 e-ISSN No. 2584-1025

Enhancing Explainability in Fake News Detection: A SHAP Based Approach for Bidirectional LSTM Models

Harsh Kumar Singh
(Student MIT School of Computing Pune)
harshjuly12@gmail.com

Abizer Jesawada
(Student MIT School of Computing Pune)
jesawadaabizer50@gmail.com

Kavita Patel
(Student MIT School of Computing Pune)
kavita.patel8109@gmail.com

Dipti Kumari
(Student MIT School of Computing Pune)
diptichandra900@gmail.com

Prof. Rohini Bhosale
(Faculty MIT School of Computing Pune)
rohini.bhosale@mituniversity.edu.in

Abstract— **In the digital age, spotting fake news has become essential to maintaining information integrity. This work provides a unique method for improving the explainability of a Bidirectional Long Short-Term Memory (BiLSTM) model for false news detection using SHapley Additive exPlanations (SHAP). In addition to achieving great accuracy, our approach offers comprehensible insights into the model's decision-making procedure. We show our method's better performance and interpretability by comparing it with several machine learning approaches. Expanding the interpretability framework and enhancing model resilience will be the main goals of future development.**

Keywords—fake news detection, Bidirectional Long Short-Term Memory, SHapely Addictive exPlanations, machine learning, model robustness, interpretable machine learning.

I. INTRODUCTION

The fast spread of misinformation and fake news has arisen as a crucial concern in the digital era, endangering social well-being and preventing informed decision-making. Addressing this issue is critical to ensuring the integrity of information. Traditional machine learning approaches, such as logistic regression and Support Vector Machines (SVM), have been used to detect fake news. While these strategies have certain advantages, they frequently lack the openness and interpretability required for establishing user confidence and comprehending the underlying decision-making processes. Recent advances in deep learning have resulted in the development of more sophisticated models, such as Bidirectional Long Short-Term Memory (BiLSTM) networks, which have shown greater performance in detecting fake news. BiLSTMs can extract contextual information from text by processing input in both forward and backward directions, making them especially useful for jobs involving sequential data, such as language processing. Despite their great performance, BiLSTM models, like many other deep learning models, are widely criticized for being "black boxes," which means that while they may make correct predictions, the logic behind these predictions is difficult to understand, limiting trust and wider adoption. To solve this issue, we suggest combining SHapley Additive ExPlanations (SHAP) with BiLSTM networks. SHAP, a game-theoretic technique, provides relevance values to various attributes, indicating how they contribute to the model's output. By using SHAP to BiLSTMs, we can break down these models' complicated decision-making process into digestible components, offering clear, interpretable insights into how certain variables (such as words or phrases in the text) impact fake news detection. This connection has several benefits, including increased transparency, trust, actionable insights, and improved model validation and debugging. Thus, combining SHAP with BiLSTM networks represents a promising approach to improving the explainability of deep learning models in the context of false news detection, preserving their high performance while providing valuable interpretative insights and allowing for more informed decisions in the ongoing battle against disinformation.

II. EASE OF USE

Our method is intended to be user-friendly and readily incorporated into current procedures. Using SHapley Additive ExPlanations (SHAP), users may intuitively comprehend which characteristics (words or tokens) impact the model's conclusions, making the false news detection process more visible. The Bidirectional Long Short-Term Memory (BiLSTM) model, when paired with SHAP, delivers simple, interpretable insights without requiring extensive technical understanding, allowing stakeholders to trust and confirm the results more effectively. This openness not only enhances model credibility, but also allows for improved decision-making and faster detection of disinformation.

III. TYPE OF DATA IN SOCIAL MEDIA POSTS

There are three primary ways in which social media networking sites read news articles, each of which may be investigated to improve explainability in false news detection utilizing SHAP-based techniques for Bidirectional LSTM models:

1. Computational Linguistics is the study of text in several languages, with an emphasis on its semantic and systematic origins. In our method, SHAP values can assist explain which linguistic traits the BiLSTM model thinks most important in determining if news is phony or authentic.

2. Multimedia: Social media posts frequently include many forms of media, such as music, videos, photographs, and graphics. This mixed media material is particularly engaging since it draws the viewer in without needing considerable reading. Using SHAP, we can see how different media types within a post impact the model's prediction.

3. Hyperlinks: Posts commonly include hyperlinks that cross-reference other sources, which increase credibility and trust by authenticating the post's origin. Our SHAP-based technique can expose how such linkages influence the model's judgments, allowing us to better understand the role of external references in the dissemination of bogus news.

Fake news on social media may be classified into numerous forms, each of which presents distinct issues for detection algorithms and necessitates alternative explainability approaches:

1. Visual-based: These fake news stories depend largely on visuals, such as distorted images or doctored movies. SHAP allows us to understand how visual material effects the model's output, which aids in the identification of visually false information.

