I.J. Modern Education and Computer Science, **2024, 1, 11-22**
Published Online on February 8, 2024 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2024.01.02

Explainable Fake News Detection Based on BERT and SHAP Applied to COVID-19

Xiuping Men*
National University, College of Computing and Information Technologies, Manila, 1008, Philippine
Anhui University of Finance and Economics, School of Management Science and Engineering, Bengbu, 233030, China
E-mail: menx@students.national-u.edu.ph
ORCID iD: https://orcid.org/0000-0001-9453-588X
*Corresponding Author

Vladimir Y. Mariano
National University, College of Computing and Information Technologies, Manila, 1008, Philippine
Email: vymariano@national-u.edu.ph

Received: 23 April, 2023; Revised: 05 June, 2023; Accepted: 13 July, 2023; Published: 08 February, 2024

Abstract: Fake news detection has become a significant research top in natural language processing. Since the outbreak of the covid-19 epidemic, a large amount of fake news about covid-19 has spread on social media, making the detection of fake news a challenging task. Applying deep learning models may improve predictions. However, their lack of explainability poses a challenge to their widespread adoption and use in practical applications. This work aims to design a deep learning framework for accurate and explainable prediction of covid-19 fake news. First, we choose BiLSTM as the base model and improve the classification performance of the BiLSTM model by incorporating BERT-based distillation. Then, a post-hoc interpretation method SHAP is used to explain the classification results of the model to improve the transparency of the model and increase people's confidence in the practical application. Finally, utilizing visual interpretation methods, such as significance plots, to analyze specific sample classification results for gaining insights into the key terms that influence the model's decisions. Ablation experiments demonstrated the reliability of the explainable method.

Index Terms: Covid-19, Fake news, Explainability, SHAP, BERT, Knowledge Distillation.

1. Introduction

In the digital age, social media platforms have removed the barriers of time and distance for people to share and disseminate information, gradually replacing traditional news media (e.g., newspapers, TV news, and radio news) as the primary source of information for more users [1]. The ease of use of social media has facilitated users to quickly create and share news and information online. However, this ease of use also provides ideal conditions for the spread and proliferation of fake news due to the lack of regulation and content censorship of social media users [2]. Fake news often uses emotional and lurid language to appeal to the public's curiosity and make otherwise absurd false news seem "interesting", thus triggering more people to share it. Fake news thus spreads farther and faster than real news, penetrates more deeply and extensively, and has caused some serious consequences in many social fields such as economy, politics, medicine, and culture [3,4].

Since 2019, the New Coronavirus has been a global pandemic, and according to the World Health Organization's official real-time epidemic statistics, the cumulative number of confirmed cases of the global epidemic has reached 670 million, and the cumulative number of deaths has reached 6,798,000. As the New Coronavirus epidemic continues to spread around the world, information on New Crown-related treatments and defensive measures has received widespread attention from thousands of users on social media, and at the same time, related fake news has been spread and disseminated indiscriminately on social media, triggering panic and further aggravating the threat of the disease [5]. New crown-related fake news spreads even faster and easier than the virus itself and is just as dangerous [6]. We are fighting not just an epidemic, but also a 'rumor epidemic'. Therefore, it is crucial to detect false news about new crowns on social media and stop their further spread in time to ensure that correct information about the disease is disseminated to the population [5].

The spread of fake news on social media has become a common phenomenon. There have been a large number of studies addressing the detection of covid-19 disinformation on social media. The utilization of deep learning models such as TextCNN, BiLSTM, and BERT has significantly enhanced the accuracy and efficiency of detecting COVID-19 fake news on social media platforms. However, limited attention has been given to interpreting the outcomes generated by these models, which presents challenges in establishing trust in their predictions. Regarding the interest in the basis of judgments about fake news on social media, there has been a growing recognition of the need for more nuanced approaches. Simple conclusions or binary classifications may not be sufficient to address the complexities of misinformation and its spread on social media platforms.

However, deep neural networks (DNNs) are often considered black-box models. The complexity of DNN architectures, which consist of multiple layers and millions of parameters, contributes to their black-box nature. The transformations and computations that occur within these models can be highly intricate, making it difficult to trace and interpret the exact features or patterns that influence their decisions. The lack of transparency regarding the decision-making process makes it difficult to understand how these models arrive at their conclusions, further hindering trust in their outputs.

There are numerous domains and applications where the issue of model prediction accuracy holds significant importance. When it comes to fake news detection, understanding the reasoning behind a model's predictions is crucial for building trust and confidence in its results. Stakeholders, such as users, regulators, and fact-checkers, need to have insights into how a model arrives at its conclusions to assess its credibility and make informed decisions. In this context, the lack of explanation or interpretability in deep neural network models can be seen as a critical drawback. The inability to explain the decision-making process of deep neural networks can hinder the adoption and acceptance of these models, especially when the consequences of misclassification or false positives/negatives are high. Users may be hesitant to trust the outputs of a black-box model without having an understanding of the factors that influenced its decisions. Furthermore, explanations can also help uncover potential biases or shortcomings in the model. By interpreting the features and patterns that contribute to the predictions, stakeholders can identify cases where the model may be making inaccurate or unfair judgments. The trustworthiness of fake news detection is inherently linked to the availability of explanations, making the absence of explanations a significant limitation.

To address these limitations, researchers are actively working on developing methods and techniques for interpreting and explaining the predictions made by DNNs. Interpreting predictions made by deep neural networks (DNNs) can be categorized into model-specific interpretation and model-agnostic interpretation [7].

