UtilitasMathematica

ISSN 0315-3681 Volume 122 (2), 2025

A Hybrid Framework for Identifying Fake News and Deceptive Information in Social Media with Mitigation Strategy

Priya Sharma[1], Mohd Waris Khan[2,*]

[1,2]Department of Computer Application, Integral University, Lucknow, Uttar Pradesh, India

priyashar@student.iul.ac.in[1], wariskhan070@gmail.com[2]

ABSTRACT: The rapid proliferation of fake news and deceptive information across social media platforms poses significant challenges to society, politics, and the global economy. This paper presents a hybrid framework for identifying and mitigating fake content, integrating principles of digital forensics with advanced detection strategies. The study highlights the scope of digital forensics—including disk, network, email, wireless, database, malware, and memory forensics—demonstrating their role in preserving, retrieving, and documenting digital evidence for legal and investigative purposes. Beyond forensic applications, the research underscores the disruptive influence of misinformation on democratic processes, financial markets, and public trust. A key focus is placed on the development of a Fake Content Detection System that leverages both forensic techniques and computational intelligence to address the evolving nature of online deception. The study also emphasizes the role of mitigation strategies, where people, governments, digital firms, and civil society collectively contribute to combating misinformation through coordinated policies and technological safeguards. By analyzing the mechanisms of social media communication and unverified content dissemination, the framework offers practical pathways for strengthening trust, accountability, and resilience in digital ecosystems. The findings emphasize the urgency of updating technological solutions to ensure accurate information dissemination and mitigate the adverse societal impacts of fake news.

Keywords: Fake News Detection, Digital Forensics, Social Media, Deceptive Information, Machine Learning.

1. INTRODUCTION

The rise of social media has profoundly transformed the way information is generated, consumed, and disseminated across the globe [1]. The majority of application areas, such as cloud computing, supply chains, manufacturing, smart cities, smart homes, and agriculture, provide a clear picture of the Internet of Things [2]. Platforms such as Facebook, Twitter (X), Instagram, and YouTube have emerged as primary avenues for news distribution, public debates, and social interaction, often surpassing traditional mass media in terms of reach and influence [3]. This digital transformation has democratized information sharing, empowering individuals to voice their opinions and access diverse perspectives. However, the same characteristics that make social media powerful—its openness, immediacy, and global accessibility—also render it highly vulnerable to the proliferation of fake news and deceptive content. Unlike conventional media outlets, where editorial checks regulate the flow of information, social media

allows unfiltered and unverified content to circulate rapidly, blurring the boundary between truth and fabrication [4, 5].

The consequences of misinformation are far-reaching, cutting across multiple domains of society [6]. In politics, fake news has been strategically deployed to manipulate electoral outcomes, influence public opinion, and propagate extremist ideologies. In healthcare, the spread of deceptive claims—ranging from misleading medical advice to false narratives about vaccinations—can jeopardize public health and even result in fatalities [7]. In financial markets, fabricated stories and rumors have been exploited to manipulate stock prices, destabilize investor confidence, and cause severe reputational and economic damage to individuals and organizations. These examples illustrate that misinformation is not a trivial byproduct of online interaction but a complex socio-technical threat with social, economic, legal, and security implications [8, 9].

Beyond its direct consequences, misinformation has a corrosive effect on trust and credibility in the digital ecosystem. The persistent circulation of deceptive content erodes public confidence in credible media outlets and undermines the perceived reliability of digital communication platforms. As users encounter conflicting or ambiguous information, skepticism often extends to authentic sources, resulting in widespread uncertainty and the weakening of collective trust. This phenomenon fuels social polarization, disrupts democratic processes, and amplifies cyber threats, as misinformation is frequently intertwined with malicious activities such as phishing, identity fraud, and coordinated disinformation campaigns [10]. Consequently, the urgent need for robust detection and mitigation frameworks is evident to safeguard societal stability, information integrity, and cybersecurity.

In addressing this challenge, digital forensics has emerged as a crucial discipline for ensuring transparency, accountability, and justice in digital ecosystems. By systematically retrieving, preserving, and analyzing digital evidence, forensic investigations provide insights into the origin, dissemination, and authenticity of information [4]. Specialized branches such as disk forensics, network forensics, email forensics, and malware forensics enable investigators to trace the pathways of misinformation, identify malicious actors, and expose deceptive campaigns. Such capabilities are indispensable not only for law enforcement and legal proceedings but also for reinforcing user trust in online interactions. The investigation of the events and conditions that led to an intelligent system's failure, including determining whether malicious activity was the cause of the failure and identifying the culpable entity or entities in such a scenario, is known as Digital forensics [11, 12].

Building on these foundations, this paper proposes a hybrid framework with mitigation strategies to counter fake news and deceptive content in social media. The novelty of the framework lies in its integration of digital forensics and machine learning techniques to create a comprehensive detection and response system. On one hand, machine learning and natural language processing (NLP) techniques enable large-scale analysis of user behavior, content patterns, and semantic structures to detect anomalies, fraudulent activities, and deceptive narratives [13]. On the other hand, digital forensic methods ensure that the evidence is preserved, verifiable, and legally admissible, thus bridging the gap between technological solutions and real-world applicability.

The framework also incorporates continuous monitoring mechanisms to identify evolving misinformation tactics and applies fraud mitigation strategies to neutralize malicious campaigns before they escalate. Moreover, it emphasizes evidence collection optimization through advanced computational methods, ensuring accuracy, scalability, and efficiency. Importantly, the study introduces guidelines for mitigation strategies,

focusing not only on detection but also on proactive measures, such as user awareness, authentication protocols, and platform-level interventions-that can reduce the impact of deceptive information.

