ARTIFICIAL INTELLIGENCE-BASED FAKE NEWS DETECTION

1 Ishwari Santosh Jadhav, 2 Prof. Vrushali D. Kawade, 2 Prof. Satav Pratik K.

1 Department of Artificial Intelligence & Data Science (AI&DS)

1 Shri Swami Samarth Institute, Malwadi Bota, India

2 Faculty of Engineering & Technology

Abstract: The rapid circulation of information through websites, social networks, messaging services, and video platforms has increased access to news while also enabling misleading or fabricated content to spread quickly. This review examines artificial intelligence (AI)-based fake news detection, focusing on machine learning (ML), natural language processing (NLP), and text-classification techniques. It reviews content-based, linguistic, social-context, knowledge-supported, and hybrid approaches, and discusses a practical baseline pipeline using text preprocessing, Term Frequency–Inverse Document Frequency (TF-IDF), and supervised classifiers such as Multinomial Naive Bayes, Logistic Regression, and linear Support Vector Machines (SVM). Transformer-based approaches are also considered as contextual models. The review highlights dataset quality, data leakage, evaluation measures, temporal generalization, and ethical deployment. An automated classifier can help prioritize items for human verification, but its predicted dataset label is not proof of factual truth. No new experiment was conducted for this review; therefore, no numerical performance results are claimed.

Index Terms— Fake news detection, artificial intelligence, machine learning, natural language processing, TF-IDF, text classification, misinformation, supervised learning.

I. INTRODUCTION

Digital news is distributed through websites, social networks, messaging services, and video platforms. These channels help people access current events, but they also make it easier for inaccurate claims, fabricated stories, manipulated headlines, and old reports presented out of context to reach large audiences. The scale and speed of online distribution create a need for computational tools that can assist journalists, researchers, and fact-checkers.

Fake news detection is commonly treated as a classification task in which a system assigns a label to a news item based on patterns learned from labelled examples. NLP transforms language into representations that algorithms can process. Traditional approaches include Naive Bayes, Logistic Regression, and Support Vector Machines. Neural and transformer-based approaches aim to represent context more effectively.

However, language patterns are not equivalent to factual verification. A false story may imitate professional journalism, while a legitimate article may contain emotional wording or unusual expressions. Reliability can depend on source provenance, publication date, corroborating evidence, and context not present in the article text. An automated detector should therefore be treated as a screening aid, not a final judge of truth.

II. REVIEW OBJECTIVES AND SCOPE

This review summarizes computational perspectives used in fake news detection; discusses NLP preprocessing and TF-IDF; compares traditional classifiers and transformer approaches at a conceptual level; and identifies evaluation, research, and responsible-use challenges. The scope is text-based detection of English-language headlines and article content using labelled datasets and common Python tools. It does not independently verify external claims or examine manipulated images and videos in detail.

III. REVIEW OF RELATED LITERATURE

Shu et al. survey fake news detection from a data-mining perspective and emphasize news content, social context, and user engagement as relevant dimensions [1]. Their perspective indicates that the problem extends beyond the wording of an article. Conroy, Rubin, and Chen review automatic deception-detection methods, including linguistic and network-based signals [2]. Rubin, Chen, and Conroy discuss different types of deceptive news, reinforcing the importance of genre and context [3].

Wang introduces the LIAR benchmark dataset for fake-news detection [4]. Benchmark datasets support model development and comparison, but their label definitions, collection processes, topic coverage, and time periods must be considered before generalizing results. Pedregosa et al. describe scikit-learn, a Python library that supports common machine-learning workflows [5]. Joachims discusses Support Vector Machines for text categorization in high-dimensional feature spaces [6]. Technical documentation from scikit-learn describes TF-IDF vectorization and model evaluation [7], [8].

A. Main Research Perspectives

Content-based methods: analyze words, phrases, headlines, and article text. They are practical for classification but cannot establish truth from wording alone.

Linguistic methods: examine stylistic or deception-related language cues. Such cues may be influenced by satire, genre, author style, and topic.

Social-context methods: use sharing patterns, engagement, and network information. These may add evidence beyond article text but introduce platform dependence and privacy considerations.