Model-specific interpretation relies on the specific model architecture and internal operations. It provides information about the structure, weights, and relationships between neurons in the model. Some model-specific interpretation methods include activation visualization, weight visualization, and neuron activation heatmaps. These methods help explain how the model makes predictions by visualizing the internal features and learned patterns. Model-agnostic interpretation does not depend on a specific model architecture, but rather analyzes the relationship between the input and output to explain the predictions. Model-agnostic interpretation methods focus more on identifying which features play a crucial role in the decision-making process. Examples of model-agnostic interpretation methods include feature importance analysis, gradient-based attribution methods, and decision tree classifiers. These methods help identify influential features in the model's decision-making process, thereby providing more intuitive explanations. Model-specific interpretation is specific to a particular model and can provide more accurate and detailed explanations. However, it involves the internal structure of the model, and it may be challenging for non-experts to comprehend these methods. Model-agnostic interpretation methods, unlike model-specific ones, are not dependent on specific model structures and have broader applicability. It primarily concentrates on analyzing the relationship between inputs and outputs to identify influential features that significantly impact prediction results. This characteristic makes the explanations more straightforward and easily comprehensible.

The motivation behind this work is to bridge the gap between accurate fake news detection and interpretability of the underlying models. To make the black-box AI model more transparent and understandable to humans, this paper proposes a false news detection solution with high accuracy and interpretability based on the combination of knowledge distillation and SHAP without changing the original model structure. A BERT-based knowledge distillation guided BiLSTM model is first used to classify fake news to improve the classification performance. Afterwards, SHAP leverages Shapley values from cooperative game theory to quantify the contribution of each feature to the prediction in a fair and consistent manner. It takes into account the interactions and dependencies between features, making it suitable for models with complex interactions, such as neural networks. Finally, the results of the interpretable analysis are presented through visualization methods.

2. Background Study

Since the outbreak of the COVID-19 pandemic, there has been extensive research conducted on the prevalence and impact of false information related to the pandemic, as well as strategies to combat it [8]. Researchers have studied the sources of COVID-19 fake news and the psychological factors that contribute to its spread. They have also developed machine learning and natural language processing tools to automatically detect and classify fake news on social media. Earlier researchers discovered clues to false messages through anomalous high-frequency words, mainly through

shallow features extracted from texts manually, such as word count, symbol count, lexical statistics, linguistic features, and other statistical information [9,10,11]. Extracting statistical information related to word order and writing style in the text also becomes an important basis for identifying false news [12]. Clues to disinformation are also hidden in the theme, semantics, and emotion of the text. Ajao finds a correlation between fake news and the sentiment of the text and uses sentiment analysis to automatically detect new crowns of fake news [13].

Gautam combines topic distributions from Latent Dirichlet Allocation (LDA) with contextual representations from XLNet, which achieved an F1 score of 0.967 on the shared task regarding COVID-19 English fake news detection [14]. By adding semantic features, the accuracy of fake news classification can be significantly improved [15]. These methods take the extracted features and feed them into Bayesian [16], support vector machine (SVM) [17], Random Forest, and other types of machine learning classification models, and finally get the prediction results [18]. Traditional machine learning methods have numerous advantages, including fast training, insensitivity to noise, low data requirements, and interpretable outputs.

These methods, however, have some drawbacks. Tedious feature engineering needs to be performed manually, especially when faced with large datasets. The extracted features are highly dependent on relevant domain knowledge, and the wording and propagation patterns of disinformation vary from domain to domain, making it difficult to generalize to other domains [19]. Moreover, misinformation is a diverse dynamic phenomenon that changes rapidly [20], it will invalidate the originally discovered feature patterns. Researchers need to re-mine new feature patterns. In conclusion, the success of such methods depends on the selection of appropriate features, a powerful machine learning model, and the type of problem being solved. Deep learning can transform the text into low-dimensional feature vectors by learning from a large library of precursors, thus breaking the reliance on manual features [19]. Moreover, domain-specific Bert models can learn linguistic features of different domains, such as COVID-Twitter-BERT, which is pre-trained based on a large corpus of Twitter messages on COVID-19 topics; it shows 10-30% improvement compared to the basic model BERT-Large [21]. The Hugging Face API also provides pre-trained COVID-BERT and COVID-SciBERT models on the CORD-19 dataset [22]. Practically, BERT-like models have shown strong performance across a range of natural language processing (NLP) tasks [21,23,24].

Recently, researchers began to extract insights and knowledge from multiple sources or modalities, such as text, images, videos, or social relation data. Multi-modal analysis can provide more comprehensive and accurate insights than single-modal analysis. Kaliyar examines potential correlation words between authors and news stories. Using social activity as a feature to detect fake news [25]. Semantic features of text and vision are extracted and mapped to the same space to achieve a common representation of cross-modal features [26]. Although it can bring some performance improvements, increasingly complex network structures require more time to train the models. It raises the cost of applying the model while making it more opaque and reducing trust in the model. Model interpretability reveals the predictive logic of the model and fosters trust among the organizations and industries that use it [27].

Deep neural networks with attention mechanisms are often used in the field of fake message detection to interpret the output of the model [28]. Attention can be interpreted as the importance of words or features [29]. Different levels of interpretability can be obtained from multiple layers of attention. Using both word and sentence hierarchies, important words in sentences and important sentences that affect document classification can be identified [30]. By constructing sentence-comment co-attentive subnetworks, the semantic similarities between news content and user comments are captured, and attention weights are obtained by joint learning, which is used as interpretable information for the model [31]. However, it is not clear what relationship exists between attention weights and model output. The learned attention weights often do not correlate with the gradient-based feature importance measure, and the standard attention module does not provide a meaningful interpretation [32].

Therefore, in this paper, we choose the SHAP model to explain the model, firstly, the model is compressed based on knowledge distillation issued to simplify the model parameters and reduce the computation time, and then the results of the model are analyzed by SHAP to make the process transparent. SHAP provides the interpretable output and visual presentation of arbitrary machine learning in a unified way and has been successfully transferred to the field of deep learning [33,34].