Ultimately, the contributions of this research extend beyond the academic domain to practical implementation. By combining the investigative rigor of digital forensics with the predictive intelligence of machine learning, the proposed hybrid framework aspires to strengthen defenses against fake news, safeguard democratic processes, and preserve the integrity of social and financial systems. The findings provide actionable insights for policymakers, forensic investigators, media regulators, and technology developers, offering a pathway toward a more resilient and trustworthy digital communication environment.

The main objectives of this research paper are as follows:

- To explore various types of digital forensics, including disk forensics, network forensics, and malware forensics, for enhancing the retrieval and analysis of digital evidence.
- To address the challenges of fake news and misinformation on social media by identifying and categorizing different types of fake content.
- To apply machine learning techniques for anomaly detection and fraud mitigation in social media, with emphasis on continuous monitoring and user authentication.
- To evaluate and optimize digital evidence collection methods using advanced technologies for improved effectiveness.

The remainder of this paper is structured as follows: Section 2 provides an overview of digital forensics, outlining its key principles, types, and role in supporting misinformation detection. Section 3 reviews previous research on fake news detection, digital forensic methods, and hybrid approaches, highlighting existing gaps. Section 4 introduces the proposed hybrid framework, detailing its architecture and functional components. Section 5 explores potential applications of the framework across political, financial, healthcare, and corporate domains. Section 6 discusses mitigation strategies, emphasizing the collaborative role of individuals, governments, digital platforms, and civil society. Section 7 addresses challenges and limitations of current detection systems, including technical, ethical, and scalability issues. Finally, Section 8 concludes the study by summarizing key findings and outlines directions for future research and practical implementation.

2. DIGITAL FORENSICS OVERVIEW

Digital forensics is a specialized discipline within cybersecurity that focuses on the identification, preservation, retrieval, and documentation of digital evidence for investigative and legal purposes. It encompasses a wide range of forensic domains, including disk forensics, network forensics, wireless forensics, database forensics, malware forensics, email forensics, and memory forensics. Each of these domains addresses unique aspects of digital investigation, enabling forensic experts to analyze diverse data sources and uncover valuable evidence [14].

The growing dependence on technology and the expansion of social media platforms have amplified the importance of digital forensics in contemporary society. Social media, in particular, impacts various dimensions such as politics, commerce, and cybersecurity, often becoming both a source of valuable evidence and a platform exploited for malicious activities. Consequently, digital forensics tools and techniques

play an indispensable role in assisting forensic teams as they navigate increasingly complex cases involving multiple devices, platforms, and communication channels.

a) The Difficulties Digital Forensics Faces:

Despite its growing relevance, digital forensics encounters several challenges that hinder effective investigations. One significant issue is internet dependency, as the widespread use of online platforms has created an immense volume of data, complicating evidence collection and analysis. In addition, the proliferation of personal computers (PCs) further adds to the complexity, as the sheer number of devices increases the difficulty of tracing, retrieving, and managing digital artifacts across diverse systems and environments. These challenges highlight the urgent need for improved forensic methods capable of handling the vast, distributed, and ever-evolving nature of digital information.

b) Hacking Tools and Data Challenges;

Another pressing set of challenges arises from the accessibility of hacking tools and the difficulty of prosecution in the absence of solid evidence. Hacking tools and malicious software have become widely available, empowering cybercriminals to exploit vulnerabilities with relative ease. At the same time, legal prosecution becomes problematic without reliable and admissible forensic evidence, undermining judicial outcomes.

Additionally, modern systems are often equipped with large storage capacities measured in terabytes, which significantly increases the time and resources required for data analysis. This problem is compounded by the pace of technological advancements, which demand continuous updates and adaptations of forensic tools and methodologies to remain effective against new threats and techniques.

c) Implementation and Data Quality Assurance:

In addition to addressing technical challenges, the implementation of digital forensic techniques must also consider the quality and reliability of data. This study emphasizes the use of third-generation deepfake datasets to support model implementation, ensuring that experimental setups reflect realistic and emerging threats [15]. Ensuring data quality is paramount, as flawed or incomplete datasets can compromise the validity of findings and lead to erroneous conclusions.

Furthermore, effective knowledge base management with comprehensive and well-structured information is essential for building robust forensic solutions. By integrating high-quality data and systematic management practices, digital forensics can achieve greater accuracy, reliability, and scalability in combating the complex challenges posed by fake news, deceptive content, and broader cybersecurity threats.

2.1 Different Kinds of Fake Content:

Fake content manifests in multiple forms, each crafted with the intention of deceiving, misleading, or manipulating audiences. One of the most common types is deliberate misinformation, where fabricated stories are intentionally circulated to mislead, such as falsely accusing someone of committing a crime. Another prevalent form is clickbait headlines, which use sensational or exaggerated titles to attract readers, often directing them toward misleading or false information. Closely related are false headlines, where the title of a news piece is misleading or misrepresents the actual content of the article, thereby creating false impressions among readers.

Satirical content is another category, where humor, parody, or exaggeration is used to distort reality, often with the aim of critiquing societal issues. Although satire is not

always intended to deceive, it can still mislead those who interpret it literally. False context arises when genuine content is presented in a misleading manner, such as associating real images with unrelated events to manipulate narratives. Similarly, manipulated content involves altering authentic media, such as editing photos or videos, to misrepresent events. This is common in the form of memes and viral images widely shared across social media.

A particularly advanced and dangerous form is deepfake content, where artificial intelligence and deep learning technologies are employed to create forged images, videos, or audio clips. Deepfakes can convincingly depict individuals saying or doing things they never actually did, leading to misinformation, privacy violations, and reputational harm [16, 17]. Collectively, these categories of fake content illustrate the wide spectrum of tactics used to mislead audiences and reinforce the critical need for effective detection mechanisms [18].

