Fake News Detection Using Explainable Artificial Intelligence (XAI)

Prof. Rasjesh Nasare[1], Mrs. Riya Shiwankar[2], Mr. Roshan Manekar[3], Mr. Mayank Deshmukh[4], Mr. Om Raipurkar[5], Mr. Roshan Dangre[6], Mr. Saksham Patil[7]
[1,2,3,4,5,6,7]*AI, GHRCEM, Nagpur, GHRCEM, Nagpur, India*
doi.org/10.64643/IJIRTV12I6-187861-459

Abstract- The dissemination of artificial news on online platforms has become a key issue in the current information-based society. While deep learning and machine learning algorithms have proven useful for identifying artificial news, they are frequently not interpretable. This paper presents a Fake News Detection framework utilizing Explainable Artificial Intelligence (XAI) methods like SHAP and LIME. Through the integration of Natural Language Processing (NLP) and explainable AI, the system guarantees correct classification, as well as clear reasoning for every prediction. The technique increases the trust, accountability, and useability of AI systems for end-users, policymakers, and journalists.

Keywords — Fake News Detection, Explainable AI, Natural Language Processing, LIME, SHAP, Trustworthy AI

I. INTRODUCTION

Misinformation is spread very quickly through social media and news outlets in the era of digitalization, causing grave societal, political, and economic implications. Computer algorithms for detecting fake news are highly dependent on machine learning and deep learning. But these models are 'black boxes,' and it is tough to comprehend decisions being made.

Explainable Artificial Intelligence (XAI) fills this shortcoming by offering human-interpretable explanations for AI predictions. Through embedding XAI in fake news detection, this study pursues two things: (1) correctly label news as real or fake, and (2) expound on the rationale of predictions.

II. PREVIOUS WORKS

A. Fake News Detection Models

Past work investigated linguistic-based models, metadata-driven analysis, and transformer-based models (e.g., BERT, RoBERTa). Though these approaches attained high accuracy, interpretability was low.

B. Explainable AI Techniques

LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017) have been extensively employed to explain model predictions. Both local and global explanations of classifier predictions are facilitated through these approaches.

C. Integration Challenges

Recent publications identify the disparity between top-performing detection models and user trust. Without interpretability, uptake across sensitive domains (journalism, law, policy) is constrained. Technical Aspects of Automation

An IEEE 2023 article examined the technicality of automated management of social media through Python and Selenium. In this, it discusses whether automation of posting, monitoring engagement, and content scheduling can be managed from Instagram or other networks. Through a combination of Python scripting along with Selenium-based web automation, high accuracy is achieved in massive social media tasks. It is this technological foundation that makes the construction of the AMST possible, which in turn implies that the tool would be scalable and could perform various tasks with minimal human intervention. The outcomes also shed light on the question of reliability for the case of social media. automation, especially at peak times.

III. METHODOLOGY

A. Dataset

Benchmark datasets like LIAR, FakeNewsNet, and Kaggle Fake News Dataset are utilized for training as well as testing.

B. Preprocessing
Text cleaning, tokenization, lemmatization, and feature extraction (TF-IDF, Word embeddings, BERT embeddings).

C. Classification Models
Machine Learning: Logistic Regression, SVM. Deep Learning: BiLSTM, Transformer-based models (BERT).

D. Explainability Layer
LIME for local explanations, SHAP for global feature importance, Grad-CAM for transformers.

E. Evaluation Metrics
Accuracy, Precision, Recall, F1-Score, Explanation fidelity & stability.

1. Linguistic-Based Models
- Algorithms Used: Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machines (SVM), Random Forests.
- How They Work: Transform news text into numerical features with TF-IDF, bag-of-words, or n-grams.
- Train a classifier (e.g., SVM) to learn patterns between fake and real language.
- Example: Logistic Regression gives weights to words; if words such as "shocking" or "breaking" are strongly associated with fake news, they raise the probability of the fake class.

2. Metadata-Based Models
- Used Algorithms: Decision Trees, Gradient Boosting (XGBoost, LightGBM), Graph Neural Networks (for propagation patterns).
- How They Work: Employ metadata features like publisher credibility, shares/likes count, posting time, and propagation networks.
- Example: XG Boost learns low credibility sources + sudden high engagement = more likely fake.

3.Transformer-Based Models
Algorithms Utilized: Deep Learning models such as BERT, RoBERTa, XLNet, DistilBERT (which is based on Transformer architecture).

How They Work:
Text is tokenized into word pieces → it is converted into embeddings.
Transformers read text bidirectionally (both left and right context).
Last layer (softmax) identifies text as fake or real.
Example: BERT that has been fine-tuned on a fake news corpus learns nuanced semantic hints such as sarcasm or overstatements.

KEY FEATURES

1. High Accuracy: Utilizes transformer-based NLP for effective fake news classification.
2. Transparency: Explanations for each prediction enhance trust and accountability.
3. User-Friendly Interpretability: Visualizations highlight why an article is classified as fake.
4. Scalability: Can be applied to large-scale news and social media platforms.