Knowledge-supported methods: compare claims with trusted sources, reference knowledge, or external evidence. Their quality depends on the accuracy and currency of those sources.

Hybrid methods: combine content with contextual or external evidence. They may offer a broader view but increase data and system complexity.

IV. TEXT-BASED DETECTION METHODOLOGY

A baseline pipeline accepts a headline, article body, or both; preprocesses the text; extracts numerical features; and applies a trained classifier. The output is a predicted dataset category. It should be accompanied by a caution that the prediction is not independent proof of truth.

A. Dataset Preparation

A labelled dataset is required for supervised classification. Its licence, label definitions, sources, collection dates, and limitations should be checked. Records should be inspected for missing text, duplicates, and class imbalance. Related articles should be grouped where possible to reduce the risk of near-duplicate stories appearing in both training and test sets. Features that directly reveal labels should be avoided.

B. Text Preprocessing

Preprocessing should be consistent during training and prediction. Common steps include handling missing values, normalizing whitespace, and removing malformed records. Lowercasing, stop-word removal, stemming, and lemmatization can be tested, but should not automatically be assumed to improve every dataset. Meaningful negation and phrases should be preserved.

C. TF-IDF Feature Extraction

TF-IDF assigns weight to terms according to their frequency in a document and their relative rarity across the corpus. Word unigrams and bigrams can represent individual terms and short phrases. The vectorizer must be fitted only on training data to prevent test-set information leakage [7].

D. Classification Algorithms

Multinomial Naive Bayes: a fast probabilistic baseline commonly used for text classification.

Logistic Regression: a linear classifier that works with sparse TF-IDF features and provides a useful baseline.

Linear SVM: a margin-based classifier suited to high-dimensional text categorization; probability estimates require calibration when needed [6].

Transformer models: pretrained language models can be fine-tuned to learn contextual representations, but require careful evaluation and computational resources.

V. CONCEPTUAL SYSTEM ARCHITECTURE

The reviewed system can be represented as a sequence of stages: news-text input, preprocessing, TF-IDF feature extraction, a trained machine-learning classifier, and a predicted label accompanied by a caution. This is a conceptual architecture, not a claim that a working system was experimentally validated.

VI. EVALUATION AND COMPARATIVE DISCUSSION

Accuracy measures the proportion of correct predictions. Precision measures how often positive predictions are correct, recall measures how many actual positive cases are detected, and the F1-score combines precision and recall. The positive class must be defined clearly. A confusion matrix and per-class metrics should accompany aggregate accuracy because accuracy alone can conceal poor performance on one class [8].

All classifiers should be evaluated on identical data partitions. Parameter tuning should use cross-validation within the training data, while the held-out test set should remain untouched until final evaluation. Random splits can overestimate performance when near-duplicate articles occur across partitions. Time-based or publisher-held-out testing can provide stronger evidence of generalization.

A. Comparative Characteristics

Naive Bayes offers a simple and computationally efficient baseline. Logistic Regression provides a linear decision boundary for sparse text features. Linear SVM is established for high-dimensional text categorization [6]. Transformer models can represent

context more richly, but added complexity does not guarantee greater reliability. Every approach depends on representative data and valid labels.

B. Experimental Status

No model was trained and no dataset was experimentally evaluated for the source paper or this review. Therefore, no measured accuracy, precision, recall, F1-score, or model ranking is reported. The following values must be filled only after conducting a documented experiment.

VII. RESEARCH GAPS AND ETHICAL CONSIDERATIONS

Dataset labels may be incomplete or tied to particular publishers, events, and collection practices. Models can learn source- or topic-specific shortcuts instead of general misinformation indicators. Satire and genre further complicate interpretation. Misinformation tactics and language change over time, so models may lose reliability when applied to new events or sources.

False positives can wrongly flag legitimate reporting, while false negatives can allow misleading content to pass through. Systems should communicate uncertainty, retain human review, and avoid automatic censorship based only on model output. User submissions should be protected, unnecessary personal information should not be collected, and data use should be explained. Performance should be assessed across topics and sources where possible.

VIII. FUTURE RESEARCH DIRECTIONS

Future work can compare transformer models with traditional TF-IDF baselines using controlled data partitions; connect claims to trusted fact-checking databases and primary sources; use publication dates and source provenance to reduce leakage; extend evaluation to Marathi and other regional languages with labelled datasets; and develop a web interface that communicates uncertainty and monitors performance drift.