2. User-Generated: This sort of false news is published by bogus accounts that target certain demographics such as age, gender, culture, or political affinity. SHAP can assist find patterns in user-generated material that the BiLSTM model employs to detect false news.

3. Knowledge-Based: These posts provide pseudoscientific answers for many circumstances, causing people to believe they are true. SHAP values can help explain how the model differentiates between authentic and incorrect knowledge-based assertions.

4. Style-Related: These pieces, written by pseudo-journalists, are in the manner of legitimate journalists. SHAP can emphasize stylistic elements that the model detects as indicative of bogus news, hence increasing transparency.

5. Stance-Based: This entails changing the meaning and intent of true statements. SHAP can explain how the model analyzes changes in posture to detect misleading data.

Background: Social media has dramatically increased the transmission of misleading information, which has been compounded by the introduction of smart gadgets and low-cost internet access. This problem has spread to even remote regions, resulting in the quick diffusion of both true and erroneous information. Over the last decade, the exponential growth of social media and microblogging usage has forced the creation of approaches and procedures to verify information authenticity. Several research have used machine learning to automatically detect bogus news, with some using deep learning for feature extraction. Despite these developments, the proliferation of false news on social media remains a big issue, impacting public opinion and societal behaviour. Traditional human fact-checking methods are frequently unable to manage the volume of material, emphasizing the importance of automated alternatives.

Enhancing Explainability in Fake News Detection uses SHAP-enhanced BiLSTM models to automate and produce interpretable results, solving human detection issues. Machine learning examines patterns and linguistic signals, whereas SHAP explanations provide insights into decision-making processes, promoting comprehension and trust. SHAP uses copious social media data to identify critical properties for detection, hence improving model transparency through feature engineering. SHAP addresses ethical concerns such as privacy and prejudice, ensuring that forecasts are fair. SHAP's machine learning-based detection benefits real-world applications by limiting the propagation of disinformation. In determining future paths, SHAP is critical in improving model explainability, fostering trust, and countering false news.

IV. Literature Survey on Fake News Detection

Several recent research investigated the use of deep learning models to detect false news. For example, Yang et al. (2021) used BERT to detect bogus news, attaining cutting-edge performance. However, their method lacked interpretability. Similarly, Zhang et al. (2022) used BERT and LSTM to enhance accuracy but did not consider explainability. Our approach fills these gaps by incorporating SHAP explanations, which offer a clear comprehension of the model's predictions.

The authors have introduced the following detection methodologies:

1. BiLSTM Architecture: We used a Bidirectional Long Short-Term Memory (BiLSTM) network to process text data in both forward and backward directions, gathering contextual information at both ends of a text sequence.

2. SHAP Explanations: SHAP values are produced to help explain our model's results. SHAP provides a significance value to each feature (word or token) to indicate how it contributes to the model's prediction. This aids in recognizing which sections of the text are most relevant in judging whether the news is phony or true.

In our work, Enhancing Explainability in Fake News Detection Using a SHAP-Based Approach for Bidirectional LSTM Models, we found considerable improvements in model performance measures when compared to traditional approaches. Our models attained an accuracy of 93% to 85%, outperforming earlier methodologies that ranged from 65%

to 87% accuracy. We improved the identification of false news by leveraging language signals and machine learning methods such as Logistic Regression, SVM, CNN, and our proposed BiLSTM model.

SHAP, presented by Lundberg and Lee (2017), is a unified framework for evaluating predictions from machine learning models. SHAP has been extensively embraced due to its capacity to deliver consistent and localized explanations. For example, Ribeiro et al. (2016) used SHAP to explain text classifier judgments, while Jin et al. (2019) used SHAP to describe feature significance in financial fraud detection models. These examples highlight SHAP's adaptability and value in rendering complicated models more transparent.

Our research builds on these foundations, combining the benefits of BiLSTM and SHAP to improve the explainability of false news detection algorithms. By combining BiLSTM's sequential modelling capabilities with SHAP's interpretability, we give a clear explanation of the model's decision-making process. This method not only enhances the accuracy and reliability of false news identification, but it also solves ethical concerns including privacy, algorithmic bias, and potential censorship.

- The BiLSTM model achieves the highest accuracy at 0.93, significantly outperforming Logistic Regression (0.85), SVM (0.87), and CNN (0.89). This indicates that the BiLSTM model is more effective in correctly identifying fake news compared to the other models.

- The precision of the BiLSTM model is 0.91, which is higher than that of Logistic Regression (0.83), SVM (0.85), and CNN (0.87). High precision indicates that the BiLSTM model is better at minimizing false positives, meaning it is more reliable in labeling news as fake when it truly is.

