prediction and assigning a specific weight value to each feature. From figure 11 we can assume a linear relationship between each feature and prediction. For example, individual features are negative (decrease) except *Switzerland* and *code* in figure **??**. It is clear that classifying false news with help of inspecting *banned*, *someone*, *named*, other blue texts in figure **??**. In figure **??**, *Together*, *stop*, *Learn*, *way* increase model probability of prediction on real news to be high.

5 Conclusion and Future Work

To address the COVID-19 pandemic, we presented an overview of the COVID-19 misinformation detection system to reduce the risk of the "COVID-19 Pandemic".

This study collected news articles about COVID-19 from multiple sources and a few posts spreading

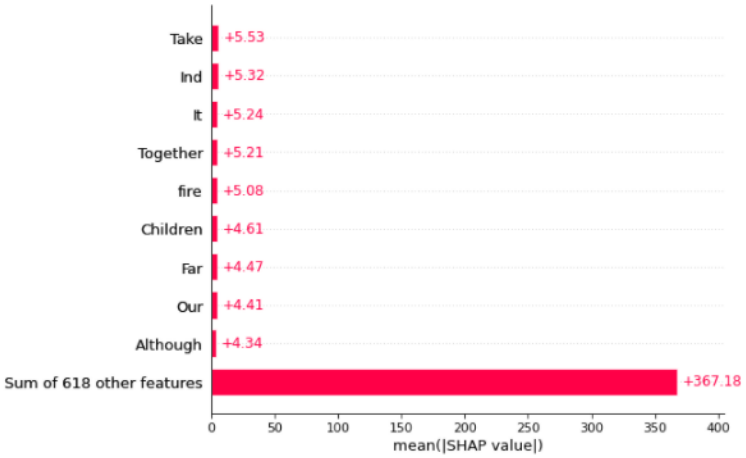
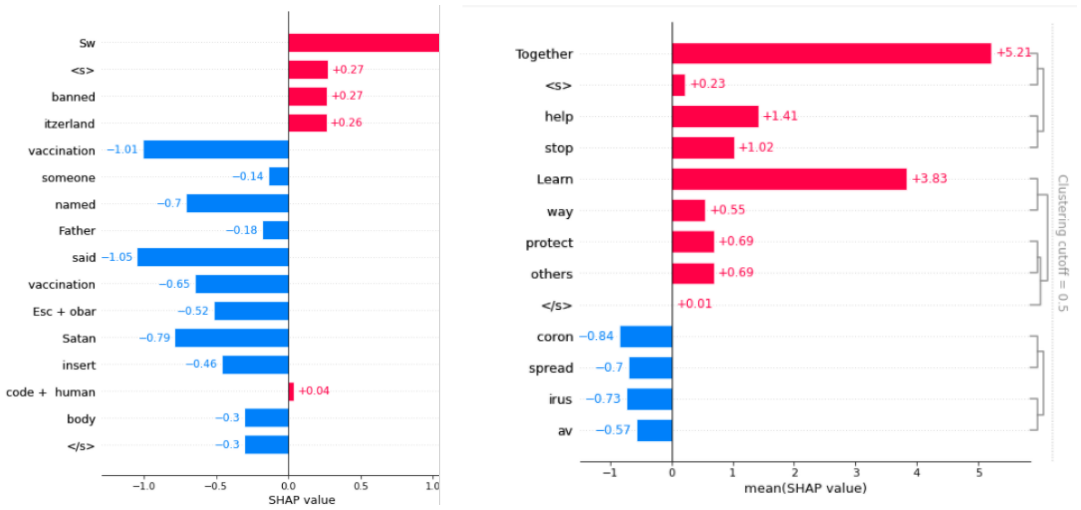
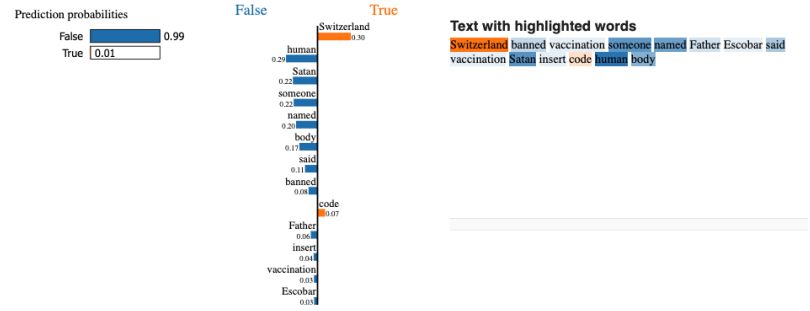


