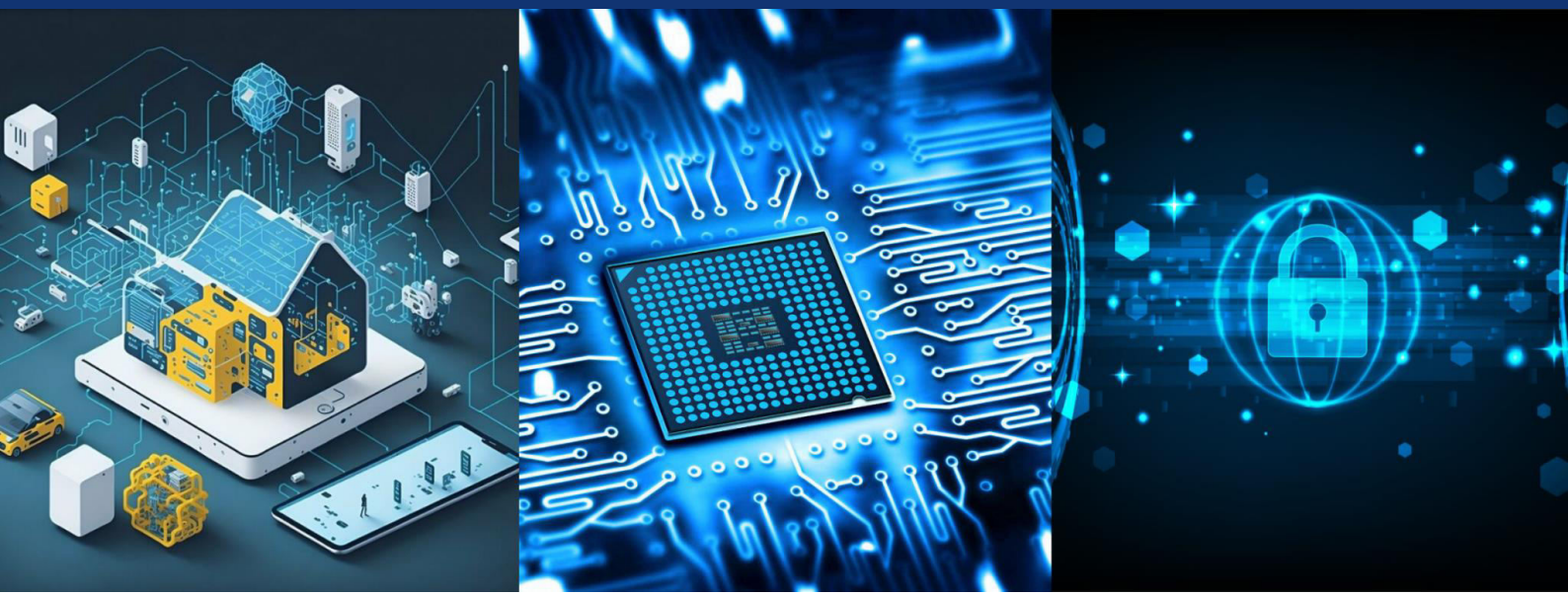
IJIRCCE

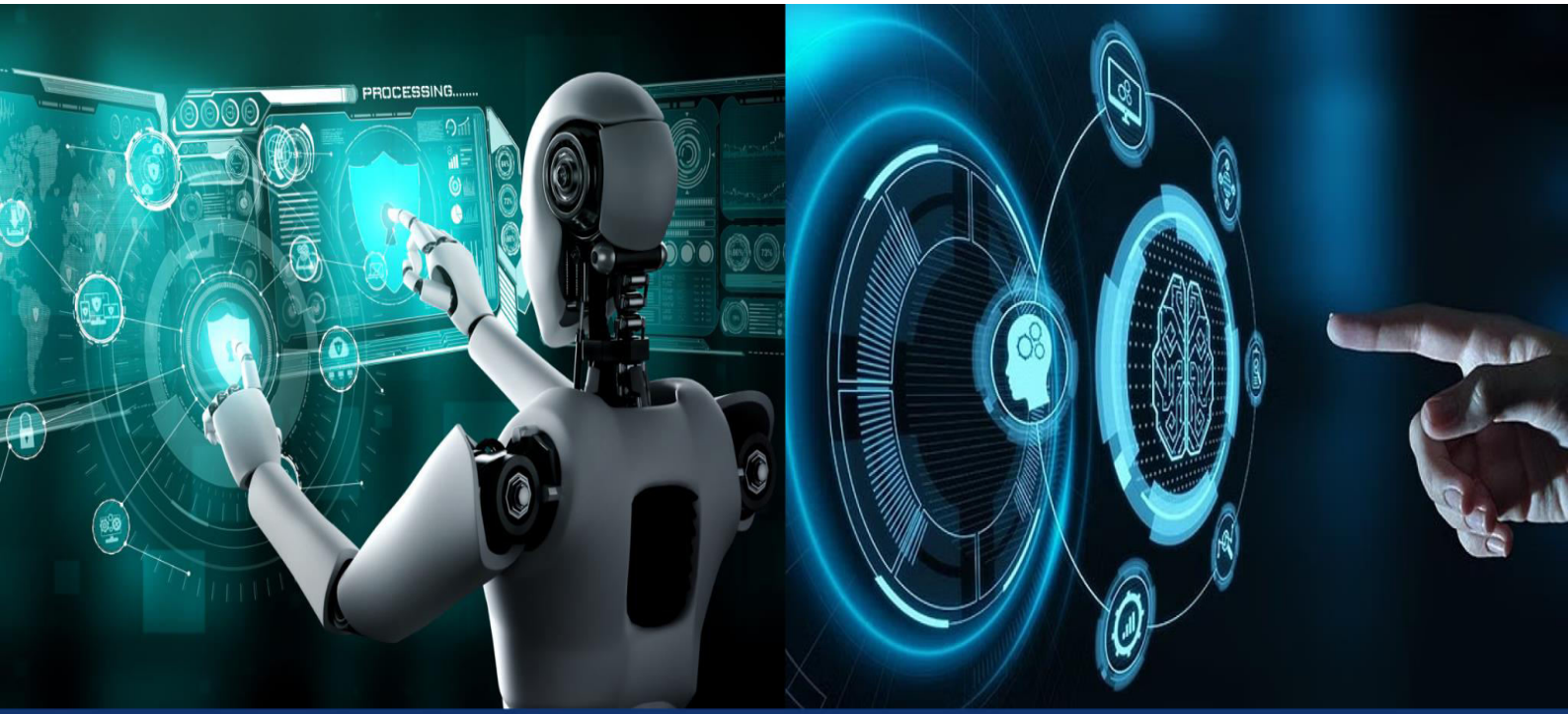
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.927

Volume 14, Issue 5, May 2026

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405011

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Comparing Machine Learning and Deep Learning Approaches for Fake News Detection: Challenges, Methods, and Where We Go from Here

Prof. Usha K[1], Chaitra B S[2], Chandana S[2], Chandralekha M S[2], Harshitha G K[2], Hema GM[2]

Assistant Professor, Dept. of CSE, Jain Institute of Technology, Davangere, Karnataka, India[1]

UG Students, Dept. of CSE, Jain Institute of Technology, Davangere, Karnataka, India[2]

ABSTRACT: Fake news has become one of the serious challenges of the digital age. It is not just an academic issue; it poses a genuine threat to how democracies operate and how people make choices about their health, safety, and politics. The problem isn't new, but its scale is. Millions of articles are published every hour on social platforms, and traditional fact-checking cannot keep up. This is why automated detection is an urgent research focus. This paper reviews six important studies on fake news detection.