Existing works on explainable fake news detection often face challenges related to the use of large models like BERT, which can be computationally intensive and require high-performance equipment. This poses limitations, especially for low-performance equipment vendors or platforms with limited computing resources. On the other hand, directly adopting smaller models such as TextCNN or LSTM may offer computational efficiency but can come at the cost of lower accuracy compared to larger models. This compromises the performance of subsequent interpretable methods since they heavily rely on the accuracy of the underlying model for effective explanations.

By addressing the challenges associated with computational requirements and the trade-off between accuracy and efficiency, we propose to apply model compression techniques to guide smaller student models by distilling knowledge from large models such as BERT, which improves the accuracy of small models while keeping the computational requirements low and improving the interpretable performance of SHAP.

3. Methodology

3.1 Proposed approach Method overview

In this article, I would like to introduce an interpretable social media fake news classification model based on knowledge distillation and SHAP (Shapley Additive Explanations). A pre-trained BERT-like model will be employed as the complex model and transfer its knowledge to a BiLSTM model for classifying covid-19 fake news on social media. Furthermore, to enhance the interpretability of our model, we will incorporate the SHAP method. SHAP provides us with a quantitative measure of the relevance and impact of each feature on the model's decision-making process. This enables us to identify the key characteristics the model considers when classifying fake news, allowing us to comprehend the rationale behind its predictions. The model structure is shown in Fig 1.

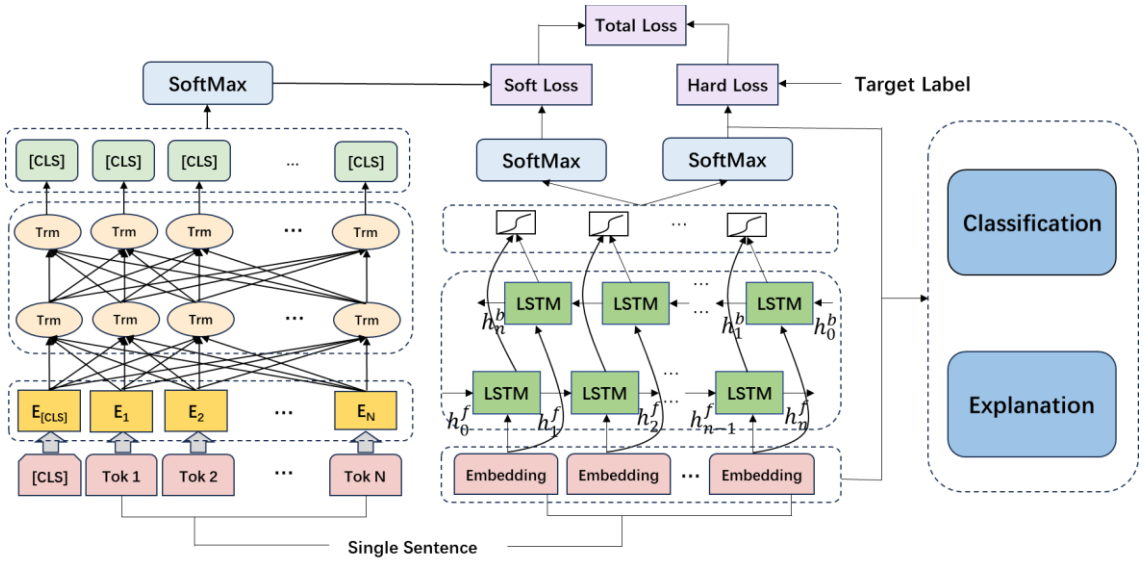

Fig. 1. The overview of proposed covid-19 fake news detection process

3.2 Dataset Description and Preprocessing

"COVID Fake News Dataset", developed by Sumit Banik and published in Coronavirus Disease Research Community-Covid-19. The dataset contains 10202 COVID fake news shared on social media platforms, including Facebook, and Instagram posts and news on social media with keywords novel coronavirus pneumonia, coronavirus, and pandemic. The dataset is divided into two columns. The first column heading is the title and the second column heading is the result. The first column contains string attributes and the second column contains binary labels 0 and 1. 0 means the title is false and 1 means the title is true.

3.3 Knowledge distillation

Deep neural networks are widely employed in numerous research disciplines and have achieved remarkable success. However, there are still some challenges. To solve complex problems and improve the training effect of the model, the network structure of the model is gradually designed to be deep and complex, which is difficult to adapt to the low resource and low power requirements of mobile computing development. Knowledge distillation was initially used for model compression as a learning paradigm to transfer knowledge from large teacher models to shallow student models to improve performance. It refers to the "knowledge distillation" of the feature representations learned by a complex, highly capable network termed the teacher model and passing them to a network with a small number of parameters and a low learning capacity. This results in a faster and more capable network which is termed the student model, so it is a conceptual model compression scheme.

We usually use complex or ensemble models to get the best results, but more complex models produce more severe parameter redundancy. For instance, there are 300 million parameters in BERT. Therefore, in forward prediction, complex computation of the model is required, which increases the running cost of the model and is not utilized for application in low-performance environments. Knowledge distillation techniques transfer the knowledge learned from a complex model (Teacher) to another lightweight model (Student). This makes the model lighter (easier to deploy) with minimal loss of performance [35] .

On the other hand, distillation allows students to learn the soft knowledge in the Teacher model, which contains inter-category information that is not available in the traditional one-hot label. Because of the nature of soft labeling in

distillation, distillation can also be considered a regularization strategy.

In this work, we opt for the BiLSTM model as our base model due to its simplicity and efficiency. To enhance the classification effectiveness of the BiLSTM model without altering its structure, we integrate knowledge distillation from BERT into the training process. Knowledge distillation allows us to transfer the knowledge learned by BERT to guide the BiLSTM model, improving its classification performance. By leveraging the knowledge distilled from BERT, the BiLSTM model can harness the insights and representations derived from a more complex model, thereby benefiting from improved classification accuracy.