Figure 9: SHAP: Global feature importance

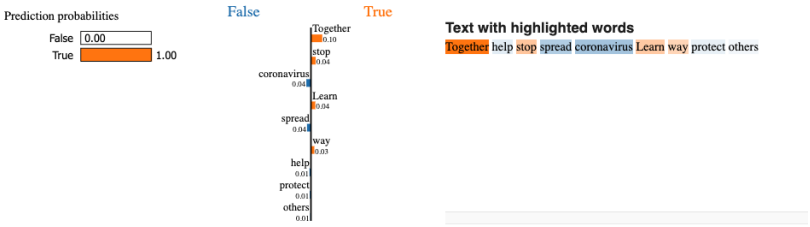


(a) SHAP Explanation of Example 1 (b) SHAP Explanation of Example 2

Figure 10: SHAP Visualization on Example of COVID-19 News in table 6



(a) LIME explanation on COVID-19 News Example 1



(b) LIME explanation on COVID-19 News Example 2

Figure 11: LIME Visualization on Example COVID-19 News

across social media. An essential explanatory is performed on the dataset section. For each COVID-19 news, we also included meta info, such as original source link, counties were to spread misinformation, etc. For this dataset crossing county in the world, We manually detected the language they used.

The researcher built up both traditional ML algorithm and state-of-the-art BERT likely model to classify misinformation about COVID-19. The study demonstrated that the BERT family showed outstanding results compared to LinearSVC, Random Forest, Logistic regression. RoBERTa outperformed other models and achieved 0.9412 of average F1-score on our collected COVID-19 dataset 0.96845 of the average score on the merged COVID-19 dataset.

Moreover, the study retrieved RoBERa's prediction output about the COVID-19 news dataset and then used SHAP and LIME to explain the given news easily. Our explanation is to improve the public's

ability to classify COVID-19 misinformation and reduce the risk of misleading social media news.

A limitation in building such a detection system is to potentially train on a large of the up-to-date dataset, not only including news but also involving data from social media. Our COVID19 misinformation was mostly collected from medium outlets and only a few from Twitter, Youtube, etc. Besides, COVID19 misinformation data has the property of Velocity, which means news can be in different formats such as video, picture and podcast, etc. Therefore, it might only have limited generalization ability to classify and explanation on various COVID-19 news.

The explanation for individual prediction improve the interpretability of model prediction. However, SHAP official document is not adapted to operate on state-of-the-art transformer-based neural networks such as BERT. In addition, visualization of explanation in the form of lists of most important features does not take into account the sequential and structurally dependent nature of the text. Moreover, it is not reasonable to expect the public to comprehend why the prediction was made, even if global important feature inspected in sample. And Model Agnostic Method (MAM) study each feature separately, ignoring the feature interaction. In NLP real-word tasks, each word of the sentence has an impact on each other, which may violate the assumption in MAM about feature independence. The primary drawbacks of utilising LIME for NLP are that it has been discovered to be unstable, which means that different sampling around the same local data might result in drastically different explanation findings, and that LIME only gives local interpretability. Furthermore, while LIME builds a linear local model around the point of interest, it presupposes that the local instant can be read linearly, which isn't necessarily the case.

In further work, it suggested that bench marking COVID-19 misinformation dataset should include data format and source where they originate from as many as possible. In addition, the MAM model also needs to consider feature interaction when we use the interpretable model to approach the original model locally.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al., *Tensorflow: Large-scale machine learning on heterogeneous distributed systems*, arXiv preprint arXiv:1603.04467, (2016).

[2] M. Abdul-Mageed, A. Elmadany, E. M. B. Nagoudi, D. Pabbi, K. Verma, and R. Lin, *Mega-COV: A billion-scale dataset of 100+ languages for COVID-19*.

[3] A. Aggarwal, *Liar! liar! pants on fire!* https://learningenglish.voanews.com/a/liars-liars-pants-on-fire/3084832.html, Dec. 2015.

[4] N. S. Altman, *An introduction to kernel and nearest-neighbor nonparametric regression*, The American Statistician, 46 (1992), pp. 175–185.

[5] J. Ayoub, X. J. Yang, and F. Zhou, *Combat covid-19 infodemic using explainable natural language processing models*, Information Processing & Management, 58 (2021), p. 102569.

[6] M. Bramer, *Principles of data mining*, vol. 180, Springer, 2007.

[7] L. Breiman, *Random forests*, Machine learning, 45 (2001), pp. 5–32.