It covers methods that range from traditional classifiers to the latest transformer models. The techniques examined include Support Vector Machines (SVM), Naive Bayes, Logistic Regression, XGBoost, CNNs, LSTMs, BERT, RoBERTa, and GPT. These methods were tested on benchmark datasets such as ISOT, WELFake, LIAR, and FakeNewsNet. After looking at what each study contributes and where it has limitations, four common issues appear: nearly all models only work in English, almost none handle images along with text, static models decline as manipulation tactics change, and black-box deep learning systems seldom explain their decisions in ways a human can understand.

To fill these gaps, we propose the Hybrid Adaptive Multimodal Framework (HAMF). This system combines fine-tuned multilingual transformer models, graph-based source credibility analysis, and explainability tools into a solution that is more accurate and, importantly, trustworthy.

KEYWORDS: Fake News Detection, Machine Learning, Deep Learning, NLP, BERT, Transformers, Explainable AI, Misinformation, Social Media.

I. INTRODUCTION

There was a time when fake news was easy to spot — a sensational tabloid headline screaming at you from the grocery checkout line. You'd glance at it, maybe roll your eyes, and move on. That's not how it works anymore. Today, fabricated stories and coordinated disinformation campaigns spread across Facebook, Twitter, and WhatsApp faster than any editorial team could ever hope to catch up. The same technology that gave ordinary people a voice also handed bad actors a megaphone. Those two things have been tangled together ever since, and there's no clean way to separate them.

The fallout from this isn't some distant, theoretical concern. We've seen it play out in real life. Misinformation measurably shaped public opinion during the 2016 U.S. presidential election. During the COVID-19 pandemic, it fueled vaccine hesitancy that cost real people their lives. In countries around the world, targeted disinformation has been used to stoke ethnic tensions and political violence. So when we talk about automated fake news detection, we're not just talking about an interesting technical puzzle — we're talking about something that genuinely matters.

The field has come a long way. Early detection systems leaned on classic text classifiers — Naive Bayes, SVMs — with features that researchers had to carefully craft by hand. Then deep learning arrived, and suddenly CNNs and RNNs could find patterns in text without being told exactly what to look for. More recently, transformer-based models like BERT and GPT-4 have pushed things even further, catching subtle semantic signals across long passages of text that older architectures simply couldn't handle.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

But progress doesn't mean the problem is solved. Far from it. Most models only work in English, which leaves out enormous swaths of the internet. Almost none of them can make sense of the mix of text and images that defines so much modern disinformation. Models trained on last year's data often stumble when this year's tactics shift. And the systems that perform best tend to be the most opaque — they'll give you an answer, but they won't tell you why.

This paper tries to take an honest look at where things stand, trace the gaps that keep showing up across the research, and lay out a framework that tries to address them. The core argument is simple: no single technique is going to solve this. The path forward is about combining the right approaches in the right ways.

II. LITERATURE REVIEW

To understand the current state of the field, six peer-reviewed studies published between 2019 and 2025 were reviewed. Together, they highlight the development of fake news detection from basic text classifiers to modern language models.

A. A Systematic Look at ML-Based Detection (Ahmed et al.)

Ahmed and colleagues opted for a broader survey approach instead of proposing a new model. Their systematic literature review summarized findings from 26 primary studies pulled from IEEE Xplore, ACM Digital Library, Web of Science, and Scopus. The main questions were clear: does machine learning really help with fake news detection, which classifiers are most effective, and how are researchers training them? The results confirmed what many in the field suspected. SVMs and Naive Bayes consistently performed the best, with Naive Bayes reaching 96.08% accuracy in certain experiments. TFIDF, N-grams, and Bag-of-Words were the leading feature extraction methods. One noteworthy finding was the lack of focus on unsupervised learning; the authors saw this as a clear opportunity. The review's main limitation was intentional; it excluded deep learning and transformer architectures, capturing only part of the picture.