By adopting this approach, we strike a balance between classification performance and computational efficiency. The BiLSTM model, guided by BERT's distilled knowledge, allows us to achieve effective classification of fake news on social media without the need for a complex model, ensuring efficient deployment and practical usability.

The process of knowledge distillation consists of 2 stages.

(1) Primitive Model Training: The "Teacher model" is trained on the original dataset using hard targets (true labels of the samples). The predicted output of the teacher network is divided by the temperature parameter (T) and then SoftMax is calculated to obtain a soft probability distribution (soft target or soft label) with values between 0 and 1, with a more moderate distribution of values. In this paper, the BERT model will be selected as the teacher model.

(2) Simple Model Training： The output of the Teacher model is used as the SOFT target to train the Student network so that the result of the Student network is close to the output of the Teacher model. Total loss is designed as a weighted average of the cross entropy corresponding to the soft and hard targets, and TextCNN and LSTM, BI-LSTM are used as student models in this paper.

The larger value of T indicates that the transfer learning process is more dependent on the contribution of the teacher network; therefore for more difficult classification or detection tasks, T is usually taken as 1 to ensure the contribution of correct predictions of the teacher network, because the prediction accuracy of the teacher network is usually better than that of the student network, and the higher the inference accuracy of the teacher network, the more beneficial the learning of the student network. So, the goal of distillation is to let Student learn the generalization ability of Teacher, and the results obtained will theoretically be better than those of Student who simply fit the training data.

In this paper, typical knowledge distillation is adopted, which uses the logits as the source of teacher knowledge. The final output of the teacher model guides the student to mimic the predictions of the teacher model. The loss function is defined as the difference between the logits of the student and the teacher model respectively. The formula is as follows.

$$L_{distill} = \left\| z^{(B)} - z^{(S)} \right\|_2^2 \tag{1}$$

where $z^{(B)}$ and $z^{(S)}$donate the teacher's and student's logits, respectively. The overall loss function is defined as a conjunction of cross-entropy and distilling objective(Tang et al., 2019):

$$L = -\alpha \sum t_i \log y_i{}^{(S)} - (1-\alpha) \left\| z^{(B)} - z^{(S)} \right\|_2^2 \tag{2}$$

Where t is the real label.

3.4 SHAP Explainability

SHAP value is a model interpretable method based on the approximation of maximum information effective and Kullback-Leibler (KL) distance approximation proposed by Suzon S, Scott Lundberg, and Kiros Halls in 2017. SHAP constructs an additive explanatory model in which all features are considered as "contributors". The core idea is to calculate the marginal contribution of features to the model output, and then interpret the "black box model" at both global and local levels.

Global interpretability is used to measure the closeness between model behavior and the true variables, and local interpretability is used to measure the magnitude of the effect of specific input variables on model behavior. SHAP values can combine these two concepts to represent model interpretability in a unified framework.

SHAP belongs to the approach of post hoc interpretation of models, which treats each feature variable in the dataset as a player from a game theory perspective, and many players cooperate to complete a project. The marginal contribution of each feature when added to the model is calculated, and then the different marginal contributions of the feature in the case of all the feature sequences are taken as the mean value, which is the SHAP baseline value of the feature, and this value indicates the degree of contribution of each feature to the prediction.

Since a deep learning model based on BERT is used, SHAP is used to decipher the classification decision process from a local perspective.

In this paper, the performance of the algorithm will be evaluated in three aspects:

1) Performance metrics

Performance metrics are used to quantify the performance of all models in this study. Common metrics include accuracy, precision, recall, and F1 score. The choice of performance metric depends on the problem being solved and the specific requirements of the system. Accuracy, precision, recall, and f1 score are chosen in this work.

2) algorithm efficiency

This study uses a model compression technique based on knowledge distillation. the goal of knowledge distillation is to take advantage of the superior performance of a larger model while maintaining the efficiency and simplicity of a smaller model. Therefore, we compare and analyze the number of parameters and the running time of the models to determine the efficiency of the algorithms.

3) Explainability

XAI (explainable artificial intelligence) refers to the ability of AI models to provide transparent and interpretable explanations for their predictions and decision-making processes. XAI can help identify potential errors or biases in AI models, which can improve safety and reduce risk. it becomes more and more critical to ensure that we understand how these systems work and why they make the decisions they do. This study will demonstrate the interpretability of the model through visualization.

3.5 Evaluation

The SHAP model provides a visualization method for intuitively presenting the importance of features and their impact on model predictions. However , due to the differences in social media users' knowledge backgrounds and cognitive abilities, their acceptance of the results may vary. For professionals with relevant domain knowledge, they may intuitively grasp the significance of these findings in explaining the rationality and effectiveness of model predictions. However , for general users or those lacking domain-specific knowledge, understanding and accepting these results may be more challenging. Therefore, the validity of the results for the interpretation cannot be verified from the users.

In order to quantify the effectiveness of the SHAP method, we conducted an elimination study. In this study, we systematically removed words from the input text in the order of their Shapley scores as determined by the SHAP model, from highest to lowest. After removing the keywords from the text, the modified text is re-fed into the BiLSTM model to observe the changes in its classification performance. This process aims to assess the impact of the removed keywords on the model's predictions.

4. Data Analysis and Findings

4.1. Experiment setup

The program was implemented on Google Colab, using Tesla T4 for training. All teacher models were implemented using TensorFlow. BiLSTM were used as student models, and SHAP was used to calculate the importance of each word on the predicted outcome. We report the prediction accuracy, recall, and F1 score of the distillation models, and analyze the model parameters and computing time of the distillation models. Finally, the interpretable analysis results are shown by the visualization method of the SHAP package.

