International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3015-3019
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0380

Explainable AI-Based Detection of Misinformation Spread in Online Social Networks

Gunasheela R[1], Muhammed Nisham V[2], Aditi Menon [3], Adithyan V[4], Muhammed Fahim [5], Muhammed Musfar M K[6]

[1]Assistant Professor, Dept. of CSE, Yenepoya Deemed to be University. Bangalore, India

[2,3,4,5,6]Undergraduate Student, Dept. of CSE, Yenepoya Deemed to be University. Bangalore, India

Emails: *gunasheelar.blr@yenepoya.edu.in[1], nisham13v@gmail.com[2], aditi.menon07111@gmail.com[3], adithyan.v.adi2004@gmail.com[4], fahiiimm05@gmail.com[5], musfarmk007@gmail.com[6]*

Abstract

Social media is rapidly expanding and has completely transformed how people exchange and receive information. Social media platforms like Facebook, Instagram, and Twitter allows us to connect with others across the globe instantly and make the communication very easy and fast than ever before. Social Media platform is very convenient to share the useful information as well as fake information is spreading swiftly. In case the information is not true, people get influenced by the public opinion and act accordingly. and even lead to serious social or political consequences. As a result, identifying this false information in huge amounts of social media data has emerged as a significant research topic. Machine learning and artificial intelligence have found widespread application in the automated identification of misinformation. Nevertheless, many existing models function as "black-box" systems, thereby obscuring the processes underlying their decision-making When automated systems don't clearly show how they make decisions, people may start to question how reliable they really are. To find this issue, this research suggests using Explainable AI (XAI) in misinformation detection tools. By making the decision-making process more transparent and easier to understand, XAI can help build trust and make these systems feel more dependable. This framework clarifies the prediction generation process, utilizing machine learning and natural language processing methods, and employing tools like SHAP and LIME. These indicate that Explainable AI enhances the performance and transparency of misinformation detection. As a result, this methodology can bolster the dependability of detection systems, thereby fostering a more thorough comprehension of false information dissemination for researchers, social media platforms, and policymakers.

Keywords: *Explainable artificial intelligence; Fake news detection; Machine learning; Misinformation analysis; social media.*

1. Introduction

In today's digital era, the internet is one of the most important channels for communication. Millions of people are sharing the information, accessing news, and exchange ideas through social media platforms. Due to Internet speed and accessibility, information spreads across the globe. Meanwhile, the rapid flow of information offers benefits, it can also create an were, false information spreads rapidly. Features such as reposts, shares, and automated recommendation systems make it easier for misleading content to reach a large audience. This highlights the false information at the earliest. Traditional techniques, such as manual fact-checking, are efficient but time-consuming, which makes it challenging to implement it in the large volume of content produced daily on social media platforms (Castillo et al., 2011). In order to recognize false information, researchers have investigated the application of machine learning and natural language processing techniques. Early research was classified news based on the linguistic patterns using traditional methods like Support Vector Machines and Naïve Bayes. Subsequent methods provided deep learning models that could analyses textual context. Despite the fact that these models increased detection accuracy, many of the function are complex systems

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3015-3019
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0380

which is very hard to understand and taking decision. Explainable Artificial Intelligence (XAI) has been proposed as a solution to black box models being unable to explain their predictions. XAI techniques can be used to highlight the features used by the model to generate a specific decision; SHAP and LIME are two such techniques capable of identifying the words or substrings that contributed to the decision. Combining XAI with misinformation classification systems can increase the Explainability and trustworthiness of the system. This work aims to investigate how the Explainability of the comparison models may be increased by the utilization of XAI. The resulting models are not only expected to decrease the error of the classification, but also to increase the confidence of the user in the classification [1-5].

1.1. Misinformation in Online Social Networks:

False or misleading information that is spread without fact-checking is known as misinformation. This will be possible by the architecture and sharing methods of social media platforms. Users frequently repost unconfirmed content, which makes it possible for false information to spread quickly. Misinformation spreads faster than verified information because it is typically more sensational or stimulating, according to numerous studies. By doing likes, comments, and shares, this kind of material attracts greater attention and fosters user participation. Because of this, false information can spread far before being verified or rectified. Researchers, policymakers, and social media networks are becoming concerned about the widespread distribution and impact of false information. As a result, identify and stop the spread of false information is now an important research area.