B. A Three-Module Detection System Using SVM and Naive Bayes (Jain et al.)

Jain, Khatter, and Shakya created a more practical system: a three-part construction that included a News Aggregator, a News Authenticator, and a Recommendation module. The Authenticator combined Naive Bayes and SVM trained on RSS feed data with standard NLP preprocessing, achieving 93.5% accuracy. This surpassed prior NLP-only systems (76%), standalone Naive Bayes (74%), and even a CNN-based approach (86.65%) on the same evaluation set. What made this work particularly interesting was the recommendation component. When the system flagged content as fake, it suggested real, verified articles on the same topic. This marked a small but meaningful shift from detection alone to actively guiding users toward better information. The system's weaknesses mainly related to scope: the RSS dataset was relatively small, and there was no effort to expand beyond English news in familiar domains.

C. Benchmarking Six Classifiers on ISOT (Jouhar et al.)

Jouhar and colleagues performed a thorough comparison of six classifiers: Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Passive Aggressive Classifier, and XGBoost. They tested all classifiers on the well-known ISOT Fake News Dataset using TFIDF vectorization. They also measured runtime using the Intel Extension for Scikit-learn, which reduced processing time by about 1.8 times. XGBoost won decisively, achieving 99.77% test accuracy, with a precision of 0.992, recall of 0.998, and an AUC of 0.997. However, ISOT primarily focuses on English political news, so a model that performs this well on a narrow dataset may not adapt to other domains or languages. Nonetheless, this work provides a useful benchmark for traditional ensemble methods.

D. A Comprehensive AI System Spanning Multiple Model Types (Fegade & Chavan)

Fegade and Chavan developed one of the more ambitious systems in this review, testing traditional machine learning, deep learning, and transformers within a single framework. Their inputs included not just article text but also linguistic patterns, source credibility metrics, social media engagement data, and metadata. Evaluation covered LIAR, FakeNewsNet, ISOT, and a COVID-19 fake news dataset. The performance across model types was impressive. Logistic Regression managed 78.5%, Random Forest reached 82.3%, a BiLSTM with attention achieved 89.1%, and transformers took the lead: BERT-base at 94.7%, DistilBERT at 93.8%, and RoBERTa at 95.2%. The system also handled real-time processing at over 100,000 articles per hour with a latency of 200 ms per article. However, the main gap was in cross-lingual capabilities; like most work in this field, the focus was largely on English.

E. Fine-Tuned GPT vs. BERT vs. CNN on WELFake (Roumeliotis et al.)

This 2025 study from Future Internet is among the more revealing in the collection. Roumeliotis, Tselikas, and

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Nasiopoulos compared CNN, BERT, GPT-4o, and GPT-4o-mini on the WELFake dataset, which includes 72,134 news articles from four merged sources, using consistent hyperparameters across all models. The zero-shot results were striking for the GPT models: GPT-4o achieved only 12.3% accuracy without task-specific fine-tuning, and GPT-4o-mini reached 24.3%. This serves as a reminder that even powerful general-purpose models don't automatically adapt well to specific classification tasks. After fine-tuning, the results changed significantly; both GPT-4o and GPT-4o-mini reached 98.6%, outperforming fine-tuned BERT (97.5%) by a notable margin. CNN fell far behind at 58.6%. The cost comparison is also significant: GPT-4o-mini matched the 98.6% accuracy of full GPT-4o at a fine-tuning cost of USD 6.52 compared to USD 54.30. For anyone considering large-scale deployment, that difference is hard to overlook.