[8] J. S. Brennen, F. M. Simon, P. N. Howard, and R. K. Nielsen, *Types, sources, and claims of COVID-19 misinformation*, PhD thesis, University of Oxford, 2020.

[9] W. contributors, *List of iso 639-1 codes*, 2021. Online; accessed 16-Jan-2022.

[10] C. Cortes and V. Vapnik, *Support-vector networks*, Machine learning, 20 (1995), pp. 273–297.

[11] L. Cui and D. Lee, *Coaid: Covid-19 healthcare misinformation dataset*, arXiv preprint arXiv:2006.00885, (2020).

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, arXiv preprint arXiv:1810.04805, (2018).

[13] M. K. Elhadad, K. F. Li, and F. Gebali, *Detecting misleading information on covid-19*, Ieee Access, 8 (2020), pp. 165201–165215.

[14] M. Fatourechi, R. K. Ward, S. G. Mason, J. Huggins, A. Schloegl, and G. E. Birch, *Comparison of evaluation metrics in classification applications with imbalanced datasets*, in 2008 seventh international conference on machine learning and applications, IEEE, 2008, pp. 777–782.

[15] C.-W. Hsu, C.-C. Chang, C.-J. Lin, et al., *A practical guide to support vector classification*, 2003.

[16] V. Kecman, *Support vector machines–an introduction*, in Support vector machines: theory and applications, Springer, 2005, pp. 1–47.

[17] S. Le Cessie and J. C. Van Houwelingen, *Ridge estimators in logistic regression*, Journal of the Royal Statistical Society: Series C (Applied Statistics), 41 (1992), pp. 191–201.

[18] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, *Roberta: A robustly optimized bert pretraining approach*, arXiv preprint arXiv:1907.11692, (2019).

[19] S. M. Lundberg and S.-I. Lee, *A unified approach to interpreting model predictions*, in Proceedings of the 31st international conference on neural information processing systems, 2017, pp. 4768–4777.

[20] C. Molnar, *Interpretable Machine Learning*.

[21] M. Müller, M. Salathé, and P. E. Kummervold, *Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter*, arXiv preprint arXiv:2005.07503, (2020).

[22] S. Nakatani, *Language detection library for java*, 2010.

[23] D. Q. Nguyen, T. Vu, and A. T. Nguyen, *Bertweet: A pre-trained language model for english tweets*, arXiv preprint arXiv:2005.10200, (2020).

[24] 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*, 2020.

[25] J. R. Quinlan, *Induction of decision trees*, Machine learning, 1 (1986), pp. 81–106.

[26] M. T. Ribeiro, S. Singh, and C. Guestrin, *Model-agnostic interpretability of machine learning*, arXiv preprint arXiv:1606.05386, (2016).

[27] M. T. Ribeiro, S. Singh, and C. Guestrin, *"why should i trust you?" explaining the predictions of any classifier*, in Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.

[28] L. Richardson, *Beautiful soup documentation*, April, (2007).

[29] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, *Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter*, arXiv preprint arXiv:1910.01108, (2019).

[30] G. K. Shahi and D. Nandini, *FakeCovid – A Multilingual Cross-domain Fact Check News Dataset for COVID-19*, arXiv:2006.11343 [cs], (2020). arXiv: 2006.11343.

[31] L. S. Shapley, *17. A value for n-person games*, Princeton University Press, 2016.

[32] A. Ullah, A. Das, A. Das, M. A. Kabir, and K. Shu, *A survey of covid-19 misinformation: Datasets, detection techniques and open issues*, arXiv preprint arXiv:2110.00737, (2021).

[33] T. Wilner, *We can probably measure media bias. but do we want to?* https://www.cjr.org/innovations/measure-media-bias-partisan.php, accessed December 29, 2021.

[34] T. Wolf, J. Chaumond, L. Debut, V. Sanh, C. Delangue, A. Moi, P. Cistac, M. Funtowicz, J. Davison, S. Shleifer, et al., *Transformers: State-of-the-art natural language processing*, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 38–45.

[35] L. Yang, Y. Zhao, X. Niu, Z. Song, Q. Gao, and J. Wu, *Municipal solid waste forecasting in china based on machine learning models. front*, Energy Res, 9 (2021), p. 763977.

[36] W. Zhao, T. Joshi, V. N. Nair, and A. Sudjianto, *Shap values for explaining cnn-based text classification models*, arXiv preprint arXiv:2008.11825, (2020).

[37] X. Zhou, A. Mulay, E. Ferrara, and R. Zafarani, *ReCOVery: A multimodal repository for COVID-19 news credibility research*, pp. 3205–3212.