4.2. Analysis of Fake news detection

We tested the proposed knowledge distillation-based classification model on the dataset. The teacher model used BERT, RoBERTa, DeBERTa, and XLM-RoBERTa. They were chosen because of their high classification performance in all types of NLP tasks. BiLSTM was used for the student model. BiLSTM is a powerful tool for NLP tasks that require capturing complex relationships between words in a sequence.

The prediction results of teacher models are reported in Table 1. The deBERTa model is proved to be the best-fitting model, obtaining an f1 score of 96.8% on the test set. The Roberta model had the worst prediction results among the four BERT-like models but also obtained an f1 score of 95.5%. This result again validates the excellent performance of the BERT model in NLP tasks.

To provide a comprehensive evaluation, we conducted tests on the student model BiLSTM using the dataset. Remarkably, the student models achieved an impressive F1 score exceeding 91%, as illustrated in Figure 2. The prediction results of the knowledge distillation model are shown in Fig 3. More detailed classification results are reported in Table 2.

Table 1. Performance Score of Various Teacher Models on Validation and Test Set

Teacher models	Validation Set				Test set			
	ACC	PREC	REC	F1	ACC	PREC	REC	F1
BERT	0.968	0.968	0.968	0.968	0.964	0.964	0.964	0.964
roBETTa	0.968	0.968	0.968	0.968	0.955	0.955	0.955	0.955
deBERTa	0.973	0.973	0.973	0.973	0.968	0.968	0.968	0.968
XLM-roBERTa	0.962	0.962	0.962	0.962	0.961	0.961	0.961	0.961

Table 2. The best acc of distillation models and student models

Distill Model	**ACC.**
BERT+BiLSTM	0.927
RoBERTa+BiLSTM	0.928
deBERTa+BiLSTM	**0.932**
XLMRoBERTa+BiLSTM	0.916

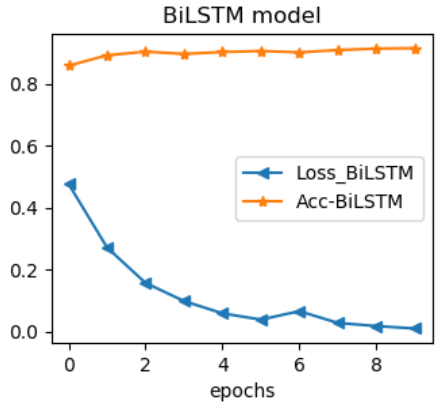

Fig. 2. Classification Accuracy and Cross-Entropy Loss Using BiLSTM

Our findings indicate that by distilling the knowledge learned by BERT and incorporating it into the training process of the BiLSTM model, we capitalize on the comprehensive understanding and representations acquired by BERT. This enables the BiLSTM model to leverage the distilled knowledge and make more accurate predictions. The observed increase in prediction accuracy highlights the effectiveness of knowledge distillation as a strategy for improving the performance of simpler models. By leveraging the strengths of a complex model like BERT, while still utilizing a more lightweight and efficient model like BiLSTM, we strike a balance between accuracy and computational efficiency.

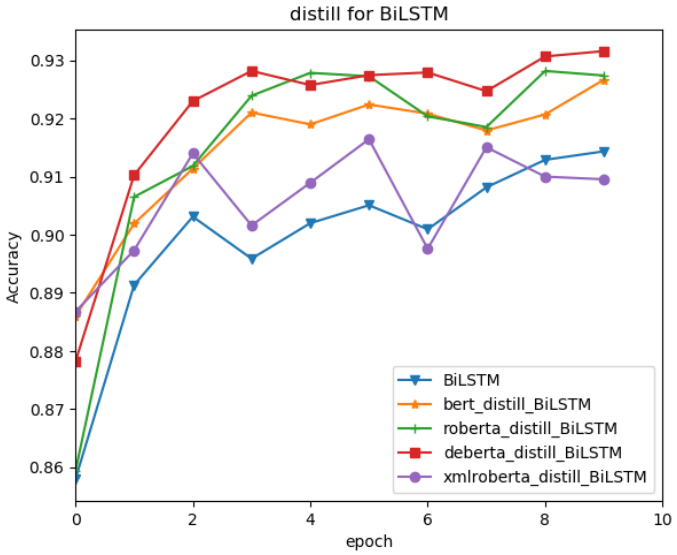

Fig. 3. Performance for Distillation Models of BiLSTM model on Test Set

4.3. Computational efficiency

Although the prediction performance of distillation magic is lower than that of the BERT model, there is a significant improvement in the computational efficiency. For all models used in the experiments, their number of parameters and inference times were compared. The results are shown in Table 3. The method presented in this paper can effectively transfer knowledge from the teacher model to the student model without increasing the parameters of the student model. The teacher model is first trained on the dataset and its parameters are saved. Then the student model is trained with the ground truth labels as hard targets and the predictions of the teacher model as soft targets. The parameters of the teacher model are kept fixed during the training process. The testing process is carried out entirely by the student model using the test set, independently of the teacher model. By adopting this approach, the student model can benefit from the rich knowledge and representations learned by the teacher model without incurring the computational requirements of the teacher model.

4.4. SHAP Explanation

Shapley values can be difficult to calculate, so we concentrate on elucidating only the top-performing models. We randomly selected two positive samples and two negative samples from the test set to see how SHAP explained the prediction results of individuals, as shown in Figure 4. The base value in the figure is the average value of the target variable on the data set. The red bars indicate that the associated words contribute to the target vector close to the base value, i.e., they have a positive impact on the predicted outcome, while the blue bars indicate that the associated words contribute to the target vector far from the base value, i.e., they have a negative impact on the predicted outcome, and the length of the color of the gradient is proportional to the contribution score of the marker.