2.2 The Necessity of Fake Content Detection Systems:

The widespread dissemination of fake content highlights the urgent need for effective detection and mitigation systems. The impact of false information is particularly evident in politics, where misinformation campaigns have been strategically employed to manipulate public opinion and electoral outcomes [19]. For instance, during the 2016 U.S. presidential election, groups in Macedonia deliberately created pro-Trump content that gained viral traction, while Russian propaganda organizations fabricated stories to portray an artificial sense of widespread support for Trump's campaign [20]. Although the exact influence of such activities remains debated, the use of fake news as a political tool poses significant risks to democratic stability.

Another major consequence is the undermining of reputable media sources. Politicians and interest groups frequently exploit public susceptibility to fake news to delegitimize credible outlets, casting doubt on factual reporting that contradicts their narratives [21, 22]. This tactic erodes trust in journalism and damages the broader integrity of information ecosystems. Furthermore, political parties in countries such as the U.S. and the U.K. have weaponized the term "fake news" itself, using it to dismiss valid criticisms, obscure evidence, and discredit expertise.

These trends underscore the necessity of fake content detection systems that not only identify misleading content but also safeguard democratic institutions, protect public trust, and maintain the credibility of reliable information sources [23, 24]. Without such mechanisms, societies remain vulnerable to manipulation, polarization, and the erosion of informed decision-making.

2.3 Effects on Financial Markets:

The influence of misinformation extends deeply into the financial domain, where false information can trigger volatility, disrupt investor confidence, and damage reputations. A notable example occurred in 2015, when a false report about a French-English truce caused a sudden 5% surge in stock prices on the London Stock Exchange, highlighting the market's sensitivity to misinformation. Similarly, cryptocurrency markets are particularly vulnerable, with misleading information spread via platforms such as Telegram and Twitter artificially inflating the value of specific digital currencies.

Businesses are also frequent targets of politically motivated fake news campaigns designed to discredit organizations or harm their public image. Such misinformation not only causes financial losses but also erodes stakeholder confidence. Celebrities and public figures are likewise vulnerable, often subjected to fabricated stories and death

hoaxes aimed at damaging their reputations. With the rise of deepfake technology, reputational threats have grown more severe; manipulated videos or images are increasingly used to portray celebrities in false and harmful contexts, including non-consensual explicit content [25].

Beyond market and celebrity impacts, misinformation can cause serious harm to communities. For example, spreading false accusations, such as blaming an elderly woman for child abduction, can lead to social ostracism, panic, and long-lasting trauma for innocent individuals. These examples demonstrate that fake news is not merely an inconvenience but a serious socio-economic threat that affects individuals, businesses, and entire communities. Thus, proactive measures and robust counter-strategies are essential to safeguard the stability of financial systems and protect societal well-being from the pervasive dangers of misinformation.

3. REVIEW OF PREVIOUS WORK

Table1 presents a comprehensive review of previous research on fake news and deceptive information detection in social media. Each entry includes the paper title with its publication year, the technique or methodology applied, the problem addressed, and the key outcomes. This comparative overview offers valuable insights into existing methods, their effectiveness, and the research gaps that justify the need for developing a hybrid framework with mitigation strategies.

Table 1: Summary of Related Work

S. No.	Reference/Paper Title	Method/ Technique Used	Problem Addressed	Key Findings/Outcomes
1.	Emerging Trends in Digital Forensics: Investigating Cybercrime, 2025 [1]	Forensic tools and network security	Tracing cryptocurrency transactions to combat money laundering, ransomware payments, and illicit trading on the dark web.	AI technologies enhance forensic detection and speed up investigation processes.
2.	IoT Forensics: Addressing Challenges and Establishing a Framework for Exploring Digital Forensics, 2025 [2]	IoT and cloud computing	Reviews the current state of IoT forensics and identifies challenges faced by investigators.	A framework is developed to address challenges and improve data analysis efficiency.
3.	Text Analysis for Anomaly Detection and Fraud Mitigation in Social Media using R, 2024 [3]	LNRE (Large Number of Rare Events) model	Detecting anomalies and fraudulent activities in social media	Highlights the importance of preventive measures such as continuous monitoring and user authentication to mitigate fraudulent activities on social networks.
4.	Survey of Machine	Natural	Fake news	Provides

	Learning Techniques for Arabic Fake News Detection, 2024 [4]	Language Processing and Deep Learning	detection in Arabic, a low-resource language, to contextualize the current state of research	researchers with a roadmap for advancing studies in Arabic fake news detection.
5.	Duped by Bots: Why Some Are Better than Others at Detecting Fake Social Media Personas, 2024 [5]	Signal Detection Theory, Bot Indicators	The impact of bots amplified by anonymity and low-cost deployment	Findings suggest the use of warning indicators such as "myside" or bot scores.
6.	A Scientometric Analysis of Deep Learning Approaches for Detecting Fake News, 2023 [6]	Deep learning approaches	Identifying fake news on online platforms	Strengthened through qualitative analysis of published articles.
7.	Certain Investigations on Drug Recommendations Using Machine Learning Techniques, 2023 [7]	Deep learning techniques	Serious side effects on the human bodies due to the indiscriminate usage of medicines without the doctor's prescription	Proposed system achieves higher precision, recall, F-measure, and accuracy than other methods.
8.	A Novel Approach for Behavior Prediction Using Sentiment Analysis on Social Media Data using Machine Learning Techniques, 2022 [8]	Machine learning techniques	Analysis of customer feedback (positive, negative, neutral) on social media	Develops effective sentiment and behavioral analysis techniques.
9.	Development of an Approach for Image Forgery Detection Using Machine Learning, 2022 [9]	Machine learning algorithms	Detecting image copy-move, image splicing, image retouching, and resampling forgery	Improved the efficiency of forgery detection rate and reduced the false positive rate.
10.	Digital Forensics for Malware Classification: Binary Code to Pixel Vector Transition, 2022 [10]	Static and dynamic malware analysis	Texture-based malware classification	Deep learning methods show highest accuracy in malware detection.
11.	Digital Forensics AI: Evaluating, Standardizing and Optimizing Evidence Mining Techniques, 2022 [11]	Classification, Regression, and Clustering algorithm in DFAI technique	Comparative review of optimization techniques	Formalizes the DFAI concept with its core components.