1.2. Role of Explainable Artificial Intelligence

Explainable Artificial Intelligence (XAI) is built to make machine learning models more transparent and interpretable Usually, machine learning models are very good at predictions that they make but they do not tell anything about how they came to that prediction Lack of interpretability in models can turn off users to the power and the potential of automated decision - making XAI techniques are designed to identify the features that contribute to model predictions for example SHAP values attribute importance values to features then those features can be investigated to understand what caused the decision LIME creates local surrogate models that approximate the behavior of the large complicated models for the individual prediction Using Explainable AI techniques in the detection of misinformation will give insight into why the particular content has been classified as misleading. It will provide the researchers with a resource to validate model behavior and to verify that automated detection systems are fair and reliable [6-10].

2. Methodology

Data collection, preprocessing, feature extraction, model training, and Explainability analysis are all part of the system's organized design. Initially, social media platforms are used to gather the data. Posts, comments, and news stories classified as accurate or false information are included in this data. These datasets were collected from open sources. To make the collected data clean and prepared for analysis, it must go through the processing stage. The first phase involves removing stop words, converting all text to lowercase, and discarding any unnecessary punctuation. Tokenization is the following step, in which the phrases are divided into smaller units called tokens. The cleaned text is then transformed into numerical form so that the machine learning models can comprehend it. This process is known as feature extraction. Feature extraction techniques, such as TF-IDF, are utilized to assist. Sentiment scores and user interaction metrics (likes, shares, and comments) are also included. Lastly, machine learning models like Random Forest, Logistic Regression, and sophisticated transformer-based models can be trained using the processed data. These models are trained to determine if a piece of information is accurate or deceptive, and they are able to recognize trends. Explainability approaches are then used to understand the model's decision-making process [11-15].

2.1. Tables

Tables are mainly helpful to demonstrate the experimental information in a structured manner.

For example, Table 1 represents a sample dataset structure which is used for misinformation classification.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3015-3019
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0380

Table 1 Experimental Dataset Features Used for Misinformation Detection

Feature Type	Description	Example Value
Text Content	Raw textual data extracted from social media posts	"Breaking: Miracle cure discovered!"
Word Frequency (TF-IDF)	Measures the importance of words in a document	0.45
Sentiment Score	Detects the emotional tone of the post	Positive / Negative
User Engagement	Number of likes, shares, or comments	1200 shares
Source Credibility	Reliability score of the content source	High / Low
Hashtag Usage	Presence of trending hashtags	#BreakingNews
Label	Classification of the content	Real / Fake

2.2. Figures

Figure 1 represent the working of Proposed Systems. Data Collection → Data Preprocessing → Feature Extraction → Machine Learning Model → Explainable AI Analysis → Prediction Output

Figure 1 Work Flow

3. Results and Discussion

3.1. Results

The Experiment shows that the Machine Learning Models can predict the patterns which can lead to misinformation. When it comes to differentiating between authentic and deceptive posts, models trained on textual aspects do exceptionally well. Model performance is assessed using evaluation criteria like accuracy, precision, recall, and F1-score. Contextual feature integration increases classification accuracy, according to experimental findings. Furthermore, the explainability methods effectively draw attention to important characteristics that contribute to predictions. During analysis, words that are associated with dramatic statements, emotive language, or inflated assertions frequently obtain high priority scores.

3.2.Discussion

This research demonstrates that integrating Explainable Artificial Intelligence (XAI) with misinformation detection systems can improve both the accuracy of incorrect information identification and the comprehension of the results. Many of the current machine learning models are capable of detecting the existence of false information, but they frequently function in a way that makes it difficult for users to comprehend how they arrive at their findings. This problem is lessened when XAI is used because the model provides precise explanations in addition to its forecasts, making the system more transparent and accurate. The best part of the research is that specific components are crucial for spotting false information. These include how content is created, how people interact with it, and how quickly it spreads throughout the network. Posts that use strong emotional language, come from questionable sources, or suddenly get shared very widely are often more likely to be seen as false or misleading. Explainability techniques help improve how well the system works by making it clearer how different factors influence its decisions. One of the most important aspects of explainability is its real-world usefulness. It is very easy to understand for researchers when one algorithm is showing the labeled data as false and this information helps people to understand and take decisions based on that. This is very important to identify because if algorithm

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3015-3019
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0380

identify wrong labeling data , then it can create problems, such as unfair censorship. As a result, Explainable AI (XAI) is helping to know the gap between advanced technology and human understanding and thus it helps practically. Anyway, there are some restrictions to think about. The model's execution is dependent on the standard of the data used, and after misinformation samples keep changing, it can be challenging to continue accuracy over time. Additionally, attaching explainability characteristics can become greater computational complexity, which may affect real -time performance. There is also a chance that some users may misstate the explanations if they are not presented clearly. Overall, this learning recommends that XAI is a sspriceless approach for helping misinformation detection systems. It not only helps in recognising wrong information but also creates confidence by making the operation easier to understand. Future tasks can concentrate on enhancing scalability, helping real- time analysis, and including more different data sources to better represent how misinformation increases in online social networks shown in figure 2.

