A BERT-based Explainable System for COVID-19 Misinformation Identification

Lwin Moe, Arghya Kundu, and Uyen Trang Nguyen

Department of Electrical Engineering & Computer Science
York University, Toronto, Canada
lwinmoe@yorku.ca, arghyak@yorku.ca, utn@eecs.yorku.ca

Abstract

Misinformation related to COVID-19 can have serious consequences, such as undermining public health efforts to combat the pandemic. To address this problem, many COVID-19 misinformation detection models have been proposed. Most of them focused on classification accuracy, but none provides justifications or explanations for their classification output. In this paper, we present an explainable COVID-19 misinformation detection system that classifies whether a claim (a sentence) related to COVID-19 is true, false, or partially true, using an existing BERT-based classification model. The system then provides rationales behind the predictions using the LIME XAI tool, which allow machine learning practitioners to understand in depth how the model works to debug, fine tune, and optimize the model. Furthermore, the system provides explanations for the prediction using relevant sentences extracted from news articles (which serve as the ground truth in the classification process). The sources of the news articles are listed along with a ranking of the credibility of the publisher of each article (e.g., high or medium). These pieces of information explain to end users how a classification decision is reached, what data sources were used to arrive at the decision and whether the data sources are trustworthy. Such information and explanations will instill trust in the end users of the system. To the best of our knowledge, our system is the first explainable fact checking system designed to combat COVID-19 misinformation. We present examples of explainability output provided by our system to demonstrate the effectiveness of the above explainability features. The proposed explainability framework can be readily applied to misinformation identification models in other domains, e.g., politics, finance, sports, and entertainment.

Introduction

False information (fake news) is a persistent problem on social media and the Internet. We can broadly divide fake news into three categories based on intent: 1) misinformation, 2) disinformation and 3) malinformation (Verrall 2022). Misinformation and disinformation are both false information, but the former is spread unknowingly and the latter, intentionally to cause harm. Malinformation is in fact true, but spread to cause harm on a person or organization. For the sake of simplicity, we will refer to all categories of fake news as misinformation in this paper. Furthermore, we consider COVID-19 misinformation in various contexts, for example, medical, social, cultural, political and public health context.

Misinformation related to COVID-19 can have serious consequences, such as undermining public health efforts to combat the pandemic. To address this problem, many COVID-19 misinformation detection models have been proposed (Sanaullah et al. 2022). Most of them focus on classification aaccuracy, but none provides reasons or explanations for their classification decisions.

Justifications and explanations are crucial in a misinformation detection system. End users will want to know how a classification decision was reached, what data sources were used to arrive at the decision, and whether the data sources are trustworthy. Such information and explanations will instill trust in the end users of the system. Furthermore they allow machine learning practitioners to understand in depth how their model works to debug, fine tune, optimize and extend the model.

In this paper, we present a COVID-19 misinformation detection system that

- classifies whether a claim (a sentence) related to COVID-19 is true, false, or partially true, using a BERT-based classification model proposed by Ou (2021);
- provide explanations for the classification using relevant sentences extracted from news articles. The sources of the news articles are listed along with a ranking of the credibility of the publisher of each article (e.g., high or medium);
- provide rationales behind the predictions using the LIME XAI tool (Ribeiro, Singh, and Guestrin 2016).

To the best of our knowledge, our system is the first explainable fact checking system designed to combat COVID-19 misinformation.

The proposed fact checking system consists of the following three components:

- Component 1: Given a claim (sentence), a BERT-based model (Ou 2021) classifies the claim as true, partly true or false using a set of news articles whose contents are related to the claim (see Table 1 for a sample entry from the dataset). The set of related articles, collected from reputable sources, serves as the ground truth to assess the validity of the claim. In other words, given a claim

ID	3739
Claim	COVID-19 vaccine trial killed 7 children in Senegal.
Claimant	Social Media
Related Articles	https://factcheck.afp.com/senegalese-children-did-not-die-coronavirus-vaccine-which-does-not-yet-exist https://factcheck.afp.com/busting-coronavirus-myths https://web.archive.org/web/20200410170614/https://www.weblyf.com/2020/04/covid-19-vaccines-killed-seven-children-in-senegal-africa/ ...
Label	False
Date	2020-04-09

Table 1: Sample Data from the COVMIS Dataset (Ou 2021).

c and related articles, the goal is to predict its label $y \in \{\text{true}, \text{false}, \text{partially true}\}$ using a BERT-based classification function $f : \mathcal{C} \rightarrow \mathcal{Y}$, where $\mathcal{C}$ is the space of all possible claims and $\mathcal{Y}$ is the space of possible labels.

- Component 2: Given a claim, this module extracts sentences from the related articles that are the most relevant (similar) to the claim using TF-IDF, transformer-based embeddings and cosine similarities. These extracted sentences are also used by the BERT-based model to classify the claim. This module also includes a news source credibility checker that provides a credibility ranking (high or medium) to a news outlet. For each claim that is classified, the extracted sentences are displayed along with the URLs of the related articles, and the credibility ranking of the publishers of the articles. The sentences displayed explains to the end user the rationale behind the classification output. The user can follow the links to read the related articles in their entirety for more information. The credibility ranking allows the user to make informed decisions based on the trustworthiness of the sources.
- Component 3: We incorporated the LIME XAI framework into the BERT-based classifier to provide explanations for the predictions. The LIME framework provides local interpretations by generating a set of weighted features that contribute the most to a prediction. The output from LIME allows machine learning researchers to understand the reasons behind the predictions. This information will assist them in debugging, fine tuning, and optimizing the classifier.