12.	Digital Image Forensics Based on Machine Learning for Forgery Detection and Localization, 2021 [12]	Support Vector Machine (SVM)	Image forgery detection technique	Improves detection speed using preprocessing and feature reduction.
13.	Deep Learning on Image Forensics and Anti-Forensics, 2021 [13]	CNN and GNN models	Improving forensic and anti-forensic methods	Enhances detection accuracy and image sharpening techniques.
14.	A Survey of Machine Learning Applications in Digital Forensics, 2021 [14]	IoT and machine learning technique	Classification for feature extraction	Reduces manual effort and supports a secure working environment.
15.	Deepfake Videos in the Wild: Analysis and Detection, 2021 [15]	Supervised and unsupervised learning	Generation and analysis of Deepfakes	Improves detection performance and techniques.
16.	Deepfake Video Forensics Based on Transfer Learning, 2020 [16]	Deepfake and transfer learning	Detecting manipulated photos and videos	Enhances neural network efficiency for Deepfake detection.
17.	Detecting Deepfake Videos Using Euler Video Magnification, 2020 [17]	Spatial decomposition and temporal filtering	Identification of Deepfake videos	Extracts hidden features with Euler magnification technique.
18.	Data Mining Approach for Digital Forensics task with Deep Learning techniques, 2020 [18]	CNN-based deep learning	Data clustering and classification	Proposed CNN reduces computational load and processing time.
19.	Bio-Inspired Computing for Outlier Detection: Select Studies in Web 3.0 Domain, 2019 [19]	ML approaches and bio-inspired algorithms	Detecting/Identification of outliers in Web 3.0 domain	Combines search strategies for global convergence with higher accuracy.
20.	Improving the Classification of Tiny Images for Forensic Analysis, 2019 [20]	KNN and CNN classifiers	File identification and forensic classification	Detects policy violations with higher accuracy.

The review of previous works reveals that a variety of approaches ranging from machine learning and deep learning models to linguistic, network-based, and hybrid techniques—have been explored for detecting fake news and deceptive information in social media. While these studies demonstrate significant progress, most focus primarily on detection without adequately addressing mitigation strategies.

Furthermore, challenges such as scalability, adaptability to evolving misinformation patterns, and cross-platform applicability remain largely unresolved. These gaps highlight the need for a hybrid framework that not only enhances detection accuracy but also provides effective mitigation guidance to curb the spread of deceptive information in social media ecosystems. In line with this review, the key research gaps and findings are summarized below.

3.1 Research Gap:

- There is no clear consensus on the definition and scope of fake news, which creates ambiguity in detection and mitigation strategies.
- Existing methodologies for detecting deceptive information often fail to keep pace with the dynamic and rapidly evolving nature of social media platforms.
- While several detection techniques exist, there is a lack of holistic frameworks that integrate both detection and mitigation strategies, limiting their practical effectiveness.
- Studies focusing on low-resource languages and cross-platform misinformation remain scarce, reducing the generalizability of current solutions.

3.2 Findings:

- Previous research highlights the significant impact of fake news and deceptive information on public perception, decision-making, and social stability.
- Detection efforts increasingly rely on machine learning, deep learning, and hybrid approaches, which demonstrate improved accuracy compared to traditional methods.
- Despite advancements, most studies prioritize detection while overlooking mitigation, leaving a critical gap in controlling the spread of misinformation.
- AI-driven solutions show promise in enhancing detection accuracy and scalability, but further work is needed to adapt them to diverse languages, contexts, and platforms.

4. PROPOSED HYBRID FRAMEWORK

Fake news has become a major issue in today's digital world, influencing public opinion, political landscapes, and even financial markets. Traditional detection methods have limitations in scalability and accuracy. Hence, this research focuses on a hybrid machine learning model that improves reliability and robustness in detecting fake content. Logistic Regression is effective in identifying linear patterns in text data, while Random Forest handles non-linearity and reduces overfitting. By combining both, we leverage their strengths, ensuring better generalization and improving classification accuracy. XGBoost is good at fine-tuning, but the other models help in different ways, such as handling linear patterns or noise, making the system more robust.

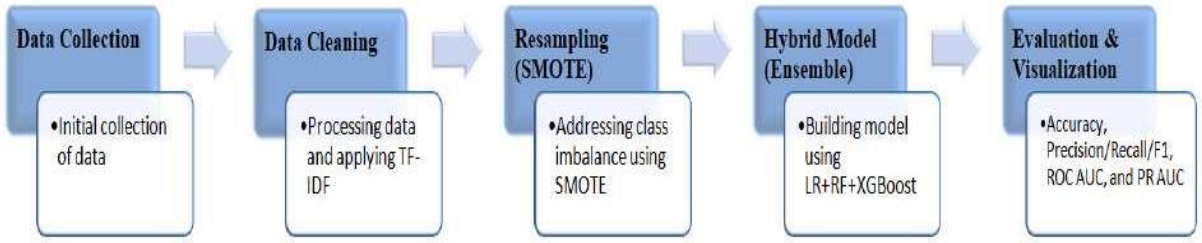

Fig. 1: Hybrid Framework

Figure 1 illustrates the Hybrid Framework, which is structured through the following sequential steps.