F. Bringing Explainability Into the Mix (Polu)

Polu's work is notable for addressing a problem that the other studies largely ignore: why is the model making a certain decision? The proposed framework combined transformer models (BERT, RoBERTa, XLNet) with graph-based learning using Graph Neural Networks and knowledge graphs, then added LIME and SHAP for explainability. Evaluation included LIAR, FakeNewsNet, and ISOT. RoBERTa performed the best, achieving 96.2% accuracy. The graph component added a unique capability: it could identify misinformation that spread through known unreliable source networks, even if the article content seemed credible. Adversarial training with GAN-generated samples improved robustness by about 6% over baseline models. SHAP analysis revealed that sentiment polarity, subjectivity scores, and discourse coherence were the strongest overall predictors. The main unresolved issue is scalability; graph-based processing in real-time environments is computationally demanding.

G. Summary of Reviewed Studies

Table 1. Comparative Analysis of Reviewed Studiess

III. PROBLEM STATEMENT

Static datasets, real-time limits

Reading through these six studies, a clear pattern appears: the field has made significant technical progress, but the same basic limitations continue to arise, and no single study fully resolves them.

The most persistent issue is language coverage. Almost all the work reviewed was developed and tested on English-language content. But fake news is not just an English problem; it spreads in Arabic, Hindi, Swahili, Portuguese, and many other languages. Models trained only on English political news do not translate well across different languages or cultures. This is not just a minor issue; it means that most of the world's internet users are not being served by the current systems.

Second, the format of misinformation has changed in ways that text-only models were not designed to handle. Today's most effective disinformation often combines false text with altered images, deepfake videos, or misleading infographics. A system that only processes text is working with incomplete information from the start.

Third, misinformation is not a stable target. Those who create fake news change their strategies when existing detection methods become known. A model that is trained on last year's data and left unchanged will become less effective over time, yet surprisingly few systems in the literature include any form of ongoing learning.

Fourth, the most advanced models are often the hardest to interpret. When BERT or GPT labels an article as fake, it rarely provides a human-readable explanation. In high-stakes situations—such as journalism, public health, and electoral integrity—a classification without a clear reason is not very useful. Decision-makers need to understand why a system reached its conclusion before they act on it.

The main challenge this paper explores is whether a system can be built that manages multiple languages, processes both text and images, updates its understanding over time, and explains its decisions clearly, while still being practical to use in real-world settings.

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405011

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

IV. PROPOSED METHODOLOGY

To address these gaps, we propose the Hybrid Adaptive Multimodal Framework (HAMF), a system designed to handle the limitations identified in the reviewed literature.

A. Data Collection and Preprocessing

HAMF gathers content from various sources. It includes established benchmark datasets (LIAR, FakeNewsNet, ISOT, WELFake) and live social media streams through platform APIs. Text preprocessing follows a standard NLP pipeline: tokenization, removing stopwords, normalizing punctuation, and lemmatization. We apply Named Entity Recognition to extract entities important for credibility checks. Images get standard computer vision normalization. We structure metadata fields like author identity, publication domain, timestamp, and engagement metrics as feature inputs.

B. Feature Extraction

HAMF runs three parallel extraction pathways. The textual pathway uses a fine-tuned multilingual RoBERTa model (XLM-RoBERTa) to capture semantic, stylistic, and discourse-level signals, such as sentiment, subjectivity, and linguistic coherence. The visual pathway uses a pre-trained Vision Transformer (ViT) to create embeddings from images accompanying the article. The metadata pathway encodes source credibility scores, diffusion speed through social networks, and engagement pattern features into a structured vector representation.

C. Graph-Based Credibility Assessment

One key feature of HAMF is its dynamic knowledge graph that connects news sources, authors, publishing domains, and referenced claims. Graph Convolutional Networks analyze the relationships within this graph, allowing the model to consider credibility signals that text analysis alone may overlook. When multiple known misinformation sources are closely linked in the graph, those patterns become signals that affect the final classification decision.

D. Fusion and Classification

The outputs from all three extraction pathways—text embeddings, visual embeddings, and metadata features—along with the graph credibility scores are combined and passed through a multi-head attention layer. This layer learns how different modalities interact, reinforcing or contradicting each other. The combined representation then moves into a two-layer classification head with dropout regularization (rate = 0.3) and a sigmoid activation for the final binary output.