The contributions of this paper are Components 2 and 3, and the integration of the BERT-based classifier (Ou 2021) to form a complete explainable fact checking system. The contributions of this study can be applied to other domains beyond COVID-19 misinformation detection (e.g., election misinformation), where explainability and transparency are crucial to building and maintaining trust in automated decision making systems.

The remainder of the paper is organized as follows. In section 'Related Work', we discuss existing work on misinformation identification, including those with XAI methods, and COVID-19 datasets and models for misinformation identification. Section 'Methodology' provides a summary of the BERT-based classifier (Ou 2021), and detailed descriptions of Components 2 and 3. In section 'Results and Discussion', we present examples of explainability output provided by our system. We summarize the paper and outline future work in section 'Conclusion and Future Work'.

Related Work

Misinformation detection has been an active area of research in recent years due to the growing problem of false information spreading through social media and online platforms. In this section, we review the related work on detecting and explaining misinformation.

Misinformation detection methods are generally divided into three categories: 1) content-based, 2) feedback-based, and 3) intervention-based (Sharma et al. 2019). Content-based methods use text and linguistic features as the primary input for detection models (Fuller, Biros, and Wilson 2009; Ott et al. 2011; Feng, Banerjee, and Choi 2012). Other types of input such as videos and images are also used as visual-based features (Gupta et al. 2013; Jin et al. 2017). In feedback-based methods, information such as propagation patterns, temporal patterns, response texts and response users are used (Ma, Gao, and Wong 2017). In intervention-based methods, network monitoring, crowd-sourcing and user behaviour modeling are used (Amoruso et al. 2017; Kim et al. 2018).

Transformer-based models were also used to encode content for a classification model to identify misinformation (Ou 2021). BERT (Devlin et al. 2019), which stands for Bidirectional Encoder Representations from Transformers, is one of the transformer-based language models. There are many pre-trained BERT models using unlabeled texts. For example Med-BERT is specifically trained on electronic health records (Rasmy et al. 2020). As a result, it may not be effective for detecting misinformation appearing in the news or social media in social, cultural or political context. Pre-trained BERT models can be fine-tuned with an additional output layer for many downstream NLP tasks including misinformation detection.

Currently most pre-trained BERT models limit the token length to a maximum of 512 tokens. There are models such as LongFormer (Beltagy, Peters, and Cohan 2020) that allows token lengths of more than 512, but they require a large amount of GPU memory.

Due to this limitation, text summarization is used to trim

the input to more relevant content to feed into the BERT-based classifier. Text summarization is used as a feature in the misinformation detection task (Esmaeilzadeh, Peh, and Xu 2019; Ou 2021). In addition, Talarico and Viviani (2022) discussed extractive summarization in which *important* sentences are extracted, and their *credibility* are also assessed at the same time to generate a credible summary.

Many studies have incorporated XAI into the classifiers. Fujita et al. (2022) proposed a machine learning model, which uses the output of XAI, to identify misclassified malicious activities in cyberspace. XFake (Yang et al. 2019) is an explainable fake news detector, which assists a user with news attribute analysis, and linguistic analysis such as noun, verb and adjective ratios. Their system assists the user with word frequency analysis and part-of-speech ratios.

During the COVID pandemic, various corpora and techniques for the detection of COVID-19 misinformation have been proposed. TweetsCOV19 (Dimitrov et al. 2020) is a corpus of tweets related to COVID-19, annotated with entities, hashtags, user mentions, sentiments, and URLs. CoAID (Cui and Lee 2020) is a dataset of COVID-19 healthcare misinformation, with which various machine learning methods were tested. Wani et al. (2021) use deep learning approaches such as CNN, LSTM, and BERT to classify the tweets from the COVID-19 Fake News Dataset (Patwa et al. 2020). COVIDLIES is a collection of tweets related to COVID-19. Existing NLP systems were evaluated, using COVIDLIES dataset, to retrieve the misconception and classified whether the tweet agrees or disagrees with the misconception (Hossain et al. 2020).

Most of the existing COVID-19 misinformation detection systems (Sanaullah et al. 2022) focus on the classification tasks. They do not have an XAI component to explain the rationales behind the classification decisions. Our proposed system provides such explanations, along with the input to the model (i.e., sentences extracted from related articles) and the credibility ranking of the news sources from which the related articles were collected. We will discuss our methodology next in the following section.

Methodology

We use the BERT-based model proposed by Ou (2021) to classify a COVID-related claim as true, false or partially true. We explain the model decision and output using the LIME XAI tool and our own user interface developed to help end users understand the rationale behind the classification. Figure 1 shows the architecture of our proposed system. Our system has three important components:

1. Classifying if the claim is true, false, or partially true using the BERT-based classifier proposed by Ou (2021).
2. Explaining the features (summary sentences) for the user to make an informed decision about the result from the model.
3. Explaining the model with LIME.