4.1. Data Collection: The first step in building any analytical or predictive model is data collection, which involves acquiring and quantifying raw data from multiple sources to serve as the foundation for analysis and model development. Much like bricks and mortar are essential for constructing a house, relevant and high-quality data is fundamental for constructing a reliable data model.
The success of the entire process is directly influenced by the accuracy, completeness, and relevance of the collected data. This stage of the data pipeline requires systematic gathering of raw information, as the principle "Garbage In, Garbage Out" emphasizes that poor-quality input inevitably leads to unreliable outcomes. Therefore, the primary objective of data collection is to ensure that all pertinent information is gathered to effectively address the research question or solve the defined problem.

4.2. Data Cleaning: Data cleaning is an essential step that ensures the dataset is accurate, consistent, and ready for analysis. It involves fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data. In the context of text processing, data cleaning is performed before vectorizing the text using TF-IDF. The process includes converting all text to lowercase, removing URLs, special symbols, and non-alphabetical characters, collapsing extra spaces, and filtering out common stopwords. The development process utilized a standard Python data science stack: textual data was cleaned with the re module, structured data operations like loading and preprocessing were handled with pandas, and the machine learning lifecycle from feature extraction to model training was managed using scikit-learn.

For Example, the Text: "OMG! This new song is FIRE 🔥! Download it here: https://music.com/best-song.mp3 #MustListen" is first converted to lowercase, then URLs, symbols, and hashtags are removed, followed by the elimination of stopwords. The final cleaned text, "omg new song fire download mustlisten", is reduced to its most informative terms, which capture the excitement ("omg", "fire") and the context ("new song", "download", "mustlisten"). This transformation produces a concise representation that is easier for machine learning models to analyze.
Once the text is cleaned, TF-IDF (Term Frequency–Inverse Document Frequency) is applied. This method efficiently represents textual data by giving greater importance to unique words within a document while minimizing the influence of frequently occurring terms across the dataset. As a result, TF-IDF helps in highlighting meaningful features that improve the accuracy and efficiency of machine learning models.

4.3. Resampling (SMOTE): The next step in the framework involves addressing class imbalance in the target variable, which is a common challenge in text classification

tasks. Since the Synthetic Minority Over-sampling Technique (SMOTE) operates only on numerical values, the text data is first converted into numerical vectors before applying the method. SMOTE works by generating synthetic examples of the minority class, thereby balancing the distribution of classes within the dataset.

In this framework, the class imbalance was mitigated using the SMOTE implementation provided by the *imblearn* library. To handle severe imbalance, a dual strategy was adopted: synthetic minority samples were generated through oversampling, while random undersampling was simultaneously applied to the majority class. In addition, class weights were adjusted during the model training phase to ensure that greater emphasis was placed on correctly classifying the minority class. This combined approach improves the robustness of the model and enhances its ability to generalize across imbalanced datasets.

4.4. Hybrid Model (Ensemble): Rather than relying on a single classifier, this framework employs a hybrid ensemble model that integrates Logistic Regression, Random Forest, and XGBoost to achieve higher robustness and accuracy. The weights assigned to each classifier were determined through empirical testing to optimize performance.

Each algorithm contributes distinct strengths to the ensemble. Logistic Regression is effective in capturing linear relationships within text data, while Random Forest manages non-linearity and reduces the risk of overfitting. XGBoost further enhances performance through fine-tuning and its ability to handle complex decision boundaries. By combining these models, the ensemble leverages their complementary advantages: Logistic Regression provides strong baseline linear classification, Random Forest improves generalization and noise handling, and XGBoost fine-tunes predictive accuracy. This ensemble approach ensures improved generalization, reduced bias and variance, and ultimately delivers more reliable classification outcomes compared to individual models.

4.5. Evaluation & Visualization: The performance of the proposed model was evaluated using multiple metrics to provide a comprehensive understanding of its effectiveness, particularly in the presence of class imbalance. Accuracy, precision, recall, F1-score, ROC AUC, and PR AUC were employed, as together they offer a holistic view of prediction quality. Accuracy measures overall correctness, while precision and recall assess reliability and sensitivity. The F1-score balances precision and recall, providing a single measure of performance in imbalanced settings. ROC and PR AUC scores further illustrate the trade-offs between sensitivity and specificity.

To complement these metrics, visualization techniques such as the confusion matrix, ROC curve, and heatmaps were used to present class-wise performance and highlight areas where the model may misclassify. All visualizations were generated using *matplotlib* and enhanced with the high-level interface of *seaborn* to ensure clarity and readability.

In addition, to preserve the trained models beyond the active session, they were serialized to disk using *joblib*, which enables efficient saving and loading of Python objects containing large NumPy arrays. The overarching goal of these evaluation and

visualization methods is not only to measure accuracy but also to gain insights into why and how the model makes errors, which is essential for guiding improvements and refining predictive performance.

Evaluation Metrics for Classification Models

A. The Foundation: The Confusion Matrix:

The confusion matrix serves as the foundation for all evaluation metrics, as it provides the raw numbers required to compute performance measures. It is a tabular representation that categorizes predictions into four outcomes: true positive (TP), false positive (FP), false negative (FN), and true negative (TN). In the context of binary classification (for example, distinguishing between *fake news* and *real news*), the positive class represents the category of primary interest (*fake news*), while the negative class represents the other category (*real news*).

- **True Positive (TP):** Correctly identified positive instances (e.g., detecting fake news correctly).
- **False Positive (FP):** Incorrectly predicted positive cases (e.g., labeling real news as fake) – also known as Type I Error.
- **False Negative (FN):** Missed positive cases (e.g., failing to detect fake news) – also known as Type II Error.
- **True Negative (TN):** Correctly identified negative instances (e.g., correctly recognizing real news).

This matrix can also be visualized as a **heatmap**, where the intensity of color highlights the magnitude of each category, enabling quick interpretation of model performance at a glance.