The graph shows the prediction result of each tweet under the current classification model and the contribution of each word in the tweet to the prediction result. The "+" sign indicates that the weight increases the chance that the sentence is false, while the "-" sign indicates the opposite. Words such as "report", "daily", and "public" have a higher contribution to the prediction results compared to other words, which are frequently found in news reports in official media and therefore have a positive effect on the probability of that news being predicted as correct.

Table 3. Distillation model size and inference speed. (# Param. denotes the number of millions of parameters, and inference time is in seconds.)

Models	# Param. (Millions)	Inf. Time (Second)
BERT(Base)	109.5	28.8
RoBERTa	124.6	38.7
DeBERTa	139.2	38.7
XLMRoBERTa	278.0	29.7
BERT(Base)+ BiLSTM	8.7	0.2
RoBERTa + BiLSTM	8.7	0.6
DeBERTa+ BiLSTM	8.7	0.6
XLMRoBERTa+ BiLSTM	8.7	0.6

Fig. 4. Visualizing the contribution of words in each class for real news

Negative emotions tend to spread quickly, so fake news publishers hide their own emotions through negative emotion words, thus characterizing the message conveyed by the fake news publisher. Thus, emotions can play a role in detecting fake news(Ajao et al., 2019). This finding is corroborated in our selected sample of fake news. As shown in Figure 5. Words of negative emotions such as poor, vulnerable, and stigmas hoaxes contributed more to the classification results.

Figure 6 shows the bar chart of keywords with reverse influence and their contribution scores in each category. Figure 7 shows the bar chart of keywords with positive influence and their contribution scores in each category.

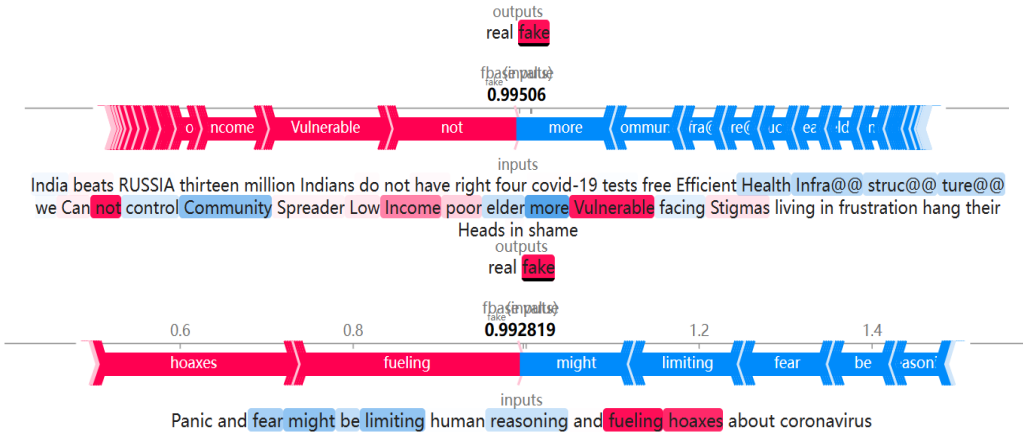

Fig. 5. Visualizing the contribution of words in each class for fake news

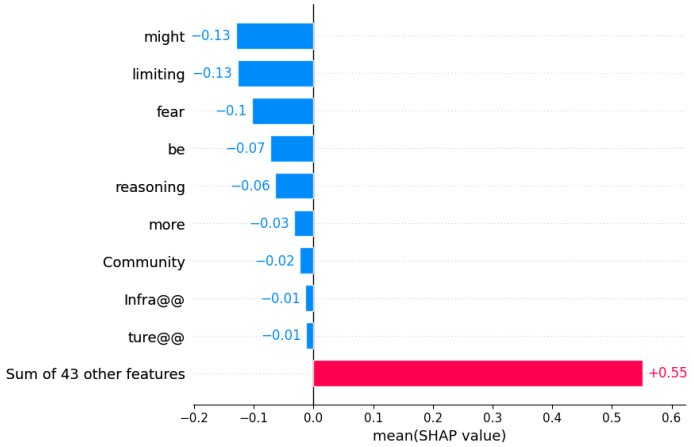

Fig. 6. Words with reverse influence and their contribution values

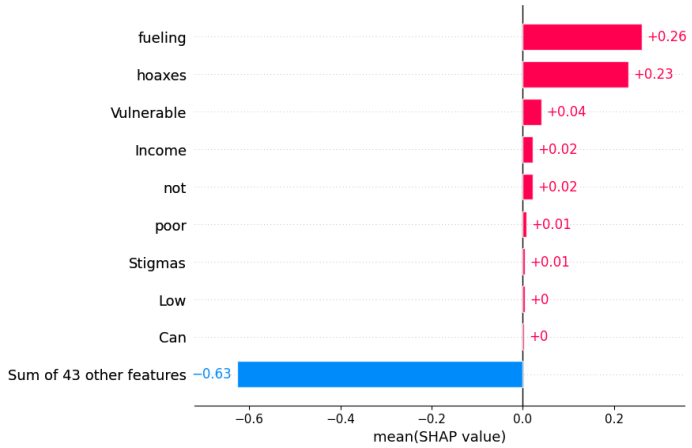

Fig. 7. Words with positive influence and their contribution values

4.5. Results of ablation experiments

The ablation study conducted in our research aimed to quantify the validity of this explainable method. By removing these important words in turn, we aimed to understand how their absence would affect the accuracy of the model. To measure this impact, we tracked the model's accuracy across our dataset as we removed the words. This analysis allowed us to quantify the contribution of these important words and ascertain their influence on the model's overall accuracy. For comparison purposes, the method of randomly deleting words was used as the baseline method.

The results (shown in Figure 8) show that the accuracy of the model decreases consistently when important words are sequentially removed from all datasets.