The first module, the BERT-based classifier is a method reported in Ou (2021). We first use TF-IDF based query-focused summarization to extract sentences most similar to a claim we are going to classify. Sentences from the articles related to a claim are extracted using TF-IDF similarity scores to the claim. We extract five sentences as a summary of the articles. In addition to TF-IDF, we also use transformer-based embedding vectors and cosine similarities to extract most relevant sentences for the summary. We then use a pre-trained BERT model to encode the summary with embedding vectors to train a classifier to classify the claim.

The second module explains the extracted summary sentences by colour-coding them based on their source credibility scores. The News Source Credibility Checker determines the credibility of the related articles, from where the sentences are extracted, and provides that information to the user as part of the explanation. This helps the user better understand the credibility of the information they are receiving and make an informed decision about trusting the model.

The third module explains the BERT classifier using LIME framework. We will discuss each component of our proposed system in details in the following sections.

We use the COVMIS dataset (Ou, Nguyen, and Ismail 2022), which is freely available on GitHub[1], for our experiments. It is a dataset of COVID-19 related claims, which are labelled as true, partly true, and false. It consists of a set of news articles related to the claims. The articles serve as the ground truth to assess the validity of the claim. The claims and related articles are collected from different fact-checking websites such as PolitiFact, Snopes, Africa Check, Poynter and Google Fact Check Tool. The dataset has a total of 14,384 claims, 10,158 of which are false, 2192 partly true and 2034 true.

Classification Model Using BERT Embeddings

In our experiments, we use a pre-trained BERT model, *bert-base-uncased*, from HuggingFace (Devlin et al. 2019). It was trained on the BookCorpus, a dataset of 11,038 unpublished books and English Wikipedia (Hugging Face 2023). For the training parameters, we used the optimal learning rate of $4e^{-5}$ reported in Ou (2021). We trained our models for five epochs.

The BERT model allows a maximum token length of 512. As a result, we need to extract the data most relevant to a claim from the set of articles related to the claim. The following section discusses the methods for extracting the most relevant data from a set of related articles.

Extractive Summarization

Sentences from the related articles of a claim are extracted based on their similarity to the claim (i.e., similarity scores). For each claim, we combined the articles related to the claim into a single document. For each sentence in the document, we calculated the similarity score of the sentence against the claim using TF-IDF. We then selected five sentences having the highest TF-IDF scores, which forms a summary of the articles. We repeated the same summarization process but used BERT embeddings and cosine similarity scores, and compared this method with the TF-IDF method.

[1] https://github.com/caryou/COVMIS

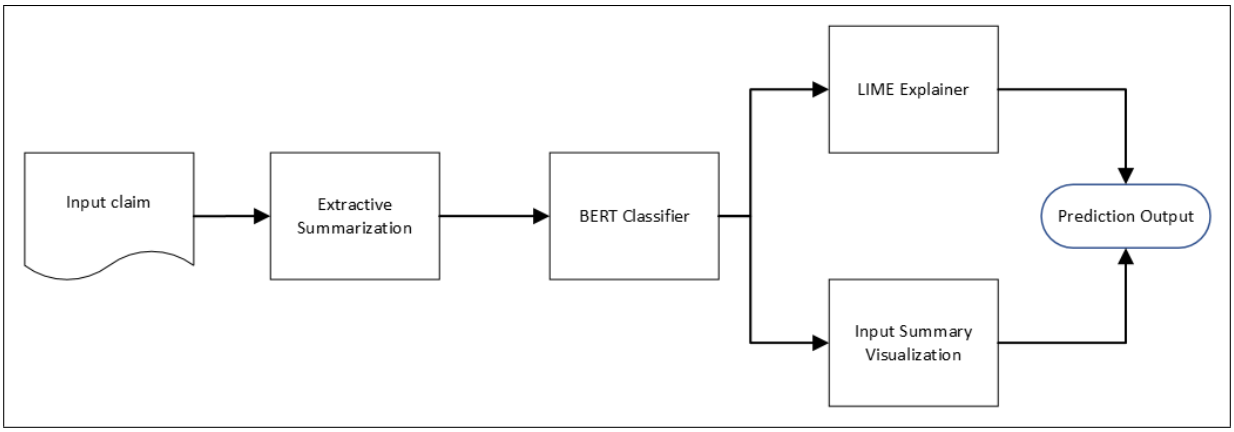


Figure 1: In our proposed system, COVID-19 related claims and summaries of articles related to the claims are classified with a BERT-based classifier. We used LIME and input sentences visualization to explain the model predictions.

Exp #	Input Selections	P	R	F1	Accu
Exp. 1	BERT+TF-IDF	0.78	0.81	0.78	0.81
	BERT+Transformer	0.78	0.80	0.78	0.80
Exp. 2	BERT+TF-IDF	0.81	0.82	0.82	0.82
	BERT+Transformer	0.82	0.83	0.82	0.83

Table 2: Results for the COVID misinformation classification using different input selection methods: 1) BERT + TF-IDF input sentence selection, 2) BERT + transformer-based input sentence selection. (P is precision, R is recall, and Accu is accuracy.)