B. Core Metrics Derived from the Confusion Matrix:

From the confusion matrix, several fundamental evaluation metrics are derived:

- **Accuracy** measures the proportion of correct predictions out of the total predictions. It is most appropriate when classes are balanced, and the cost of false positives and false negatives is roughly equal.

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

- **Precision** measures the reliability of positive predictions, indicating how often a predicted positive is actually correct. It is particularly important when minimizing false alarms (FPs) is critical.

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

- **Recall (Sensitivity or True Positive Rate – TPR)** measures the proportion of actual positives that were correctly identified. It is crucial when minimizing missed detections (FNs) is more important.

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

- **F1-Score** is the harmonic mean of precision and recall, providing a balanced metric that is especially useful for imbalanced datasets where neither precision nor recall alone is sufficient.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

C. Advanced Threshold-Based Metrics:

Beyond fixed-threshold metrics, advanced measures such as the ROC and PR curves provide insights into model performance across varying classification thresholds.

- **ROC Curve and AUC (Receiver Operating Characteristic):** The ROC curve plots the true positive rate (recall) against the false positive rate (FPR), defined as:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

  The ROC AUC score represents the area under the ROC curve, quantifying the model's ability to distinguish between classes. An AUC of 1.0 indicates a perfect classifier, while 0.5 indicates performance equivalent to random guessing. ROC AUC is best suited for balanced datasets or when overall ranking ability is the key concern.
- **PR Curve and AUC (Precision-Recall):** The PR curve plots precision against recall across different thresholds. The PR AUC score summarizes this trade-off, focusing specifically on the positive class. PR AUC is particularly informative for imbalanced datasets, where precision and recall provide more meaningful insights than ROC AUC.

D. Choosing the Right Metric:

The selection of evaluation metrics depends on the application context and the relative cost of errors. Accuracy is appropriate when classes are balanced, precision is prioritized when false positives carry high costs, and recall is emphasized when missing positive cases is unacceptable. The F1-score is valuable when a single metric is needed for imbalanced data, while ROC AUC provides a general measure of class separability. In contrast, PR AUC offers a clearer picture of performance when the dataset is highly imbalanced, as it focuses on the positive class of interest.

5. APPLICATIONS

The proposed framework has wide-ranging applications in both practical and research contexts, particularly in the field of digital forensics and the broader domain of misinformation detection. Within digital forensics, the framework can be applied to multiple branches such as disk, network, wireless, database, malware, email, and memory forensics. For instance, disk forensics benefits from retrieving misleading or manipulated content stored on digital media, while network and wireless forensics use the framework to trace the propagation of misinformation across communication channels. Similarly, database and email forensics can incorporate the model to detect tampered records or falsified communications, and memory forensics enables the recovery of processes in volatile memory associated with disinformation campaigns. In each of these areas, the model adds an intelligent layer of automated analysis that complements the manual collection and documentation of forensic evidence.

From a practical standpoint, the framework directly addresses the challenges of detecting fake news on digital platforms, thereby contributing to improving public trust, safeguarding political landscapes, and preserving democratic values. The hybrid ensemble model, by integrating TF-IDF bigram features, SMOTE resampling, and a combination of Logistic Regression, Random Forest, and XGBoost classifiers, ensures higher accuracy, reliability, and robustness in content classification. This makes it suitable for deployment in real-world systems such as social media monitoring platforms, online news portals, and governmental or regulatory agencies tasked with filtering misleading content. By reducing misclassification and improving detection rates, the framework provides a scalable solution for mitigating the influence of misinformation on society.

In terms of research applications, the model extends the body of knowledge in text classification, machine learning, and digital forensics by proposing a hybrid approach that overcomes limitations of single classifiers. It contributes to the exploration of techniques that combine linguistic feature engineering with ensemble learning to enhance classification accuracy in imbalanced datasets. Furthermore, the evaluation metrics and visualization methods integrated into the framework provide deeper insights into model behavior, offering future researchers a replicable and extendable methodology for misinformation detection. Beyond fake news, the same principles can be adapted to study other forms of deceptive or harmful content such as phishing emails, fraudulent reviews, or extremist propaganda.

Overall, the framework demonstrates value not only in practical implementations—where it can serve as a reliable detection mechanism within digital forensic investigations and online monitoring systems—but also in academic and research contexts, where it advances the understanding of hybrid machine learning models in addressing one of the most pressing issues of the digital age: the detection and prevention of misinformation.

6. MITIGATION STRATEGIES

Addressing misinformation is increasingly recognized as a multi-stakeholder responsibility, requiring the active involvement of individuals, governments, digital platforms, civil society, and the academic research community. As highlighted in global policy frameworks and digital safety reports, coordinated strategies are essential to ensure transparency, accountability, and resilience against deceptive content. The following mitigation strategies represent recommendations commonly endorsed across international policy and research domains.

6.1. Encourage Transparency:

- Indicate whether a piece of material is sponsored, paid, political, or AI-generated.
- Give credit where credit is due for political advertisements and campaigns.
- Notify users when significant changes to the algorithm impact their view.

6.2. Confirm the authenticity of the content & accounts:

- Expect stronger verification for political or high-reach accounts.
- Find and punish sockpuppet accounts, bots, trolls, and coordinated fraudulent activity.
- Employ evident tags to identify modified or altered media.

6.3. Boost Platform Accountability & Content Moderation:

- Prior to it becoming widely disseminated, identify and label misinformation.
- Don't overreact to clickbait, sensational, or unreliable content.
- Apply restrictions or fines to platforms and advertisers who violate the rules.

6.4. Incorporate Independent Fact-Checking:

- Collaborate with fact-checkers to validate posts that go viral.
- Employ visible labels such as “False,” “Misleading,” or “Disputed.”
- Encourage local and international fact-checking projects..