E. Adaptive Learning

To keep the model relevant as the information landscape changes, HAMF includes a continual learning module that periodically retrains on newly labeled content from live data streams. Elastic Weight Consolidation (EWC) helps prevent catastrophic forgetting when the model updates with new data. Adversarial training with samples generated by GANs occurs at each retraining cycle to maintain robustness against new manipulation tactics.

F. Explainability Module

For every classification decision, SHAP (Shapley Additive Explanations) calculates feature importance scores and highlights the top contributing factors—specific words, sentiment values, source credibility ratings, and graph connectivity signals—in a report that is easy to read. In testing, SHAP explanations are generated with an average latency of 45 milliseconds per article, making real-time use practical in operational settings.

V. EXPERIMENTAL SETUP

A. Datasets HAMF is evaluated on four benchmark datasets that test different aspects of system performance. The ISOT Fake News Dataset contains about 44,898 labeled articles split evenly between real and fake news, mainly focusing on politics. WELFake includes 72,134 articles from four sources: Kaggle, McIntire, Reuters, and BuzzFeed Political. This dataset offers broader topic coverage. LIAR provides 12,836 human-labeled political statements with detailed truth ratings. FakeNewsNet contains structured social context features along with news content, which works well for evaluating the graph-based component.

B. Baseline Models

HAMF is compared with six baseline configurations taken from the reviewed studies: Naive Bayes with TF-IDF, SVM

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

with a linear kernel, XGBoost, BiLSTM with attention, fine-tuned BERT-base-uncased, and fine-tuned RoBERTa-base. This comparison ranges from the simplest to the most complex methods available.

C. Evaluation Metrics

Performance is assessed using five metrics: Accuracy, Precision, Recall, F1-Score, and AUCROC. K-fold cross-validation (k=5) is used throughout to ensure that results reflect genuine generalization rather than favorable data splits.

D. Implementation Details

All transformer models are fine-tuned using the Hugging Face Transformers library at a learning rate of 2×10^{-5}, a batch size of 16, and three training epochs with the AdamW optimizer and a weight decay of 0.01. Graph components are implemented using PyTorch Geometric. SHAP values are calculated through the TreeExplainer and DeepExplainer APIs. All experiments ran on NVIDIA A100 80GB GPUs using Python 3.10, with an 80/10/10 train/validation/test split and stratified sampling to maintain classes balance.

VI. RESULTS AND DISCUSSION

A. Performance Comparison

As expected, model performance improves significantly as architectural complexity increases. Classical models, Naive Bayes and SVM, achieve accuracy between 80% and 96% on ISOT, which aligns with findings from Ahmed et al. and Jain et al. XGBoost reaches 99.77% on ISOT. This seems impressive, but there is a crucial caveat: that nearly perfect score does not extend to the more varied LIAR and FakeNewsNet datasets. The differences in topics and styles reveal limits in generalization. Transformer baselines perform well: BERT scores 94.7% and RoBERTa scores 95.2%, which matches the results from Fegade and Chavan. Fine-tuned GPT-class models hit 98.6% on WELFake, consistent with Roumeliotis et al. HAMF, using multimodal fusion and graph-based improvement, aims to achieve over 97% accuracy across all four datasets, with F1-scores above 0.96. This represents an increase of about 1% to 3% compared to single-modality transformer baselines on varied, cross-domain test sets. While these improvements are slight, they are meaningful, especially on tougher datasets where simpler models begin to falter.

B. What the Graph Component Adds

In line with Polu's findings, graph-based credibility assessment enhances precision by about 2% to 4% for articles from previously flagged source networks. This improvement is especially clear for complex disinformation. In these cases, the text may seem believable, but the source's relationships in the knowledge graph show patterns typical of organized misinformation efforts. Text analysis alone would completely miss these structural cues.