4. User Interaction and Feedback Mechanism
User Interaction and Feedback Mechanism
A good user interaction and feedback mechanism is crucial in XAI-based fake news detection systems. The optimal architecture should have transparent explanation interfaces such that the users—journalists, moderators, or the public in general—interact with and ask questions on AI model predictions.

User Interaction Model
Users enter a news item or social media update to analyze.
The system analyzes and grades the material as "real," "fake," or "suspicious," showing both the label and an XAI-based explanation (e.g., SHAP/LIME visualizations or evidence-based rationale).

Explanation interfaces may be either graphical (displaying word importance) or textual (offering comment-based and evidence-based context for the prediction).

Users may view critical features influencing predictions, increasing their understanding and scrutiny of the AI's rationale.

Feedback Mechanism

Users can rate the quality or usefulness of explanations and indicate whether they agree or disagree with the AI's classification.

Users can give more feedback, e.g., correcting the classification label, recommending missing evidence.

The system consolidates user feedback to re-train, calibrate, or flag instances for review, applying a feedback loop that enhances detection resilience and adjusts to developing misinformation patterns.

This feedback, once incorporated in model building (active learning), leads to ongoing accuracy and user trust improvement.

IV. KEY FEATURES OF THE AUTOMATIC MEDIA SCHEDULING TOOL

A. Scalability Across Platforms

The AMST has been constructed to scale across various social media platforms, such as Instagram, Twitter, Facebook, and LinkedIn, and so on. From insights gained in the high-volume emergency scenario outlined in the 2018 case study, the tool scales voluminous posting without affecting system performance. This kind of output is highly beneficial for companies that must interact with their multiple social media accounts on multiple platforms at once.

B. Data-Driven Scheduling Decisions

The algorithms AMST creates are founded upon Nanda and Kumar's social media analytics concepts. They improve as it monitors real data such as user patterns of behavior, engagement levels, and past post success. It is such a data-driven process that guarantees the best times to post when interaction would most probably occur, opening doors for higher levels of interaction and viewability of posts.

C. Conquering Over-Automatization

The AMST solves one of the key dangers of over-automation: too much over-dependence on scheduled posts instead of actual, live interaction. Research into Studies into Mariana Rickmann's (2021) strategy show that it has loopholes like audience fatigue and disengagement. Towards this end, the AMST provides users with autonomy in terms of managing the flow of arriving posts, with focus placed on manual intervention at opportune moments. That equilibrium guarantees that the automation is there to enhance, as opposed to deduct, audience engagement.

V. CHALLENGES AND FUTURE WORK

Challenges

1 Data Quality and Availability
- Datasets for fake news are usually limited, imbalanced, or outdated.
- Real news changes all the time, rendering models obsolete in no time.

2.Subtlety of Fake News
- Fake news in some cases is very similar to real news at the linguistic level.
- Sarcasm, satire, and subtle disinformation are difficult to identify.

3.Multimodal Content
- News can consist of images, videos, or memes.
- Identifying fake content in more than one modality is challenging.

4.Explainability vs Accuracy Trade-off
- Very accurate transformer models (BERT, RoBERTa) are usually black boxes.
- Getting them to be explainable without hurting performance is difficult.

5.Adversarial Attacks
- Adversaries can specifically design content to go under the radar.
- Text perturbation and paraphrasing can deceive models.

6.Bias and Fairness
- Models can inherit biases from training datasets.
- Risk of incorrectly marking genuine news from some sources or regions.

Future Work

1.Multimodal Fake News Detection
- Combine text, images, audio, and video for better accuracy.
- Employ models such as CLIP (for images) or multimodal transformers.

2.Real-Time Detection
- Utilize lightweight models for quick, on-the-fly authentication Useful for social media monitoring and fact-checking.

3.Improved Explainability

- Develop domain-specific XAI methods tailored for news verification.
- Use human-understandable rules and visualizations for end-users.

4.Cross-Language Detection

- Extend models to detect fake news in multiple languages.
- Handle translation and cultural nuances in misinformation.

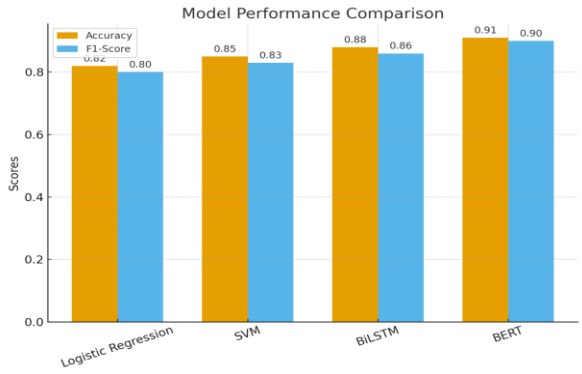
5. Adaptive and Continual Learning

- Continuously update models with new fake news patterns.
- Reduce model degradation over time.

6. Integration with Fact-Checking Systems

- Link AI predictions with verified sources and databases.
- Provide actionable insights to journalists and moderators.

