Journal of Science Research International (JSRI)
ISSN: 2456-6365 (Online)

OPEN ACCESS

TRUTH IN THE AGE OF ALGORITHMS: A COMPREHENSIVE SURVEY OF MACHINE LEARNING TECHNIQUES FOR FAKE NEWS DETECTION

Rashmi Subodh Chavan* and Avinash Baban Kekan

Department of Information Technology,

Pillai College of Arts, Commerce & Science (Autonomous), Navi Mumbai, Maharashtra, India 410206

**Corresponding author E-mail: rashmi@mes.ac.in*

Received: 25 December 2025	Revised: 19 January 2026	Accepted: 21 February 2026	Published: 28 February 2026

DOI: **https://doi.org/10.5281/zenodo.19043037**

Abstract:

The proliferation of digital media has precipitated an unprecedented crisis of information integrity, commonly termed the "infodemic". Fabricated or intentionally misleading content spreads globally in minutes, posing a significant threat to democratic processes, public health, and social cohesion. This paper presents a comprehensive survey of the machine learning (ML) and natural language processing (NLP) techniques developed to address this challenge. We trace the field's evolution through a critical analysis of foundational datasets including LIAR, ISOT, and FakeNewsNet that have shaped research paradigms. Detection methodologies have progressed from early statistical models reliant on linguistic cues to contemporary deep learning architectures such as Long Short-Term Memory (LSTM) networks and Bidirectional Encoder Representations from Transformers (BERT), which capture complex contextual and semantic nuances. Our analysis highlights critical trade-offs between accuracy, precision, recall, and interpretability. This survey concludes that fake news detection is an evolving, adversarial socio-technical challenge.

Keywords: *Fake News Detection, Machine Learning (ML), Natural Language Processing (NLP), Infodemic, Deep Learning, BERT, Transformers, Social Propagation.*

1. Introduction

In the contemporary digital ecosystem, information propagates at a velocity and scale previously unimaginable. Social media platforms, instant messaging applications, and online news portals have democratized content creation, but they have also cultivated a fertile environment for the rapid spread of misinformation and disinformation. This "infodemic" encompasses a broad spectrum of content, from intentionally fabricated articles to subtly misleading headlines designed for engagement. The societal impact is profound, observed in the corruption of democratic elections, the erosion of public trust in foundational institutions like journalism and

science, and the spread of life-threatening health misinformation. Historically, professional journalists and fact-checking organizations served as the primary gatekeepers. However, the rise of the decentralized, user-driven web has dismantled this traditional gatekeeping model. With billions of posts shared daily, human factchecking is a logistical impossibility. The viral nature of digital content means misinformation can achieve widespread reach long before a manual correction can be disseminated. This inherent latency underscores the inadequacy of traditional methods and establishes a pressing imperative for automated, real-time solutions.

ML and NLP have emerged as promising frontiers for developing scalable countermeasures. These models offer the capacity to analyze massive volumes of data at machine speed, identifying subtle patterns, linguistic anomalies, and behavioral signals that indicate manipulative content. This paper aims to synthesize the vast body of research into a coherent narrative, tracing the evolution of the field from foundational concepts to current socio-technical challenges.

2. Literature review: The co-evolutionary trinity

The history of automated fake news detection is a story of co-evolution between three interconnected elements: the conceptual framing of the problem, the nature of the data collected, and the design of the models built to solve it. This "Co- evolutionary Trinity" reveals a self-reinforcing cycle where a more nuanced understanding of the problem drives more sophisticated datasets, which in turn necessitates more advanced models.

2.1 Stage 1: Fake news as a document classification problem

Initially, the field framed the challenge as a straightforward binary document classification task. The goal was to analyze an article's text and assign a label of "real" or "fake" based on intrinsic linguistic properties. This approach treated articles as static, self-contained documents.

The ISOT Fake News Dataset perfectly exemplifies this stage. Released between 2016 and 2017, it contains over 12,600 real articles from Reuters.com and a similar number of fake articles from unreliable websites. A critical feature of ISOT was the preservation of linguistic artifacts in fake articles, such as grammatical errors, unusual punctuation, and sensationalist language, which provided strong discriminative features for early classifiers.

2.2 Stage 2: Fake news as claim-level factchecking

The second stage reframed the problem to acknowledge that truth is often not a simple binary and that entire articles are rarely completely fabricated. The focus shifted to verifying specific, discrete statements or claims embedded within larger texts.

The **LIAR dataset** (2017) catalyzed this shift by providing a large-scale benchmark for nuanced factchecking. It contains 12,836 short statements from PolitiFact.com with a six-class labelling scheme: pants-fire, false, barely true, half-true, mostly true, and true. Crucially, LIAR included rich metadata for each claim, such as the speaker's job title, political affiliation, and "credit history" of truthfulness, acknowledging that source credibility is integral to veracity.