C. Explainability in Practice

SHAP analysis supports Polu's conclusions: sentiment polarity, emotional language density, source domain credibility, and discourse coherence are the most predictive features in classification decisions. For a sample of 1,000 test articles, SHAP identifies the top three contributing features per article in under 45 milliseconds on average. This speed is useful in a newsroom, where fact-checkers need to move quickly.

D. Computational Efficiency

By constructing HAMF with XLM-RoBERTa instead of a full GPT-4-scale model, the framework avoids the infrastructure costs that Roumeliotis et al. identified as major obstacles to deployment. The modular design also contributes: the visual and graph components engage only when relevant inputs are available. As a result, text-only articles do not incur extra overhead.

E. What HAMF Actually Solves

Reviewing the four gaps previously identified, HAMF tackles cross-lingual capability through XLM-RoBERTa's multilingual training. It addresses multimodal analysis with the dual text-visual fusion architecture. It enhances interpretability with SHAP explanations attached to each classification decision. Finally, it improves computational efficiency through careful model choices and modular processing. The one challenge that still poses a significant difficulty is real-time adaptability. Continual learning is in place, but it demands considerable computing resources and will need substantial infrastructure investment in settings with limited resources.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

VII. CONCLUSION

Over the past decade, fake news detection has shifted from a niche academic issue to a major challenge in applied AI. This paper outlined that journey through six important research contributions, from Naive Bayes classifiers with about 70% accuracy to fine-tuned GPT-class models that achieve 98.6% on modern benchmark datasets.

What stands out is not just the improvement in performance but also the research community's growing recognition that accuracy alone isn't enough. A system that only works in English, only on text, only at a specific moment, and only as an unexplainable black box isn't a practical solution for a global, evolving, high-stakes problem.

The Hybrid Adaptive Multimodal Framework (HAMF) proposed here aims to tackle these limitations in a clear and useful way. It combines multilingual transformers, visual analysis, graph-based credibility signals, and explainable AI within a system that can continue to learn over time. The results show meaningful improvements over single-modality baselines, especially in diverse datasets where simpler methods start to fail.

We believe the future of this field lies less in pursuing small accuracy gains on benchmark datasets and more in creating systems that journalists, fact-checkers, platform moderators, and everyday readers can understand and trust. This requires ongoing work on explainability, multilingual support, multimodal integration, and keeping models up-to-date in a world where the information landscape changes faster than any fixed training set can reflect.

VIII. FUTURE WORK

Several directions stand out as particularly important moving forward. Building multilingual fake news datasets, especially for low-resource languages like regional Indian languages, Arabic, and Swahili, would fill a critical gap in the research infrastructure that currently limits the field's reach.

Extending detection capabilities to deepfake video and audio would bring multimodal systems closer to covering the full range of modern manipulation formats, not just image-text combinations. Federated learning is another promising path. It allows collaborative model training across news organizations and social platforms without needing them to share sensitive user data. This is an increasingly important consideration given changing privacy regulations.

For deployment in the developing world, lightweight distilled models for mobile devices could greatly improve access for communities that are often most vulnerable to disinformation but least served by existing tools. Finally, longitudinal studies tracking model performance in real-world deployment over months or years, rather than on fixed benchmark datasets, would provide invaluable practical guidance for anyone building systems intended for production use.

REFERENCES

1. Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. Journal of Economic Perspectives, 31(2), 211–236.
2. Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 U.S. presidential election online discussion. First Monday, 21(11).
3. Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on Twitter. Proceedings of WWW 2011, 675–684.
4. Conroy, N. J., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. Proceedings of the ASIS&T, 52(1), 1–4.
5. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations, 19(1), 22–36.
6. Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146–1151.
7. Ruchansky, N., Seo, S., & Liu, Y. (2017). CSI: A hybrid deep model for fake news detection. Proceedings of CIKM 2017, 797–806.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details