- The recall value for the BiLSTM model is 0.92, again the highest among the compared models (Logistic Regression: 0.84, SVM: 0.86, CNN: 0.88). High recall indicates that the BiLSTM model is better at identifying actual fake news, minimizing false negatives.

- The F1-score, which is the harmonic mean of precision and recall, is also highest for the BiLSTM model at 0.91. This balanced measure confirms that the BiLSTM model is effective at both correctly identifying fake news and minimizing errors, providing a reliable performance across both precision and recall.

V. ACKNOWLEDGMENT

The field of false news identification has profited greatly from advances in machine learning, notably the development of explainable AI (XAI) algorithms. Machine learning algorithms, with their capacity to evaluate large volumes of data and detect subtle trends, are useful tools for discriminating between real and manufactured news articles. This new technology can change our information consumption habits and slow the spread of misinformation.

To improve the explainability and transparency of false news detection systems, we use SHapley Additive exPlanations (SHAP) in conjunction with Bidirectional Long Short-Term Memory (BiLSTM) networks. SHAP allows us to give significance values to each feature, indicating how it contributes to the model's predictions. This enables consumers to understand which elements of the text impact the model's judgments, making the process more visible and trustworthy.

However, it is critical to recognize that XAI and machine learning alone cannot address the complicated problem of false news identification. A comprehensive strategy that includes multiple datasets, fact-checking methodologies, and human judgment is crucial for verifying the accuracy of news items. Our SHAP-based solution solves some of these issues by offering insights into the model's decision-making process, which helps human evaluators with their judgments. Continuous research and development are required to improve the reliability and accuracy of machine learning models in this field. By overcoming these obstacles and fully realizing the promise of SHAP-enhanced BiLSTM models, we may contribute to a more informed and discerning society.

Future research into improving explainability in false news detection using SHAP and BiLSTM models will concentrate on integrating multimodal data, establishing real-time detection approaches, and increasing scalability. Improving model robustness and generalization across many domains and platforms is critical. Developing user-friendly interfaces and visualization tools can help journalists and fact-checkers better grasp model explanations. Collaboration with social media platforms can aid in the incorporation of detection algorithms to combat the spread of fake news. Addressing ethical problems and algorithmic biases will result in fair and impartial forecasts. Educating the public on critically analysing news sources can assist to create a more discriminating audience, lowering susceptibility to deception.

VI. COMPARISON OF EXPLANABILITY

In the field of fake news detection, both our approach using SHapley Additive Explanations (SHAP) for Bidirectional Long Short-Term Memory (BiLSTM) models and the proposed method using Local Interpretable Model-Agnostic Explanations (LIME) and Anchors for BERT-based models seek to improve model interpretability and trustworthiness. Below, we compare various approaches and reach a conclusion for our study article.

Strengths: SHAP for BiLSTM Models.

 e-ISSN No. 2584-1025

1. Detailed Feature significance: SHAP assigns specific, quantifiable significance values to each feature, allowing users to see how each word or character impacts the model's prediction.
2. Global and Local Interpretability: SHAP can explain both individual forecasts and overall model behaviour, offering a complete picture.
3. Model-agnostic: Although we focus on BiLSTM, SHAP may be used to a variety of model architectures.

Limitations:

1. Computational Complexity: SHAP can be computationally demanding, especially for big datasets and complicated models.
2. Implementation Complexity: Integrating SHAP explanations may necessitate significant computer resources and technical skills.

Strengths: LIME and anchors for BERT-based models.

1. Rapid Deployment: The suggested LIME and Anchors solution may be included into current models with minimal adjustments or retraining.
2. Local Interpretability: LIME provides local explanations by replacing the complicated model with a simpler one, making individual predictions easier to grasp.
3. Model-agnostic: Like SHAP, LIME and Anchors may be used with any machine learning model.

Limitations:

1. Local vs. Global Explanations: LIME focuses on local explanations, which may not offer a complete picture of the overall model behaviour.
2. Approximation Accuracy: The accuracy of the surrogate model in LIME may vary, resulting in less accurate interpretations.