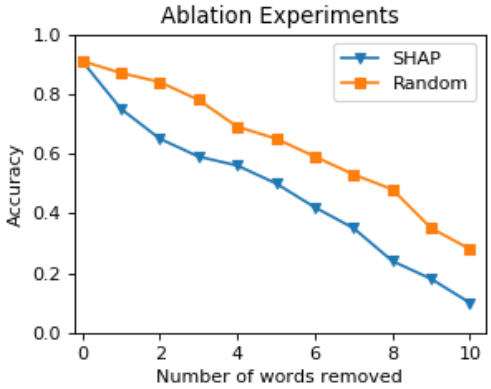

Fig. 8. Plot of number of words Removed vs. model Accuracy

Based on our findings, we observed that the SHAP method resulted in different trends in terms of changes in performance when compared to randomly dropping words. Since our method considers the importance of words in the context of the model's predictions, it is likely to capture and remove more influential words from the input text. This can have a more pronounced effect on the model's accuracy compared to randomly dropping less important words.

5. Discussion

In this paper, a generalized interpretable detection method is proposed. It leverages knowledge distillation to improve the accuracy of a smaller model without introducing additional training parameters. This allows for interpretable analysis with high classification accuracy and low computational cost, making it suitable for deployment on low-resource devices such as the cell phones. By making the method accessible to a wide range of users, including social media users, it increases the potential for widespread adoption and utilization. So, it will have a significant impact on reducing the spread of fake news on social media platforms.

However, there are a few limitations to consider. Firstly, the method relies on a teacher model, which means a high-performing teacher model needs to be identified to guide the training of the student model. The quality and performance of the teacher model can greatly impact the effectiveness of the overall approach. Secondly, the paper predominantly evaluates the reliability of interpretation results through ablation experiments. While this provides valuable insights, it does not address potential variations in how interpretation results are perceived by social media users with different knowledge backgrounds. This means the actual recognition and understanding of the interpretation results by users may differ significantly based on their individual knowledge. These limitations highlight the need for further research and experimentation to mitigate potential biases and variable interpretations among users. Additionally, exploring alternative approaches for interpretability assessment, such as user studies or surveys, could provide more comprehensive insights into the reliability and user perception of the interpretation results.

Overall, while the proposed method shows promise in terms of interpretable fake news detection with improved accuracy and low computational cost, addressing these limitations will be crucial for its effectiveness in real-world applications.

6. Conclusions and Future Work

In this paper, an efficient disinformation prediction model is developed to reveal the disinformation about covid-19 on social media by BERT-based model compression algorithm and SHAP. In conclusion, the approach presented in this paper is a promising approach to combat the spread of fake news. It contributes to the existing knowledge in the following ways:

1) By utilizing the knowledge refinement process, the student model can benefit from the expertise and rich representations learned by the teacher model. This enables efficient knowledge transfer without the need to increase the parameters of the student model during the training process. Thus, this approach provides a computationally efficient alternative to training a large and complex student model directly and independently.

2) The interpretability aspect of this approach allows users to understand why a certain news article is classified as fake, by providing clear and concrete explanations based on the features and patterns detected by the model. This not only helps individuals make informed decisions but also enables researchers and experts to further analyze and improve the detection system.

Our findings have shown how well the suggested approach works, and they have important implications for identifying COVID-19 misinformation and boosting public confidence.

With this knowledge distillation and SHAP-based model, we can achieve high accuracy while obtaining explanations for fake news classification. This will help to improve the accuracy of fake news classification on social networks by creating simplified models with lower computational requirements, which can be extended to real-time fake news detection on various platforms such as social media, where a large number of news articles are published every second. The interpretability of this approach increases the credibility and transparency of the fake news detection system. Users can gain insights into the reasons behind the classification decisions, which increases the credibility of the system and reduces the impression of black-box decisions. It also provides valuable explanations to researchers and experts. These explanations can help identify specific features, patterns or biases that contribute to fake news, thus facilitating further analysis and improving the overall understanding of the problem. The methodology can be generalized to areas other than fake news detection. Knowledge transfer and interpretability techniques have the potential to be applied in a variety of areas such as fraud detection, sentiment analysis or content management.

One of the limitations in this work is using a small COVID-19 Fake News Detection dataset can restrict the model's ability to detect and verify new fake news pertaining to COVID-19 or fake news in other medical fields. A method based solely on limited datasets and pre-trained models may not cover all domains of fake news.

To address this limitation, it is necessary to incorporate more domain-specific expertise and expand the scope of the training dataset to enhance the model's ability to detect and verify false information in the medical domain. This would help improve the accuracy in identifying and combating the spread of false information in the medical field.

References

[1] Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. Journal of economic perspectives, 31(2), 211-236.

[2] Rangel, F., Giachanou, A., Ghanem, B. H. H., & Rosso, P. (2020). Overview of the 8th author profiling task at pan 2020: Profiling fake news spreaders on twitter. In CEUR workshop proceedings (Vol. 2696, pp. 1-18). Sun SITE Central Europe.

[3] Beysolow, T. (2018). Applied natural language processing with python.

[4] Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. science, 359(6380), 1146-1151.

[5] Vijjali, R., Potluri, P., Kumar, S., & Teki, S. (2020). Two stage transformer model for COVID-19 fake news detection and fact checking. arXiv preprint arXiv:2011.13253.

[6] Bavel, J. J. V., Baicker, K., Boggio, P. S., Capraro, V., Cichocka, A., Cikara, M., ... & Willer, R. (2020). Using social and behavioural science to support COVID-19 pandemic response. Nature human behaviour, 4(5), 460-471.

[7] Rodr ́guez-P ́erez, R., & Bajorath, J. (2019). Interpretation of compound activity predictions from complex machine learning models using local approximations and shapley values. Journal of Medicinal Chemistry, 63(16), 8761-8777.