To compare the TF-IDF and BERT-based sentence extraction methods, we ran two experiments by dividing the COVMIS dataset into two halves. The first half of the dataset contains 5753, 719, and 720 claims for training, validation and testing, respectively. We then compared the performance between the TF-IDF and transformer-based input selection methods. We repeated the experiment with the second half of the dataset. The results are in Table 2. The performance for the TF-IDF based selection was better in the first experiment, but the transformer-based selection was better in the second. We conclude that the TF-IDF selection method performs as well as and requires less computing resources than the BERT-based method.

Explanation Using Summary Sentences

We built a user interface that displays the summary sentences related to a claim to help an end user obtain information about how the prediction was make and what data was used in the classification. (The summary sentences are also the input into the BERT-based classifier.) The goal is to help the user make an informed decision about the output of the model for transparency, and help them understand the output. The explanation includes the summary sentences along with the links to the related articles, from which the sentences were extracted. The sentences are also colour-coded based on the credibility of the related articles, using the News Source Credibility Checker discussed below, so users can make their own judgments about the credibility of the source materials that the model used to classify the claim.

News Source Credibility Checker

Checking the credibility of a news source is crucial for detecting and mitigating the spread of fake news (Allcott and Gentzkow 2017). By identifying reliable sources of information, we improve the accuracy and reliability of our classification system, and help prevent the dissemination of false or misleading information (Pennycook and Rand 2021). Thus, we develop a module that finds the credibility of a news site by crawling Media Bias Fact Check (MBFC) website[2] using the domain name of the related articles and using the metrics provided by the MBFC website. The module performs a search query, on the MBFC website, for the news site under consideration, and retrieves the relevant metrics, including bias rating, factual reporting, press freedom rating, traffic/popularity, and MBFC credibility rating.

The MBFC metrics used in this module have been chosen for their relevance and credibility in evaluating the credibility of a news site. The bias rating and factual reporting metrics assess the site's political leaning and accuracy of reporting. The press freedom rating evaluates the degree of freedom of the press in the country where the news site operates. The traffic/popularity metric provides an indication of the site's influence and reach, while the MBFC credibility rating is a composite score that summarizes the site's overall credibility.

To account for inconsistencies and noise in the MBFC website, the system uses several additional methods. First, we use regular expressions to filter and process the scores from the MBFC result page and make the values consistent. Second, we compare the URL from the related articles with the source URL listed on the MBFC webpage to ensure that we are retrieving the correct webpage. Third, if the scores are still not determined, it is likely that the search query failed due to issues with the MBFC search module. In such cases, we retrieve the webpage title from the news

[2]https://www.mediabiasfactcheck.com

site, and use it to search the MBFC website to obtain the correct MBFC result. Finally, we also classify credible websites that may not have information on MBFC, such as government websites, by checking the Top-Level Domain (TLD) section of the website address against a manually curated list of TLDs that are assured to be backed by credible educational, government, or international organizations.

It should be noted that all the articles related to a claim are supposed to be collected from trustworthy websites, i.e., those with a ranking of 'high' or 'medium'. If an error is made in the data collection process that includes a website with a 'low' ranking in the set of related articles, the displayed ranking will help mitigate the impact of the error. That is, the end user would discount or discard the low ranked article based on the provided 'low' ranking. The article will then be removed from the dataset.

If a news source is not currently on our list of credibility ranking, it will be displayed as 'Unknown' (credibility). We will then manually evaluate the source (using information on the Internet, if any, and our own judgment), assign a credibility rank (high, medium or low) and update the list.

Explanation Using LIME

We use LIME to explain the BERT-based classifier. LIME builds a surrogate model based on the original BERT-based model to explain the model's behaviour.

In our experiments, we use the model trained with a maximum length of 48 tokens as an input. We tried the maximum length of 512, but we were not able to load the LIME explainer into memory because of the limitation in our computing environment. To solve the problem, we ran a series of experiments to compare the performance of the model with different token lengths 48, 64, 128, 256 and 512. As shown in Table 3, the model that uses a token length of 48 exhibits comparable performance to the model that uses a token length of 512. We settled with a token length of 48 for the purpose of LIME explanation as the model performed as well as the one with 512 token lengths.

Max-lengths	Precision	Recall	F1	Accuracy
48	0.82	0.84	0.81	0.84
64	0.83	0.82	0.83	0.82
128	0.80	0.82	0.81	0.82
256	0.82	0.83	0.82	0.83
512	0.83	0.85	0.84	0.85

Table 3: Classifying COVID-related claims using different maximum token lengths for BERT. For our dataset, the token lengths do not impact the performance. Since the model performs as well with the token length of 48 as 512, LIME framework will be used with the token length of 48 for efficiency.

It should be noted that that although short and long token lengths yielded similar performance, the token length had a significant influence on the model's training time. By decreasing the token length, the training time in our computing environment dropped from 20 minutes to two minutes per epoch, with only a slight decrease in performance. The token length of 48 was the only model we were able to use with the LIME explainer because of the memory limitation in our computing environment.