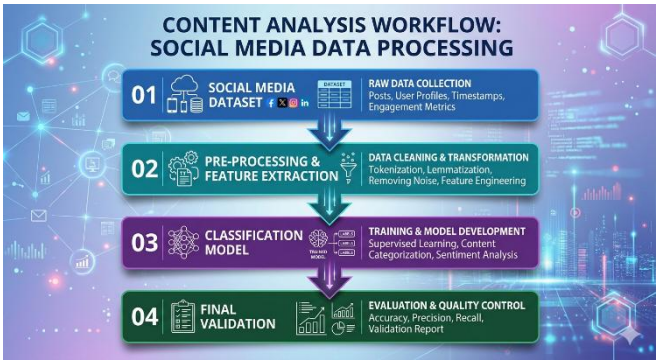

Figure 2 Process of the dataset used for misinformation detection.

Conclusion

This study examined how AI models detect misinformation in social networks. XAI, or Explainable AI, is an AI/ML integrated set of processes and methods for providing a human-understandable explanation of why information spreading through social media networks is misinformation. The effectiveness of XAI in misinformation detection is rooted in its ability to clarify the reason behind an AI system's classification of content as misinformation. As AI becomes more advanced, humans are challenged to comprehend and retrace how the algorithm came to a result. The whole calculation process is turned into a black box that is impossible to interpret. These black box models are created directly from the data. And, not even the engineers or data scientists who create the algorithm can understand or explain what exactly is happening inside them or how the AI algorithm arrived at a specific result. XAI uses features like words, network patterns, and user signals to identify fake news and provide explanations to gain trust, accountability and promote transparency for users, where the deep neural network models solely drive the predictions of the decision process. In essence, explainable AI misinformation detection uses techniques such as feature-importance methods like Shapley Additive explanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME) models, or graph-based explanations to reveal why content was labelled fake. These help users to get an idea about why the information is fake, rather than a simple prediction. Misinformation propagates rapidly through social media networks through a few well-known mechanisms, like bots, coordinated accounts, and the echo chamber effect. So, it's critical to create XAI systems that are reliable and safe. Another challenge in detecting misinformation spread is the scalability and interpretability of large graph-based models. These models are useful for studying how misinformation spreads, with their ability to capture connections and patterns in networks, they are very hard to understand when they become large and complex. Looking ahead, Future progress in Explainable AI (XAI) for misinformation detection depends on a few key areas. Firstly, XAI methods should be developed alongside social science ideas for finding how misinformation actually spreads across time, networks, and misleading languages.

Acknowledgement

The authors would like to express their gratitude to the academic mentors and faculty members who helped them prepare this research project.

References

[1]. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3015-3019
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0380

SIGKDD Explorations Newsletter, 19(1), 22–36.

[2]. Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities ACM Computing Surveys, 53(5), 1–40.

[3]. Shu, K., Mahudeswaran, D., & Liu, H. (2019) FakeNewsNet: A data repository with news content, social context, and dynamic information. Big Data, 8(3), 171–188.

[4]. Rashkin, H., Choi, E., Jang, J., Volkova, S., & Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news. EMNLP Conference Proceedings.

[5]. Bondielli, A., & Marcelloni, F. (2019). A survey on fake news and rumour detection techniques. Information Sciences, 497, 38–55.

[6]. Lundberg, S. M., & Lee, S. I. (2017) A unified approach to interpreting model predictions (SHAP). Advances in Neural Information Processing Systems.

[7]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier (LIME). KDD Conference.

[8]. Ahmed, H., Traore, I., & Saad, S. (2018). Detecting opinion spams and fake news using text classification. Security and Privacy, 1(1).

[9]. Castillo, C., Mendoza, M., & Poblete, B. (2011) Information credibility on Twitter. Proceedings of WWW Conference.

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

[11]. Shu, K., Wang, S., & Liu, H. (2019). Beyond news contents: The role of social context for fake news detection. WSDM Conference.

[12]. Monti, F., Frasca, F., Eynard, D., Mannion, D., & Bronstein, M. (2019). Fake news detection on social media using geometric deep learning arXiv preprint arXiv:1902.06673.

[13]. Pan, J., & Yang, Y. (2019). A survey on transfer learning.IEEE Transactions on Knowledge and Data Engineering.

[14]. Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable Artificial Intelligence: Understanding, visualizing and interpreting deep learning models. ITU Journal.

[15]. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable AI (XAI). IEEE Access, 6, 52138–52160.