© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1715927

# Fake News AI-Based Detection System

DR. AMANDEEP SINGH ARORA[1], SWATI GUPTA[2], DR. SHALU TANDON[3], DR. VIKAS RAO VADI[4], CHARANPREET KAUR[5], DR. ASJAD USMANI[6]

[1, 5, 6]*Associate Professor, Don Bosco Institute of Technology, Delhi*
[2]*Assistant Professor, Don Bosco Institute of Technology, Delhi*
[3, 4]*Professor, Don Bosco Institute of Technology, Delhi*

***Abstract-*** ***It is the time when a single false story can travel the globe in minutes, shaping opinions before anyone has a chance to verify it. Fake news is no longer just a nuisance — it actively destabilises democracies, fuels communal tensions, and erodes public trust in institutions. Manual fact-checking, as noble an effort as it is, simply cannot keep up with the flood of content produced every day on social media. This paper explores how Artificial Intelligence can step in to help. We built a Fake News Detection System using Natural Language Processing (NLP) and Machine Learning (ML) that reads a news article and decides whether it is real or fabricated. The system uses TF-IDF vectorization to convert text into meaningful numerical features, and a Passive Aggressive Classifier to make the final call. Trained on roughly 20,000 news records from Kaggle, the system reached 92% accuracy, 90% precision, and 91% recall — results that genuinely surprised us with how well a lightweight approach could perform. Beyond the numbers, this paper also digs into what existing research has missed and where future systems need to go.***

***Keywords: Fake News Detection, Artificial Intelligence, Natural Language Processing, Machine Learning, TF-IDF, Misinformation.***

## I. INTRODUCTION

Think about the last time you saw a shocking headline on WhatsApp or Instagram. Did you stop to check whether it was true? Most of us don't — and that is precisely the problem. Fake news spreads because it is designed to trigger emotional reactions before our rational minds kick in. The Reuters Institute Digital News Report (2023) found that more than half of people worldwide — 56% to be precise — are genuinely unsure of their ability to tell real news from fake. That number should alarm all of us.

India offers some of the most sobering examples of what happens when misinformation goes unchecked. Between 2018 and 2022, false rumours circulating on WhatsApp were directly linked to mob lynchings in several states. During the COVID-19 pandemic, the World Health Organization had to fight not just the virus but what it called an 'infodemic' — a parallel epidemic of health misinformation that convinced people to avoid vaccines and try dangerous home remedies [7].

So what can technology do about this? Quite a lot, it turns out. AI systems built with NLP and ML techniques can scan thousands of articles per second, picking up on the subtle language patterns that distinguish fabricated stories from genuine ones [1]. They are not perfect, and they are certainly not a complete solution on their own — but they can act as a powerful first line of defence, flagging suspicious content before it goes viral. That is the goal of the system we present in this paper.

## II. LITERATURE REVIEW

Researchers have been working on this problem for over a decade, and the journey from simple rule-based filters to sophisticated deep learning models is a fascinating one.

1. Castillo et al. (2011) were among the first to treat news credibility as a computational problem. Working with Twitter data, they built a decision-tree model that looked at metadata — how many times something was retweeted, what kind of account posted it — and managed about 86% accuracy. It was groundbreaking at the time, though it could only work on short social posts, not full news articles.

2. Shu et al. (2017) took a step back and wrote the definitive survey that gave the field a common language. They grouped detection approaches into four families: knowledge-based, style-based, propagation-based, and source-based. Their key insight was that style-based methods — ones that only look at how something is written — are the

most practical, because they don't need access to a user's social network or posting history.

3. Ruchansky et al. (2017) built something more ambitious: the CSI model, which combined the article's text, how users responded to it, and the credibility of the source, all fed into a deep learning framework. It worked remarkably well, but it needed data from the social platform itself — meaning you couldn't use it if the platform didn't share that data.

4. Wang (2017) introduced the LIAR dataset — 12,800 short political statements each manually labelled with a credibility rating. He also showed that knowing who said something matters just as much as what they said. A known spreader of misinformation saying something increases the probability it is false, even before you analyse the words.

5. Pérez-Rosas et al. (2018) proved that you don't always need deep learning to do well. Using n-gram features alongside measures like sentence readability and sentiment, their ML model performed competitively — and was far easier to train and deploy.

6. Kaliyar et al. (2021) pushed the accuracy ceiling with FakeBERT, which layered multiple CNN blocks on top of BERT's powerful language representations. The results were excellent, but the hardware requirements put it out of reach for most real-world applications.