Results and Discussion

In this section, we will discuss examples of the explainability output from the proposed system, which targets both machine learning practitioners and end users. For the former, we discuss the LIME-explainer output of the model. For the latter, we present summary sentences, which are colour-coded based on the credibility ranking of the news sources from which the summary sentences are collected.

In our system to classify the claims from the COVMIS dataset, we display the claim, the prediction output, the LIME explanation and the input summary sentences along with their source credibility. The LIME explanation highlights the keywords that are important in the classification output based on their surrogate models. The summary sentences are colour-coded based on their source credibility using the News Source Credibility Checker explained in the 'Methodology' section. We will discuss the explanation output in details in the following sections.

Explanation Using LIME

The LIME framework provides an explanation of what keywords are important for the prediction model. LIME builds a surrogate model based on the original BERT-based model to explain the model's behaviour. In Figure 2, we have a LIME-based explanation for the classification of the claim: "The 'biological' lab in Wuhan where the COVID-19 virus was created was 'funded' by President Barak Hussein Obama in 2015 to the tune of $3,800,000 American dollars". We can see the highlighted keywords for the prediction. Keywords such as "biological", "wuhan", "president", "barak", "hussein" plays important roles in the model prediction process. The claim, in this case, is correctly predicted as "Fasle".

However, in Figure 3, the model incorrectly predicts the claim,"A randomized, double-blind study by Henry Ford Health System proved that hydroxychloroquine is effective against COVID-19", as Partly True. Looking at the highlighted keywords from the LIME explainer shows that the keywords used by the model are not very relevant to the claim. The summary sentences also refute the claim. The users can make an informed decision whether to trust the prediction of the model using the explanation output. The visualization of the input summary sentences also helps explain the prediction process as we will discuss next.

Explanation Using The Summary Sentences

Our system displays summary sentences for the users to be able to judge, on their own, whether the output of the classifier is trustworthy and reliable. The summary sentences from the articles related to the claim are used by the model to predict the truthfulness of the claim. In Figure 4, the claim, "U.S. coronavirus response 'slowly introducing' martial law", is correctly labelled by the model as False. We can see, in the figure, that the summary sentences from credible sources support the model's prediction. In Figure 5, on

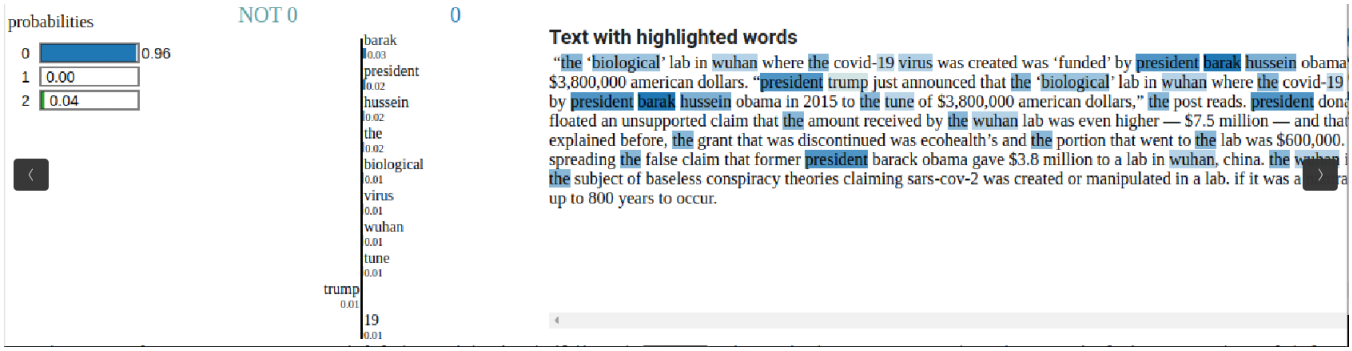
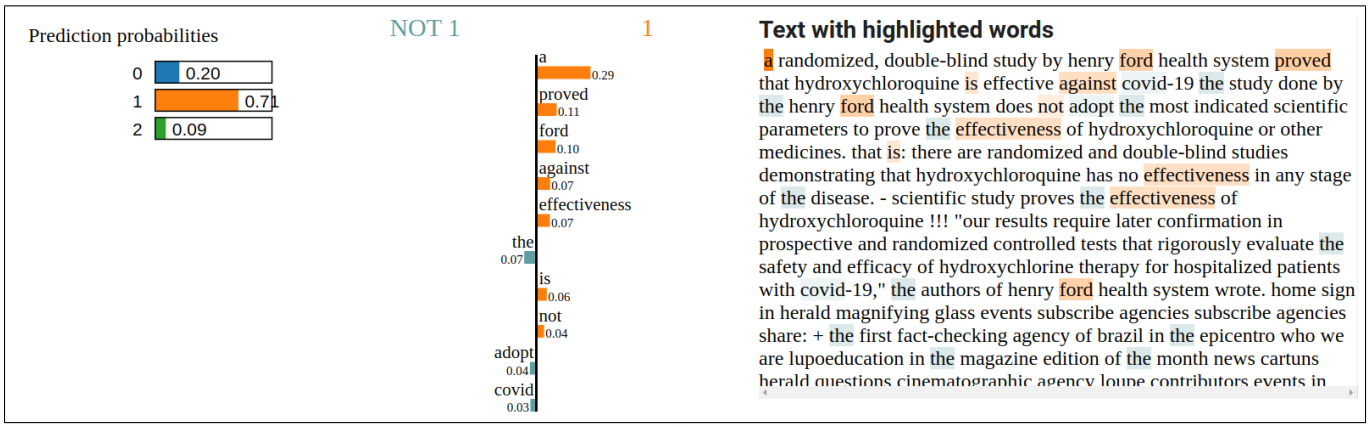