2.3 Stage 3: Fake news as a socio-technical phenomenon

The current stage expands the problem into the socio-technical domain. This framing posits that fake news is a dynamic entity whose identity is shaped by its behavior within a social network. Signals for detection lie in how a story spreads, who amplifies it, and how users engage with it online.

The FakeNewsNet repository operationalizes this perspective by linking news content to social engagement data from platforms like Twitter. It provides a view of a news item's lifecycle, including user profiles, social network

structures (followers/followees), and propagation velocity. This allows researchers to study "behavioral footprints" that distinguish fake news spread from legitimate information flows.

3. Methodology: Paradigms and text representation

The technical core of detection lies in transforming unstructured text into machine-readable features.

3.1 Statistical approaches: BoW and TF-IDF

The earliest methods were based on word frequencies. The Bag-of-Words (BoW) model represents a document as an unordered collection of word counts, disregarding grammar and order. A more sophisticated method is Term Frequency- Inverse Document Frequency (TF-IDF). TF-IDF weights words based on their frequency in a document and their rarity across the corpus.

$$TF - IDF(t, d, D) = TF(t, d) \times IDF(t, D)$$

While efficient, TF-IDF lacks semantic understanding; it cannot grasp that "king" and "queen" are related concepts.

3.2 Semantic embeddings

Breakthroughs came with word embeddings—dense vectors that capture semantic meaning. Unlike TF-IDF, embeddings map similar words close to each other in a continuous vector space.

- **Word2Vec:** Uses a shallow neural network to learn embeddings via Continuous Bag-of-Words (CBOW) or Skip-gram architectures.
- **GloVe:** Captures global word-word co-occurrence statistics from a corpus.
- **FastText:** Represents words as bags of character n-grams, allowing it to handle rare or out-of-vocabulary (OOV) words and morphological similarities.

4. Classification models: From classical to deep learning

4.1 Classical machine learning baselines

Models like Naive Bayes, Logistic Regression, and Support Vector Machines (SVM) remain crucial baselines. SVMs excel at finding separating hyperplanes in high-dimensional TF-IDF feature spaces. Tree-based ensembles like Random Forest and XGBoost combine multiple individual models to improve robustness and reduce overfitting. However, these models cannot grasp the sequential nature of language or complex phenomena like sarcasm.

4.2 Deep learning architectures

To overcome these limits, researchers turned to deep learning to automatically learn hierarchical features.

- **LSTM Networks:** A type of RNN designed to process sequences and capture long-range dependencies, making them effective at capturing "narrative tone".
- **CNNs:** While used for images, CNNs can slide filters over word embeddings to detect local n-gram patterns.
- **BERT (Transformers):** The introduction of the Transformer architecture marked a revolution. BERT's **self-attention mechanism** processes entire sequences at once, learning deep bidirectional relationships between words. BERT is pre-trained on massive corpora and then fine-tuned for specific tasks like fake news detection.

5. Results and performance analysis

Evaluating models requires moving beyond simple accuracy, as imbalanced datasets can lead to misleading results.

Table 1: Performance Benchmark of Machine Learning Models

Model	Accuracy	Precision	Recall	F1-Score
DistilBERT	0.80	0.83	0.84	0.84
BERT	0.78	0.81	0.84	0.83
SGD Classifier	0.79	0.70	0.49	0.57
Linear SVC	0.78	0.67	0.49	0.57
Logistic Regression	0.78	0.69	0.46	0.56
Random Forest	0.74	0.75	0.19	0.30

Transformer-based models achieve significantly higher F1-scores by balancing precision (avoiding false alarms) and recall (catching misinformation). Classical models like Random Forest show extremely low recall (0.19), missing 80% of fake articles. This highlights the "Interpretability-Performance Paradox": Transformers are superior but function as "black boxes," making it hard to justify moderation decisions to users or regulators.

Conclusion and future frontiers

Fake news detection is a dynamic "adversarial arms race". As models improve, creators of misinformation adapt. Future challenges include **multi-modal misinformation** (deepfakes), where text is paired with manipulated images or audio. Research must also integrate social propagation dynamics using Graph Neural Networks (GNNs) to detect coordinated inauthentic behavior. Finally, building models that are robust against adversarial examples rather than just performing well on standard benchmarks is essential for genuine factuality assessment.

References

1. Aggarwal, J., & Kumar, R. (2019). A comparative study of text representation techniques for fake news detection. *Information Processing & Management, 56*(6), 102054.
2. Bharti, R., & Sharma, D. (2021). Survey on fake news detection techniques. *International Research Journal of Engineering and Technology (IRJET), 8*(4).
3. Caselli, T. (2017). Automatic fake news detection: Are we there yet? In *Proceedings of the LREC Conference*.
4. Pérez-Rosas, V., et al. (2018). Automatic detection of fake news. In *Proceedings of the LREC 2018 Conference*.
5. Popat, K., et al. (2018). Credibility assessment of textual claims on the web. *ACM Transactions on the Web, 12*(3).
6. 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.