7. Zhou and Zafarani (2020) wrapped up a decade of research in a comprehensive survey and, crucially, pointed to what the field still gets wrong: most systems are brittle against adversarial content, almost none work in languages other than English, and very few can explain their decisions in a way humans can understand.

Table 1: Review of Literature

| Year | Paper / Study | Authors | Approach / Model | Key Contribution | Strengths | Limitations |
|---|---|---|---|---|---|---|
| 2011 | Twitter News Credibility | Castillo et al. | Decision Tree (metadata-based) | Pioneered computational modeling of news credibility using social media features | Early breakthrough; achieved ~86% accuracy | Limited to short social media posts; not applicable to long-form content |
| 2017 | Fake News Detection: A Survey | Shu et al. | Survey / Taxonomy | Proposed four categories: knowledge-, style-, propagation-, and source-based | Established foundational framework | Lacks experimental validation |
| 2017 | CSI: A Hybrid Deep Model for Fake News | Ruchansky et al. | Deep learning (content + user + source) | Integrated textual, social, and source features into a unified model | High accuracy; comprehensive modeling | Requires access to social platform data |
| 2017 | LIAR Dataset | Wang | Dataset + Machine Learning | Introduced benchmark dataset with labeled political statements | Standardized evaluation; widely adopted | Limited to short statements; lacks contextual depth |

| Year | Paper / Study | Authors | Approach / Model | Key Contribution | Strengths | Limitations |
|---|---|---|---|---|---|---|
| 2018 | Fake News Detection using Linguistic Features | Pérez-Rosas et al. | Traditional ML (n-grams, readability, sentiment) | Demonstrated effectiveness of linguistic feature-based models | Interpretable; computationally efficient | Lower scalability compared to deep learning models |
| 2020 | Fake News Detection: A Survey | Zhou & Zafarani | Survey | Highlighted challenges: adversarial robustness, explainability, multilingual gaps | Comprehensive and critical analysis | No proposed model or empirical results |
| 2021 | FakeBERT | Kaliyar et al. | BERT + CNN hybrid | Enhanced BERT with CNN layers for improved feature extraction | State-of-the-art performance | High computational cost; limited real-time applicability |
| 2023 | Fake News Detection using LLMs | OpenAI / Google / Meta (various) | Large Language Models (GPT, PaLM, LLaMA) | Applied large pre-trained LLMs for zero-/few-shot fake news detection | High accuracy; minimal task-specific training | Expensive; lacks explainability; risk of hallucination |
| 2023 | DeBERTa-based Fake News Detection | He et al. (and others) | Transformer (DeBERTa variants) | Improved contextual understanding using disentangled attention | Better contextual representation | High resource consumption |
| 2024 | Multimodal Fake News Detection | Various | Multimodal (text + image + video) | Combines textual and visual signals to detect misinformation | More robust to real-world content | Requires large multimodal datasets |
| 2024 | Explainable Fake News Detection Models | Various | XAI + Transformer models | Focus on interpretability and explainable predictions | Improved trust and transparency | Trade-off between accuracy and explainability |

## III. METHODOLOGY

Our approach was deliberately kept simple and practical. We wanted to build something that actually works on a normal laptop, without a GPU, and without needing access to any social media API. Here is how we did it:

1. Data Collection: We used the Kaggle Fake News Dataset — around 20,000 news articles evenly split between FAKE and REAL labels, covering topics from politics and health to sports and entertainment. Each article comes with a title and full body text.

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1715927

2. Text Preprocessing: Raw news text is messy. We cleaned it up by converting everything to lowercase, stripping out URLs, HTML tags, and special characters, then tokenizing the text using NLTK. We removed common stopwords (words like 'the', 'is', 'and' that don't carry meaning) and applied Porter Stemming to reduce words to their root forms — so 'running', 'ran', and 'runs' all become 'run'.

3. Feature Extraction: We used TF-IDF (Term Frequency-Inverse Document Frequency) to turn the cleaned text into numbers the model can work with. TF-IDF is clever because it gives high scores to words that are distinctive to a particular article while downweighting words that appear everywhere. We capped our vocabulary at 5,000 terms to keep things manageable.

4. Model Training: We split the data 80:20 for training and testing, making sure both splits had roughly equal proportions of fake and real articles. Our main model was the Passive Aggressive Classifier (PAC) — an online learning algorithm that is unusually well-suited to text classification because it updates aggressively when it makes a mistake and stays put when it gets things right. We also trained Logistic Regression and Linear SVM for comparison, all using Scikit-learn.