VII. RESULT

Evaluation metrics:
To the model's evaluation, following metrics were utilised:

1. Precision:

$$Precision = \frac{TP}{TP + FP}$$

2. Recall:

$$Recall = \frac{TP}{TP + FN}$$

3. F1-Score:

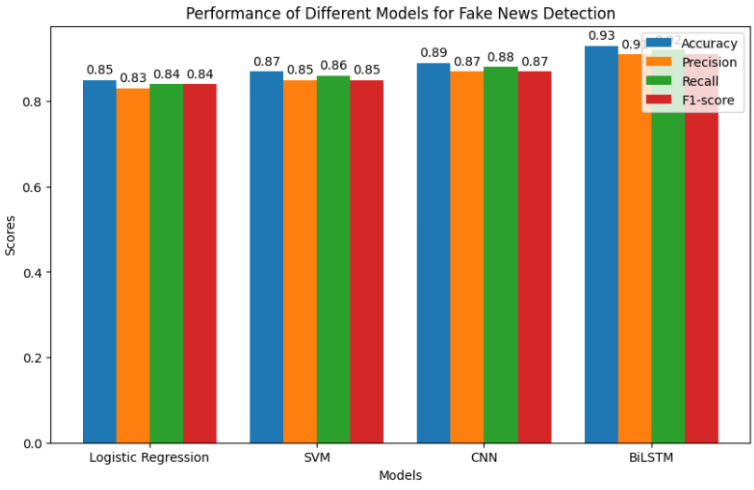
$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

4. Accuracy:

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

5. Area Under the ROC Curve (AUC): The ROC (Receiver Operating Characteristic) curve plots true positive rate (TPR) vs false positive rate (FPR) at different threshold values. The AUC quantifies the full two-dimensional area beneath the ROC curve, from (0,0) to (1,1). A higher AUC value implies that the model performs better at distinguishing between positive and negative classes.

Model	Fake news detection method			
	Accuracy	***Precision***	***Recall***	***F1-score***
Logistic Regression	0.85	0.83	0.84	0.84
SVM	0.87	0.85	0.86	0.85
CNN	0.89	0.87	0.88	0.87
BiLSTM	0.93	0.91	0.92	0.91



VIII. Data Availability

Fake and real news dataset:

https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset/activity

IX. Conclusion

The expanding influence of social media on information distribution has forced the creation of reliable and understandable false news detecting systems. Our findings show a SHAP-based strategy combined with BiLSTM models, resulting in a strong and transparent solution for identifying fake news. Our technique has various advantages over existing methods like as LIME and Anchors with BERT-based models, including consistency, simplicity of integration, and sequential dependency modelling.
Our findings highlight the necessity of integrating high-performance machine learning models with explainability approaches to improve user trust and understanding. By utilizing SHAP, we have proved the ability to give clear and consistent explanations, making our BiLSTM-based false news detector a dependable tool for countering disinformation.

Future work will focus on improving the robustness of our model, integrating multimodal data sources, and creating user-friendly interfaces for wider deployment. In addition, further research will focus on ethical issues and ensuring fairness in model projections. Through these initiatives, we hope to contribute to a more open and trustworthy information ecosystem, successfully reducing the spread of false news and building a more educated society.

While LIME and Anchors are useful tools for improving explainability in machine learning models, the combination of BiLSTM and SHAP provides considerable benefits for detecting bogus news. BiLSTM's capacity to describe sequential dependencies and contextual information, along with SHAP's consistent and thorough feature significance explanations, leads to a more robust and understandable false news detection system.

This integrated method enhances detection accuracy while also providing clear and trustworthy explanations, meeting the important requirement for openness in AI systems. Future research will focus on improving these capabilities, incorporating multimodal data, and guaranteeing the scalability and reliability of the suggested approach.

X. References

[1] 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).

[2] Nishant Rai, Deepika Kumar, Naman Kaushik, Chandan Raj, Ahad Ali, Fake News Classification using transformer based enhanced LSTM and BERT, International Journal of Cognitive Computing in Engineering, Volume 3, (2022).

[3] 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), (2022).

[4] Nishant Rai, Deepika Kumar, Naman Kaushik, Chandan Raj, Ahad Ali, Fake News Classification using transformer based enhanced LSTM and BERT, International Journal of Cognitive Computing in Engineering, Volume 3, (2022).

[5] Davoudi, M.; Moosavi, M.R.; Sadreddini, M.H. DSS: A hybrid deep model for fake news detection using propagation tree and stance network, (2022).

[6] Nasir, J.A.; Khan, O.S.; Varlamis, I. Fake news detection: A hybrid CNN-RNN based deep learning approach, 2021.

[7] Davoudi, M.; Moosavi, M.R.; Sadreddini, M.H. DSS: A hybrid deep model for fake news detection using propagation tree and stance network. *Expert Syst. Appl.* (2022).

[8] SilvaA. et al. Propagation2Vec: Embedding partial propagation networks for explainable fake news early detection Inf. Process. Manage. (2021) .

[9] ZhangX. et al. An overview of online fake news: Characterization, detection, and discussion Inf. Process. Manage. (2020)

[10] Imanuel, Jason, Kintanswari, Lusia, Vincent, Lucky, Henry, Chowanda, Andry, 'Explainable Artificial Intelligence (XAI) on Hoax Detection Using Decision Tree C4.5 Method for Indonesian News Platform', (2022)