6.5. Encourage Digital & Media Literacy:

- Promote media literacy in public awareness efforts and school curricula.
- People should be taught to notice deepfakes, identify emotional manipulation, and cross-check facts.
- Encourage users to act responsibly by pausing before sharing and posting with awareness.

6.6. Ensure Dependable & Timely Corrections:

- Make sure that as soon as lies are discovered, they are promptly retracted and corrected.
- While promoting reliable sources, decrease the algorithmic reach of proven frauds.

6.7. Empower Civil Society, Journalists & Whistleblowers:

- Invest in watchdog projects and investigative journalism.
- Journalists and whistleblowers who reveal deceptive networks should be protected.
- Promote models for community-based corrections, such as Community Notes.

6.8. Advanced AI-Based Detection & Forensic Tools;

- Develop source tracking, watermarking, and deepfake detection.
- To ensure reliable media, use digital signatures and hashing.
- Update tools frequently to prevent changing manipulation strategies.

6.9. Encourage Ethical Platform & UX Design:

- Discourage clickbait-driven sharing (e.g., limit “share without reading”).
- Create user interfaces that encourage thoughtful interaction from users.
- Create barriers to help unconfirmed viral information move more slowly.

6.10. Develop User Reporting Systems Effective:

- Make it easy and transparent to report posts that are hurtful or misleading.
- Employ real-time reaction teams and community-based review systems.

6.11. Boost Regulatory and Legal Actions:

- Protect free expression while defining and punishing harmful deception tactics.
- Require political content that is paid, state-sponsored, or produced by AI to be disclosed.
- Create independent authorities to keep an eye on platform compliance.

6.12. Encourage Collaboration in Academic and Research:

- Collaborate with academic institutions and research centres to investigate patterns in deception.
- Improve detection algorithms and test solutions.
- Distribute categorized data for separate study.

6.13. Encourage global collaboration;

- Address state-sponsored misinformation by coordinating internationally.
- Create global standards for content authenticity, such as C2PA.
- Set standards for platform responsibility and election integrity.

6.14. Monitor Emerging Threats:

- Observe new strategies such as coordinated harassment, AR/VR deception, and artificial voices.
- In real time, modify mitigating techniques.

6.15. Promote Accountability & Public Reporting;

- Regularly publish transparency reports on the effectiveness of moderation.
- Enforce advertising responsibility to stop misinformation from being funded.
- Track takedowns and the results of policy enforcement in public.

6.16. Encourage Prebunking & Cognitive Resilience;

- Employ prebunking techniques, which involve educating people about manipulation techniques before to exposure.
- Develop psychological resistance to persuasive tactics such as conspiracy framing, anxiety, and anger.

The strategies articulated above form a multi-layered framework for combating deceptive content by integrating technological tools, societal initiatives, and regulatory interventions. However, beyond large-scale measures, individual users play a crucial role as the first line of defense against misinformation. By remaining alert and practicing simple habits, users can significantly reduce the risk of being misled or contributing to the spread of false information. A practical approach involves a few key steps: pause before sharing content, verify the source and author, cross-check headlines with at least two reliable outlets, look for signs of manipulation such as exaggerated language or altered media, and prioritize trusted and fact-checked platforms over unverified sources. Building these habits develops digital resilience, enabling users not only to protect themselves but also to strengthen the collective fight against misinformation. In this way, the broader framework aligns large-scale detection mechanisms and regulatory safeguards with individual responsibility, ensuring a more sustainable and effective defense against deceptive content.

7. CHALLENGES & LIMITATIONS

Digital forensics and fake news detection face several challenges and limitations due to the rapid growth of digital technologies and the widespread use of the internet. The proliferation of personal computers, mobile devices, and online platforms has significantly increased the volume and complexity of digital evidence. At the same time, hacking tools and malicious software are now widely accessible, which complicates investigations and often leaves insufficient or inconclusive proof to support legal prosecution. Furthermore, the sheer scale of digital storage, measured in terabytes or even petabytes, creates difficulties in retrieving, analyzing, and preserving evidence in a timely and reliable manner.

Technological advancements also pose a double-edged challenge: while they enable new detection methods, they simultaneously require forensic tools and fake news detection systems to undergo constant updates. Traditional methods of detection often struggle with issues of scalability and accuracy when applied to large and dynamic datasets. This

is particularly evident in the context of misinformation, where fake content spreads rapidly across digital platforms, undermining the credibility of reputable media sources and exacerbating risks such as cyberbullying, identity theft, and coordinated disinformation campaigns.

From a computational perspective, the high dimensionality of features generated through text vectorization methods such as TF-IDF significantly increases memory usage and processing time. However, this added complexity does not always translate into improved accuracy, leading to inefficiencies in model performance. Moreover, ensuring reliability and robustness in fake news detection remains a persistent challenge, as models must contend with noisy data, subtle linguistic manipulation, and constantly evolving disinformation tactics.

To address these limitations, hybrid approaches have been adopted that combine multiple models, each contributing unique strengths. For example, linear models are effective at handling structured relationships, while ensemble methods are more resilient to noise and non-linearity. Although this improves performance, challenges remain in balancing computational efficiency, interpretability, and adaptability in real-world applications.

8. CONCLUSIONS & FUTURE SCOPE

This paper examines the impact of misinformation on financial markets, including manipulation and celebrity defamation, highlighting the urgent need for effective fake content detection systems. Digital forensics is emphasized as a critical tool for locating and documenting digital evidence for legal purposes, encompassing areas such as disk, network, and email forensics. The proliferation of social media has exacerbated societal issues, including cyberbullying, political extremism, and the rapid spread of false information. To address these challenges, a hybrid machine learning model combining Logistic Regression and Random Forest is proposed for enhanced classification accuracy, while XGBoost is employed to fine-tune the system, improving overall robustness and reliability. The study also acknowledges the challenges faced by digital forensics due to rapid technological advancements and increasing data storage complexity. The paper further discusses mitigation strategies, including the implementation of automated fake news detection systems, monitoring mechanisms, and awareness programs to reduce misinformation's reach and impact. Future research may focus on incorporating deep learning techniques, expanding datasets to include multimedia content, developing real-time detection systems for social media, and creating frameworks to mitigate the socio-economic and political impacts of misinformation.