IX. CONCLUSION

This review has discussed AI-based fake news detection as a text-classification and evidence-assessment problem. A practical baseline combines NLP preprocessing, TF-IDF features, and supervised classifiers such as Multinomial Naive Bayes, Logistic Regression, and linear SVM. Transformer models and social-context or knowledge-supported approaches provide additional directions for investigation.

Textual patterns alone cannot establish factual truth. Dataset quality, label definitions, leakage prevention, temporal and source-based generalization, and responsible communication of uncertainty are essential. Automated detection is best treated as a tool for prioritizing content for human verification, not as an autonomous authority on truth. Since no experiment was conducted, empirical performance remains to be measured in future work.

Conceptual workflow

INPUT NEWS TEXT Headline / article body
TEXT PREPROCESSING
TF-IDF FEATURE EXTRACTION
TRAINED ML CLASSIFIER
PREDICTED LABEL + CAUTION

Fig. 1. Conceptual workflow for text-based fake news classification.

Experimental comparison template

Model	Accuracy	Precision	Recall	F1
Naive Bayes	TBD	TBD	TBD	TBD
Logistic Regression	TBD	TBD	TBD	TBD
Linear SVM	TBD	TBD	TBD	TBD

Table I. Experimental comparison template; TBD = to be determined after testing.

REFERENCES

[1] S. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake News Detection on Social Media: A Data Mining Perspective,” ACM SIGKDD Explorations Newsletter, vol. 19, no. 1, pp. 22–36, 2017.

[2] N. J. Conroy, V. L. Rubin, and Y. Chen, “Automatic Deception Detection: Methods for Finding Fake News,” Proceedings of the Association for Information Science and Technology, vol. 52, no. 1, 2015.

[3] V. L. Rubin, Y. Chen, and N. J. Conroy, “Deception Detection for News: Three Types of Fakes,” Proceedings of the Association for Information Science and Technology, vol. 52, no. 1, 2015.

[4] W. Y. Wang, “Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection,” Proceedings of ACL, vol. 2, pp. 422–426, 2017.

[5] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[6] T. Joachims, “Text Categorization with Support Vector Machines: Learning with Many Relevant Features,” Proceedings of ECML, pp. 137–142, 1998.

[7] scikit-learn, "TfidfVectorizer Documentation." Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[8] scikit-learn, "Model Evaluation." Available: https://scikit-learn.org/stable/modules/model_evaluation.html

[9] X. Zhou and R. Zafarani, "A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities," ACM Computing Surveys, vol. 53, no. 5, Article 109, pp. 1–40, 2020, doi: 10.1145/3395046.

[10] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, "Detection and Resolution of Rumours in Social Media: A Survey," ACM Computing Surveys, vol. 51, no. 2, Article 32, pp. 1–36, 2018, doi: 10.1145/3161603.

[11] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media," Big Data, vol. 8, no. 3, pp. 171–188, 2020, doi: 10.1089/big.2020.0062.

[12] A. D'Ulizia, M. C. Caschera, F. Ferri, and P. Grifoni, "Fake News Detection: A Survey of Evaluation Datasets," PeerJ Computer Science, vol. 7, e518, 2021, doi: 10.7717/peerj-cs.518.

[13] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: A Large-scale Dataset for Fact Extraction and Verification," in Proceedings of NAACL-HLT, vol. 1, pp. 809–819, 2018, doi: 10.18653/v1/N18-1074.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of NAACL-HLT, vol. 1, pp. 4171–4186, 2019, doi: 10.18653/v1/N19-1423.

[15] S. Vosoughi, D. Roy, and S. Aral, "The Spread of True and False News Online," Science, vol. 359, no. 6380, pp. 1146– 1151, 2018, doi: 10.1126/science.aap9559.

Note: The references are retained from the author's supplied paper. Bibliographic details and the citation style should be verified against original publications and departmental requirements before submission.

Copyright & License:
© Authors retain the copyright of this article. This work is published under the Creative Commons Attribution 4.0 International License (CC BY 4.0), permitting unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.