Figure 2: The claim, "The 'biological' lab in Wuhan where the COVID-19 virus was created was 'funded' by President Barak Hussein Obama in 2015 to the tune of $3,800,000 American dollars", is correctly classified as False (class label 0) by the model. LIME framework highlights the keywords with high probabilities used by the model for the prediction.



(a)

claim: A randomized, double-blind study by Henry Ford Health System proved that hydroxychloroquine is effective against COVID-19
prediction: Partly True [1]
summary sentences with source:

The study done by the Henry Ford health system does not adopt the most indicated scientific parameters to prove the effectiveness of hydroxychloroquine or other medicines. [https://piaui.folha.uol.com.br/lupa/2020/07/24/verificamos-estudo-henry-ford-hidroxicloroquina-covid-19/] [Site credibility: High]

That is: There are randomized and double-blind studies demonstrating that hydroxychloroquine has no effectiveness in any stage of the disease. [https://piaui.folha.uol.com.br/lupa/2020/07/24/verificamos-estudo-henry-ford-hidroxicloroquina-covid-19/][Site credibility: High]

- Scientific study proves the effectiveness of hydroxychloroquine !!! [https://piaui.folha.uol.com.br/lupa/2020/07/24/verificamos-estudo-henry-ford-hidroxicloroquina-covid-19/][Site credibility: High]

"Our results require later confirmation in prospective and randomized controlled tests that rigorously evaluate the safety and efficacy of hydroxychlorine therapy for hospitalized patients with Covid-19," the authors of Henry Ford health system wrote. [https://piaui.folha.uol.com.br/lupa/2020/07/24/verificamos-estudo-henry-ford-hidroxicloroquina-covid-19/][Site credibility: High]

(b)

Figure 3: (a) LIME explanation for the model with highlighted keywords, (b) Explanation of the input sentences along with source credibility. In this example, the model incorrectly predicts the claim as Partly True while the ground truth is False. However, looking at the highlighted keywords from the LIME explainer shows that the keywords that are used by the model are not very relevant to the claim. The summary sentences also refute the claim. The users can make an informed decision about the prediction of the model using the explanation output.

claim: U.S. coronavirus response “slowly introducing” martial law
prediction: False [0]
summary sentences with source:

Martial law. [https://www.nbcnewyork.com/news/local/what-is-martial-law-and-why-does-gov-cuomo-keep-getting-asked-about-it/2335832/][Site credibility: High]

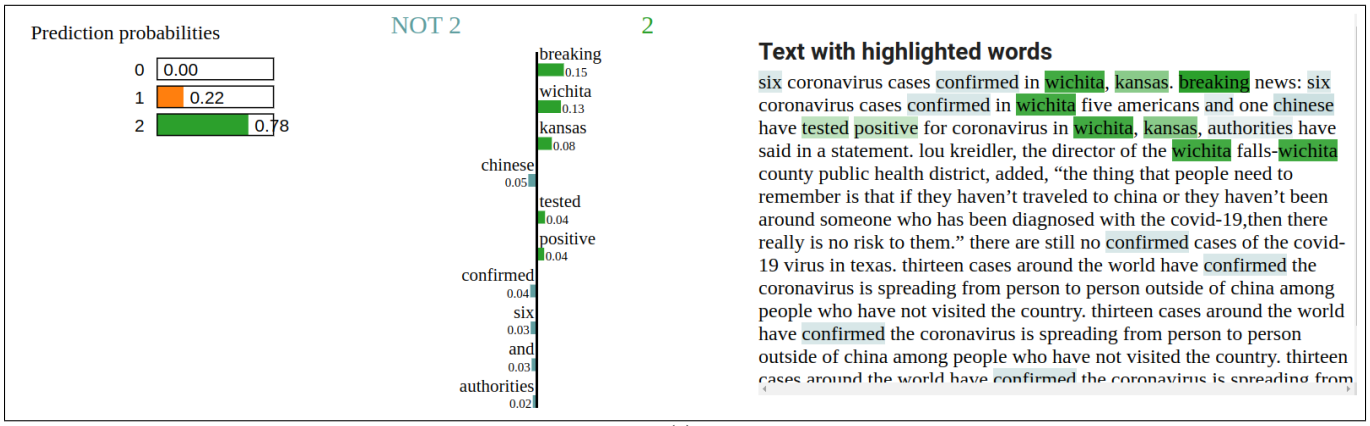
While a citizen can be governed by martial law, they cannot be governed by military law. [https://www.law.cornell.edu/wex/martial_law][Site credibility: High]

One account that shared the image wrote in the post, “Martial law…without martial law”. [https://www.reuters.com/article/uk-factcheck-coronavirus-government-cont/false-claim-government-uses-virus-to-control-all-aspects-of-citizens-lives-idUSKBN2133U9][Site credibility: High]

"I am not going to do martial law in the state of New York. [https://www.nbcnewyork.com/news/local/what-is-martial-law-and-why-does-gov-cuomo-keep-getting-asked-about-it/2335832/][Site credibility: High]

Another governor declared martial law in response to an assassination and rumors of political corruption. [https://law.jrank.org/pages/8453/Martial-Law.html][Site credibility: Unknown]

Figure 4: The claim, “U.S. coronavirus response ‘slowly introducing’ martial law”, is correctly labelled by the model as False. We can see that the summary sentences from credible sources support the model’s prediction.