ACKNOWLEDGMENT

This work is acknowledged under Integral University manuscript No: IU/R&D/2025-MCN0003932.

REFERENCES

[1] P. Narasimhan and N. Kala, "Emerging Trends in Digital Forensics: Investigating Cybercrime," IJSRCSEIT, vol. 11, no. 1, pp. 3645–3652, 2025.

[2] S. A. Solangi, "IoT Forensics: Addressing Challenges and Establishing a Framework for Exploring Digital Forensics," SJCMS, vol. 8, no. 1, 2025.
[3] N. Madhan, D. Rajan, M. Jain, "Text Analysis for Anomaly Detection and Fraud Mitigation in Social Media using R," Kurdish Studies, vol. 12, no. 1, pp. 3845–3856, 2024.
[4] I. Touahri and A. Mazroui, "Survey of Machine Learning Techniques for Arabic Fake News Detection," Springer, 2024.
[5] R. Kenny, B. Fischhoff, A. Davis, K. M. Carley, and C. Canfield, "Duped by Bots: Why Some are Better than Others at Detecting Fake Social Media Personas," 2024.
[6] P. Dhiman, A. Kaur, C. Iwendi, and S. K. Mohan, "A Scientometric Analysis of Deep Learning Approaches for Detecting Fake News," *Electronics*, 12 (4), pp. 1–31, 2023.
[7] S. Nalini, "Certain Investigations on Drug Recommendations Using Machine Learning Techniques," 2023, pp. 1–142.
[8] S. Geetha, "A Novel Approach for Behavior Prediction Using Sentiment Analysis on Social Media Data with Machine Learning Techniques," 2022, pp. 1–137.
[9] A. Doegar, "Development of an Approach for Image Forgery Detection Using Machine Learning Algorithms," 2022, pp. 1–189.
[10] M. R. Naeem, R. Amin, S. S. Alshamrani and A. Alshehri, "Digital Forensics for Malware Classification: An Approach for Binary Code to Pixel Vector Transition," Computational Intelligence and Neuroscience, 2022.
[11] A. A. Solanke and M. A. Biasiotti, "Digital Forensics AI: Evaluating, Standardizing and Optimizing Digital Evidence Mining Techniques," Künstliche Intelligenz, vol. 36, pp. 143–161, 2022.
[12] Monika and A. Passi, "Digital Image Forensic Based on Machine Learning Approach for Forgery Detection and Localization," ICMAI, IOP Publishing, 2021.
[13] Z. Shen, "Deep Learning on Image Forensics and Anti-Forensics," 2021, pp. 1– 110.
[14] H. Khan, S. Hanif and B. Muhammad, "A Survey of Machine Learning Applications in Digital Forensics," Peertechz, pp. 20–24, 2021.
[15] J. Pu, N. Mangaokar, L. Kelly, P. Bhattacharya, K. Sundaram, M. Javed, B. Wang and B. Viswanath, "Deepfake Videos in the Wild: Analysis and Detection," 2021.
[16] U. Rahul, M. Ragul, K. R. Vignesh and K. Tejeswinee, "Deepfake Video Forensics Based on Transfer Learning," Int. J. Recent Technology and Engineering, vol. 8, no. 6, pp. 5069–5073, 2020.
[17] R. Das, G. Negi and A. F. Smeaton, "Detecting Deepfake Videos Using Euler Video Magnification," 2020.
[18] L. Barik, "Data Mining Approach for Digital Forensics Task with Deep Learning Techniques," Int. J. Advanced and Applied Sciences, pp. 56–65, 2020.
[19] R. Aswani, "Bio-Inspired Computing for Outlier Detection: Select Studies in Web 3.0 Domain," 2019.
[20] R. J. Alharbi, "Improving the Classification of Tiny Images for Forensic Analysis," 2019.
[21] H. Alhakami, W. Alhakami, A. Baz, M. Faizan, M. W. Khan and A. Agrawal, "Evaluating Intelligent Methods for Detecting COVID-19 Fake News on Social Media Platforms," Electronics, vol. 11, no. 2417, 2022.
[22] N. Javed, T. Ahmed and M. Faisal, "A Comprehensive Study on Prevalence of Cyberbullying and Its Impact on Youths and Adults," J. Statistics & Management Systems, vol. 26, no. 7, pp. 1655–1672, 2023.

[23] A. A. Abdussami and M. F. Farooqui, "Optimal Feature Selection with Weight Optimised Deep Neural Network for Incremental Learning-Based Intrusion Detection in Fog Environment," J. Inf. & Knowl. Management, vol. 21, no. 3, 2250042, 2022.

[24] N. A. Farooqui, M. K. H. Mohammed, A. H. R. Noori, S. Islam, M. Haleem, S. F. Ahmad, A. Khan, F. R. Awad Ahmed, N. B. M. B. Babiker, T. E. Ahmed and A. U. R. Khan, "Hybrid Bat and Salp Swarm Algorithm for Feature Selection and Classification of Crisis-Related Tweets in Social Networks," IEEE Access, vol. 12, pp. 103908–103920, 2024.

[25] A. Malik, M. Kuribayashi, S. M. Abdullahi and A. N. Khan, "DeepFake Detection for Human Face Images and Videos: A Survey," IEEE Access, vol. 10, pp. 18757–18775, 2022.