5. Prediction: Once trained, the model takes any news article, runs it through the same preprocessing and TF-IDF pipeline, and returns a verdict: FAKE or REAL, along with a confidence score.

## IV. RESULTS

We evaluated the system on the held-out 20% test set using four standard metrics: accuracy, precision, recall, and F1-score. The Passive Aggressive Classifier delivered the following:

1. Accuracy: 92%

2. Precision: 90%

3. Recall: 91%

4. F1-Score: 90.5%

To put these numbers in context: Logistic Regression hit 88% and Linear SVM reached 91%, both respectable, but PAC edged ahead while also being the fastest to train. What we found most encouraging was the 91% recall — meaning that out of every 100 fake articles, the system correctly caught 91 of them. In a real-world setting, missing fake news (a false negative) is a much costlier error than incorrectly flagging a real article (a false positive), so a high recall matters more than a high precision. The cases the model struggled with were mostly satirical pieces and strongly worded opinion articles — content that deliberately sounds like factual reporting but isn't.

## V. DISCUSSION

Honestly, we were pleasantly surprised by how well a relatively simple pipeline performed. Much of the recent literature fixates on transformer models and deep learning stacks that require serious computing power. Our results suggest that if you invest properly in text preprocessing and feature engineering, a well-tuned traditional ML model can still hold its own — and it has the added benefit of being explainable and fast.

The system's biggest practical advantage is that it needs nothing beyond the article text itself. Many published systems require user engagement data, social graph information, or posting history — data that platforms are increasingly reluctant to share. Our approach sidesteps this entirely. It can run on a basic server and could realistically be built into a browser extension or a WhatsApp screening tool without any special infrastructure [6].

That said, we have to be honest about what this system cannot do. It was trained and tested entirely on English-language articles, so it would struggle badly with Hindi, Tamil, or Hinglish content — which is where most misinformation in India actually lives. It also cannot handle deepfakes, manipulated images, or videos, which are increasingly the medium of choice for sophisticated disinformation campaigns. And because it relies on word frequencies rather than meaning, a determined bad actor who knows how the system works could probably craft fake articles that slip through. These are not small limitations — they

are the directions where the next generation of research needs to go.

## VI. CONCLUSION

Fake news is one of those problems that technology alone cannot fully solve — it is also a question of media literacy, platform policy, and social responsibility. But technology can definitely help, and AI-based detection systems are a meaningful part of the answer. The system we built in this project shows that even a lightweight NLP and ML pipeline, if properly designed, can achieve 92% accuracy in distinguishing real news from fabricated content.

The broader lesson from our literature review is that the field is maturing quickly but unevenly. We have excellent tools for English-language text but almost nothing robust for the multilingual reality of places like India. We have models that are highly accurate but few that can explain their reasoning in a way a journalist or a fact-checker would find useful. These gaps represent real research opportunities, and we hope this paper contributes a small step toward filling them. Going forward, extending this system to support regional Indian languages, integrating a BERT-based model for improved contextual understanding, and building a real-time browser extension are the most promising next steps.

## REFERENCES

[1] Reuters Institute for the Study of Journalism. (2023). Digital news report 2023. University of Oxford. https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2023

[2] Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. Multimedia Tools and Applications, 80(8), 11765–11788. https://doi.org/10.1007/s11042-020-10183-2

[3] Kaggle. (2020). Fake News Dataset. https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset

[4] World Health Organization. (2020). Managing the COVID-19 infodemic. https://www.who.int/docs/default-source/coronaviruse/risk-comms-updates/update32-infodemic-mgt.pdf

[5] Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Computing Surveys, 53(5), 1–40. https://doi.org/10.1145/3395046

[6] Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2018). Automatic detection of fake news. Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018), 3391–3401.

[7] Ruchansky, N., Seo, S., & Liu, Y. (2017). CSI: A hybrid deep model for fake news detection. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM), 797–806. https://doi.org/10.1145/3132847.3132877

[8] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22–36. https://doi.org/10.1145/3137597.3137600

[9] Wang, W. Y. (2017). 'Liar, Liar Pants on Fire': A new benchmark dataset for fake news detection. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), 422–426. https://doi.org/10.18653/v1/P17-2067

[10] Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on Twitter. Proceedings of the 20th International Conference on World Wide Web (WWW '11), 675–684. https://doi.org/10.1145/1963405.1963500