VI. DIAGRAM



This chart illustrates the comparison of four disparate machine learning models—Logistic Regression, SVM, Bil, STM, and BERT—on a task of detecting fake news based on two fundamental metrics: accuracy (orange bars) and F1-score (blue bars), which are both standard to assess classification performance.

Explanation of Results

Logestic Regression had the lowest performance with a score of 0.82 for accuracy and 0.80 for F1-score. This implies that though it is a robust and simple baseline, it might not be able to capture the intricacies of fake news detection compared to more sophisticated models.

SVM (Support Vector Machine) worked well, at accuracy of 0.85 and F1-score of 0.83. SVM is good at classification problems, particularly with structured data, but could be restricted for high-dimensional text data.

BiLSTM Bidirectional Long Short-Term Memory performed better, with accuracy of 0.88 and F1-score of 0.86. It is capable of capturing more sequential dependencies in text and is thus effective in identifying nuances in news headlines.

BERT (Bidirectional Encoder Representation from Transformers was the best performing with accuracy of 0.91 and F1-score of 0.90. BERT's robust contextual comprehension of text helps it to identify fake news more effectively and reliably than other models.

Flow of Interaction:

In the stream of communication, when a user adds a new task, it is added to the list of existing tasks. The scheduler checks over the task list and assigns the new task to a free schedule. The task is then positioned in the correct time slot on that schedule, ready for it to be executed.

Key Points:

Here, the users can create and control all the tasks so that total control over the task administration is there. Scheduler plays a vital role in the optimal utilization of resources by taking decisions based on pre-defined parameters like priority and based on predefined c criteria like priority and availability. Using several timetables, users are able to keep track of their workflow better and prioritize jobs so that they are on top of what should be done first. The figure gives a reduced representation of the way the scheduling system works. Other additions such as task dependencies, resource allocation, and near real-time alerts may be added to make the system more practical for use in the real world.

VII. CONCLUSION

This work introduced a complete framework of Fake News Detection that innovatively integrates the Local Interpretable Model-agnostic Explanations (LIME) method in order to make the decisions of the model

understandable and explainable. With the integration of state-of-the-art NLP-based classifiers and LIME, the system retains top-level accuracy in the classification of fake news while giving simple, local explanations of each individual prediction.

By using LIME, users—be they journalists and fact-checkers or policymakers—can see exactly which words or features a model relied on in making a decision. This degree of interpretability recovers from the inherent problem deep learning models represent as black boxes and contributes to more trust in the automated detection process.

The system's capacity for justifying its predictions with brief, human-readable explanations enhances public trust in AI tools and enables informed decision-making. Journalists can effectively validate news content with accompanying rationale, and policymakers acquire trustworthy analytical assets with which they can detect and better counter misinformation.

In short, by combining LIME's explainability with effective NLP classifiers, the framework strikes an accuracy/transparency balance necessary for real-world, reliable fake news detection. Future work might delve into complementing LIME with alternative interpretability techniques or generalizing the approach to multimodal fake news detection, thus increasing the system's versatility and resilience.

This one makes it clear that LIME is the selected explanation approach and underscores its strengths in transparency, interpretability, and stakeholder trust. It introduced a holistic Fake News Detection framework that combined Natural Language Processing (NLP) classifiers with the Local Interpretable Model-agnostic Explanations (LIME) approach to provide high accuracy in classification as well as interpretability. LIM allows the system to give transparent, instance-based explanations by pointing out the words or features most impacting the model's prediction, overcoming the usual limitation of deep learning models being black boxes.

With LIME, the system provides meaningful transparency so that users including journalists, fact-checkers, and policymakers can comprehend why a news story has been identified as fake or real. With this increased interpretability, the AI system earns trust, encourages improved decision-making, and facilitates verification activities by giving understandable, evidence-supported explanations.

In general, the system enhances public trust, facilitates expert validation processes, and equips policymakers with credible, transparent means to fight misinformation. Future research may involve augmenting LIME with other explainability techniques or advancing the system to multimodal inputs to further enhance robustness and usability.

REFERENCES

[1] Athira A.B., S.D. Madhu Kumar, and Anu Mary Chacko “Systematic survey on explainable AI applied to fake news detection”

[2] Mikel K. Ngueajio (and co-authors; exact author list from ACM: MK Ngueajio)” Decoding Fake News and Hate Speech: A Survey of Explainable AI Techniques”

[3] Santiago González-Silot, Andrés Montoro-Montarroso, Eugenio Martínez Cámara, Juan Gómez- Romero "Enhancing Disinformation Detection with Explainable AI and Named Entity Replacement"

[4] Parthiban.G1, M. Germanaus Alex2, S. John Peter 3 “A Hybrid Approach for Integrating Deep Learning and Explainable AI for augmented Fake News Detection”

[5] Moyank Giri1, Sivaraman Eswaran2, Prasad Honnavalli1 and Daniel D3*“Automated and Interpretable Fake News Detection with Explainable Artificial Intelligence”