[8] Xiong, J., Lipsitz, O., Nasri, F., Lui, L. M., Gill, H., Phan, L., ... & McIntyre, R. S. (2020). Impact of COVID-19 pandemic on mental health in the general population: A systematic review. Journal of affective disorders, 277, 55-64.

[9] Giachanou, A., Rosso, P., & Crestani, F. (2019, July). Leveraging emotional signals for credibility detection. In Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval (pp. 877-880).

[10] Ksieniewicz, P., Choraś, M., Kozik, R., & Woźniak, M. (2019). Machine learning methods for fake news classification. In Intelligent Data Engineering and Automated Learning–IDEAL 2019: 20th International Conference, Manchester, UK, November 14–16, 2019, Proceedings, Part II 20 (pp. 332-339). Springer International Publishing.

[11] Azevedo, L., d'Aquin, M., Davis, B., & Zarrouk, M. (2021, August). Lux (linguistic aspects under examination): Discourse analysis for automatic fake news classification. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (pp. 41-56). Association for Computational Linguistics.

[12] Przybyla, P. (2020, April). Capturing the style of fake news. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 01, pp. 490-497).

[13] Ajao, O., Bhowmik, D., & Zargari, S. (2019, May). Sentiment aware fake news detection on online social networks. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2507-2511). IEEE.

[14] Gautam, A., Venktesh, V., & Masud, S. (2021). Fake news detection system using xlnet model with topic distributions: Constraint@ aaai2021 shared task. In Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1 (pp. 189-200). Springer International Publishing.

[15] Braşoveanu, A. M., & Andonie, R. (2019). Semantic fake news detection: a machine learning perspective. In Advances in Computational Intelligence: 15th International Work-Conference on Artificial Neural Networks, IWANN 2019, Gran Canaria, Spain, June 12-14, 2019, Proceedings, Part I 15 (pp. 656-667). Springer International Publishing.

[16] Granik, M., & Mesyura, V. (2017, May). Fake news detection using naive Bayes classifier. In 2017 IEEE first Ukraine conference on electrical and computer engineering (UKRCON) (pp. 900-903). IEEE.

[17] Aphiwongsophon, S., & Chongstitvatana, P. (2020). Identifying misinformation on Twitter with a support vector machine. Engineering and Applied Science Research, 47(3), 306-312.

[18] Cușmaliuc, C. G., Coca, L. G., & Iftene, A. (2018, November). Identifying fake news on twitter using naive bayes, SVM and random forest distributed algorithms. In Proceedings of the 13th Edition of the International Conference on Linguistic Resources and Tools for Processing Romanian Language (ConsILR-2018) pp (pp. 177-188).

[19] Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning--based text classification: a comprehensive review. ACM computing surveys (CSUR), 54(3), 1-40.

[20] Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. Information Processing & Management, 57(2), 102025.

[21] Müller, M., Salathé, M., & Kummervold, P. E. (2020). Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. arXiv preprint arXiv:2005.07503.
[22] Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., ... & Kohlmeier, S. (2020). Cord-19: The covid-19 open research dataset. ArXiv.
[23] Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., ... & Liu, Q. (2019). Tinybert: Distilling bert for natural language understanding. arXiv preprint arXiv:1909.10351.
[24] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., ... & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH), 3(1), 1-23.
[25] Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. Multimedia tools and applications, 80(8), 11765-11788.
[26] Xue, J., Wang, Y., Tian, Y., Li, Y., Shi, L., & Wei, L. (2021). Detecting fake news by exploring the consistency of multimodal data. Information Processing & Management, 58(5), 102610.
[27] Adak, A., Pradhan, B., Shukla, N., & Alamri, A. (2022). Unboxing deep learning model of food delivery service reviews using explainable artificial intelligence (XAI) technique. Foods, 11(14), 2019.
[28] Mishima, K., & Yamana, H. (2022). A survey on explainable fake news detection. IEICE TRANSACTIONS on Information and Systems, 105(7), 1249-1257.
[29] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.
[30] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016, June). Hierarchical attention networks for document classification. In Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies (pp. 1480-1489).
[31] Cui, L., Shu, K., Wang, S., Lee, D., & Liu, H. (2019, November). defend: A system for explainable fake news detection. In Proceedings of the 28th ACM international conference on information and knowledge management (pp. 2961-2964).
[32] Jain, S., & Wallace, B. C. (2019). Attention is not explanation. arXiv preprint arXiv:1902.10186.
[33] Ayoub, J., Yang, X. J., & Zhou, F. (2021). Combat COVID-19 infodemic using explainable natural language processing models. Information processing & management, 58(4), 102569.
[34] Subies, G. G., Sánchez, D. B., & Vaca, A. (2021). BERT and SHAP for Humor Analysis based on Human Annotation. In IberLEF@ SEPLN (pp. 821-828).
[35] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Authors' Profiles

Xiuping Men is currently a lecturer in the School of Management Science and Engineering at Anhui University of Finance and Economics in China. She completed his Bachelor's degree in Computer Science and Technology and Master's degree in Computer Application Technology from China University of Petroleum (East China) in 2004 and 2007, respectively. Her areas of research interest include Deep Learning and Natural Language Processing. She is currently pursuing the Ph.D. degree in College of Computing and Information Technologies at National University in Philippines.

Vladimir Y. Mariano is currently working as a professor of faculty of Computing and Information Technologies at National University in Philippines. He received the B.S. degree in statistics and the M.S. degree in computer science from the University of the Philippines Los Banos, and the Ph.D. degree in computer science and engineering from The Pennsylvania State University. His research interests include computer vision, digital image processing, and machine learning.

How to cite this paper: Xiuping Men, Vladimir Y. Mariano, "Explainable Fake News Detection Based on BERT and SHAP Applied to COVID-19", International Journal of Modern Education and Computer Science(IJMECS), Vol.16, No.1, pp. 11-22, 2024. DOI:10.5815/ijmecs.2024.01.02