(a)

claim: Six coronavirus cases confirmed in Wichita, Kansas.
prediction: True [2]
summary sentences with source:

BREAKING NEWS: Six coronavirus cases confirmed in Wichita Five Americans and one Chinese have tested positive for coronavirus in Wichita, Kansas, authorities have said in a statement. [https://archive.is/rXqQs][Site credibility: Unknown]

Lou Kreidler, the director of the Wichita Falls-Wichita County Public Health District, added, “the thing that people need to remember is that if they haven’t traveled to China or they haven’t been around someone who has been diagnosed with the COVID-19,then there really is no risk to them.” There are still no confirmed cases of the COVID-19 virus in Texas. [https://www.newschannel6now.com/2020/02/13/coronavirus-confirmed-electra-isd-not-related-china-outbreak/][Site credibility: High]

Thirteen cases around the world have confirmed the coronavirus is spreading from person to person outside of China among people who have not visited the country. [https://archive.is/QRcA9][Site credibility: Unknown]

(b)

Figure 5: (a) LIME explanation for the model with highlighted keywords, (b) Explanation of the input sentences along with source credibility. The claim, “Six coronavirus cases confirmed in Wichita, Kansas”, is incorrectly labeled by the model as True. Some of the sentences that the model uses to predict the veracity of the claim are from the sites with unknown credibility. The web page, indeed, is a fake news about the claim. The credible one in green does not mention coronavirus cases in Kansas. Instead, it is related to Texas. This highlights the data problem that the model is using for prediction. The related articles are supposed to be from the reputable websites. The explainable module catch these data issues, and inform the user whether to trust the model’s prediction.

the other hand, the claim, "Six coronavirus cases confirmed in Wichita, Kansas", is incorrectly labeled by the model as True. Some of the sentences that the model uses to predict the veracity of the claim are from a web page with unknown credibility. The web page, indeed, is a fake news about the claim. The credible one in green does not mention coronavirus cases in Kansas. Instead, it is related to Texas. This highlights the data problem that the model is using for prediction. The related articles are supposed to be from the reputable websites. The explainable module catch the data problem, and inform the user whether to trust the model's prediction or not.

Conclusion and Future Work

We have developed a system to detect misinformation and explain the classification output. In order to provide more transparency and explainability to end users, we also implemented the LIME XAI framework to explain the model's prediction process. Our approach provides two perspectives for understanding the model: one for machine learning practitioners and one for end users. The LIME framework explains the model by highlighting the keywords used in the classification process along with the weight of each word. The summary sentences along with the credibility ranking of the news sources explain to an end user how the prediction was made. The results of our study show that the proposed system is effective for identifying misinformation related to COVID-19 claims. In our future work, we will improve the accuracy of the BERT-based classifier by taking into account context and semantics of linguistic components in the claims and related articles. We will improve the user interface and displayed explanations, and have them evaluated by human participants in terms of clarity, usefulness, and user friendliness. We will also enhance the explainability of the system using knowledge graphs.

Ethical Statement

Our research aims to limit the spread of misinformation, while acknowledging the possibility of classification errors. We have implemented measures to mitigate these risks and provide users with explanations and context for informed decision-making. Additionally, we have considered the ethical implications of our data collection process, which consists of publicly available information. We recognize the broader impact and ethical considerations of our work, and strive to ensure the reliability, transparency, and trustworthiness of our model's predictions.

References

Allcott, H.; and Gentzkow, M. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2): 211–236.

Amoruso, M.; Anello, D.; Auletta, V.; and Ferraioli, D. 2017. Contrasting the Spread of Misinformation in Online Social Networks. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '17, 1323–1331. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Beltagy, I.; Peters, M. E.; and Cohan, A. 2020. Longformer: The Long-Document Transformer.

Cui, L.; and Lee, D. 2020. CoAID: COVID-19 Healthcare Misinformation Dataset. arXiv:2006.00885.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Dimitrov, D.; Baran, E.; Fafalios, P.; Yu, R.; Zhu, X.; Zloch, M.; and Dietze, S. 2020. TweetsCOV19 - A Knowledge Base of Semantically Annotated Tweets about the COVID-19 Pandemic. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, 2991–2998. New York, NY, USA: Association for Computing Machinery. ISBN 9781450368599.

Esmaeilzadeh, S.; Peh, G. X.; and Xu, A. 2019. Neural Abstractive Text Summarization and Fake News Detection. *CoRR*, abs/1904.00788.

Feng, S.; Banerjee, R.; and Choi, Y. 2012. Syntactic Stylometry for Deception Detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 171–175. Jeju Island, Korea: Association for Computational Linguistics.

Fujita, K.; Shibahara, T.; Chiba, D.; Akiyama, M.; and Uchida, M. 2022. Objection!: Identifying Misclassified Malicious Activities with XAI. In *ICC 2022 - IEEE International Conference on Communications*, 2065–2070. IEEE. ISBN 1538683474.

Fuller, C. M.; Biros, D. P.; and Wilson, R. L. 2009. Decision support for determining veracity via linguistic-based cues. *Decision Support Systems*, 46(3): 695–703.

Gupta, A.; Lamba, H.; Kumaraguru, P.; and Joshi, A. 2013. Faking Sandy: Characterizing and Identifying Fake Images on Twitter during Hurricane Sandy. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13 Companion, 729–736. New York, NY, USA: Association for Computing Machinery. ISBN 9781450320382.

Hossain, T.; Logan IV, R. L.; Ugarte, A.; Matsubara, Y.; Young, S.; and Singh, S. 2020. COVIDLies: Detecting COVID-19 Misinformation on Social Media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Online: Association for Computational Linguistics.

Hugging Face. 2023. bert-base-uncased. https://huggingface.co/bert-base-uncased. Accessed on Mar 23, 2023.

Jin, Z.; Cao, J.; Zhang, Y.; Zhou, J.; and Tian, Q. 2017. Novel Visual and Statistical Image Features for Microblogs News Verification. *Trans. Multi.*, 19(3): 598–608.

Kim, J.; Tabibian, B.; Oh, A.; Schölkopf, B.; and Gomez-Rodriguez, M. 2018. Leveraging the Crowd to Detect and Reduce the Spread of Fake News and Misinformation. In

Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM '18, 324–332. New York, NY, USA: Association for Computing Machinery. ISBN 9781450355810.

Ma, J.; Gao, W.; and Wong, K.-F. 2017. Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 708–717. Vancouver, Canada: Association for Computational Linguistics.

Ott, M.; Choi, Y.; Cardie, C.; and Hancock, J. T. 2011. Finding Deceptive Opinion Spam by Any Stretch of the Imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 309–319. Portland, Oregon, USA: Association for Computational Linguistics.

Ou, J. Y. 2021. *Misinformation Identification Using Natural Language Processing*. Master's thesis, York University.

Ou, J. Y.; Nguyen, U. T.; and Ismail, T. 2022. COVMIS: A Dataset for Research on COVID-19 Misinformation. In *2022 5th International Conference on Data Science and Information Technology (DSIT)*, 1–11.

Patwa, P.; Sharma, S.; PYKL, S.; Guptha, V.; Kumari, G.; Akhtar, M. S.; Ekbal, A.; Das, A.; and Chakraborty, T. 2020. Fighting an Infodemic: COVID-19 Fake News Dataset. *CoRR*, abs/2011.03327.

Pennycook, G.; and Rand, D. G. 2021. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, 118(15): e1921201118.

Rasmy, L.; Xiang, Y.; Xie, Z.; Tao, C.; and Zhi, D. 2020. Med-BERT: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction. *CoRR*, abs/2005.12833.

Ribeiro, M.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 97–101. San Diego, California: Association for Computational Linguistics.

Sanaullah, A. R.; Das, A.; Das, A.; Kabir, M. A.; and Shu, K. 2022. Applications of machine learning for COVID-19 misinformation: a systematic review. *Soc. Netw. Anal. Min.*, 12(1): 94.

Sharma, K.; Qian, F.; Jiang, H.; Ruchansky, N.; Zhang, M.; and Liu, Y. 2019. Combating Fake News: A Survey on Identification and Mitigation Techniques. *ACM Transactions on Intelligent Systems and Technology*, 10: 1–42.

Talarico, F. A. E.; and Viviani, M. 2022. Credible Text Summarization in Social Media. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, WI-IAT '21, 603–610. New York, NY, USA: Association for Computing Machinery. ISBN 9781450391153.

Verrall, N. 2022. *COVID-19 Disinformation, Misinformation and Malinformation During the Pandemic Infodemic: A View from the United Kingdom*, 81–112. Cham: Springer International Publishing. ISBN 978-3-030-94825-2.

Wani, A.; Joshi, I.; Khandve, S.; Wagh, V.; and Joshi, R. 2021. Evaluating Deep Learning Approaches for Covid19 Fake News Detection. In Chakraborty, T.; Shu, K.; Bernard, H. R.; Liu, H.; and Akhtar, M. S., eds., *Combating Online Hostile Posts in Regional Languages during Emergency Situation*, 153–163. Cham: Springer International Publishing. ISBN 978-3-030-73696-5.

Yang, F.; Pentyala, S. K.; Mohseni, S.; Du, M.; Yuan, H.; Linder, R.; Ragan, E. D.; Ji, S.; and Hu, X. B. 2019. XFake: Explainable Fake News Detector with Visualizations. In *The World Wide Web Conference*, WWW '19, 3600–3604. New York, NY, USA: Association for Computing Machinery. ISBN 9781450366748.