Received: 26/04/2026, Accepted: 16/05/20026

AI-Based Fake News Detection Using NLP and Machine Learning Techniques: A Review

Apoorv Mahere, Ashutosh, Ayush Srivastava, Abhay Bhardwaj
Department of CSE, Galgotias College of Engineering and Technology, Uttar Pradesh, India
madhavkumar696969@gmail.com, ashuashusingh.2004@gmail.com, kwoctcrazy@gmail.com, abhay.bh07@gmail.com

ABSTRACT

Digital communication channels have been agents of change, radically reshaping how information is created, engaged with, and disseminated within societies. This democratisation of content has allowed users around the world to gain greater power, but it has also led to the easy spread of artificial stories, known as fake news. These traditional fact-checking methods cannot keep up with the pace and scale of misinformation online, so there is a need for independent computational infrastructure that can automatically assess textual credibility at scale. This paper aims to provide a comprehensive theoretical framework for developing an AI-driven fake news detection system based on the principles of classical Machine Learning (ML) and Natural Language Processing (NLP). Empirical studies are excluded here, with emphasis on the theoretical, mathematical and linguistic aspects needed to design such a system. A full theoretical framework is developed from the characterisation of the datasets, text pre-processing, and the TF-IDF vectorisation. The authors examine in detail four classification algorithms: Logistic Regression, Decision Tree, Random Forest and Gradient Boosting; they are explored in terms of their mathematical formulation, their suitability to high-dimensional sparse text data and their ability to adhere to linguistic patterns typical of deceptive content. This paper does not present empirical findings but rather lays the groundwork for future experiments, performance benchmarking, and deployment in misinformation detection applications in a structured framework for conceptualization.

Keywords: *Fake News Detection, Natural Language Processing, TF-IDF, Logistic Regression, Decision Tree, Gradient Boosting, Random Forest, Text Classification*

1. Introduction

Digital platforms like social media, online news portals, and blogging sites have revolutionized information dissemination. This transformation has come with substantial consequences. The low barrier of content creation, combined with algorithm-driven amplification, enables fabricated narratives to propagate at unprecedented speed like wildfire. In traditional journalism, the online environment lacks editorial oversight, creating fertile ground for misinformation designed to influence the masses, divide people, or profit from sensationalism [1].

Fake news is distinguished by other types of low-quality text, like spam or text generated by bots. It is purposely made to look believable, logical and emotionally compelling. The motive is not only to cheat, but also to clutter. This renders automated detection much more challenging. The word patterns that are enshrined in deceptive text tend to ape acceptable journalistic standards artificially. blurring the line between reality and fiction [1].

The main objective of this study is to come up with a theoretically sound framework of automated detection of fake news classification with NLP and classical ML technologies. Although no model in this stage is put in place, the theoretical underpinning and foundations deposited here are necessary for the evaluation of the future in an empirical way.

*Corresponding author: Apoorv Mahere, Department of Computer Science, Galgotias College of Engineering and Technology, Uttar Pradesh, India (madhavkumar696969@gmail.com)

1.1 Research Question

To what extent can linguistic cues, statistical text features, and machine learning decision mechanisms theoretically support automated fake news detection, and how should these components be structured for future implementation?

1.2 Scope of This Work

This paper is concerned only with:

- knowledge of linguistic cues of misinformation.
- studying mathematical and theoretical characteristics of ML algorithms.
- defining the right assessment plans for future experiments.

It does not include:

- training, testing, or tuning of an empirical model,
- numerical accuracy or performance reports,
- implementation of trained systems.

It is still focused on developing a rigorous conceptual basis on which subsequent empirical researches can be based.

2. Literature Review

AI Fake news detection is a machine learning field that combines computational linguistics, psychology, and machine learning knowledge. Current literature reveals that there are theoretical dimensions that are applicable in this research.

2.1 Content-Based Linguistic Approaches

Initial studies viewed fake news as a linguistic classification issue which is motivated by lexical, syntactic, and stylistic hints. The LIAR dataset by Wang [3] showed that shallow types of ML models could be used to classify deceptive text when well-engineered features were used.

Previous research indicates the existence of a number of linguistic associations:

- Emotionally charged vocabulary: Fake news often employs a lot of emotion with the aim of evoking psychological biases [2].
- Speculative or hedging phrases: Phrases such as “sources claim” or “it is believed” appear more frequently used in deceptive narratives.

These results support the use of text-based features in ML pipelines.

2.2 Classical Machine Learning Approaches

Although eclipsed by modern neural networks, classical ML models are still good at text classification tasks. Reasons include:

- **Sparse, high-dimensional suitability:** TF–IDF vectors result in a very sparse feature space where algorithms like Logistic Regression or Decision Trees perform very well.
- **Bias–variance flexibility:** Random Forest and Gradient Boosting models have tunable trade-offs that are suitable in the real world.
- **Interpretability:** Classical ML provides clearer insight into which words or patterns contribute to classification decisions.

Experiments like [7] indicate that ensemble algorithms are better at capturing nonlinear interactions among linguistic cues than single classifiers.

2.3 Neural and Transformer Architectures

Transformer-based models (BERT, RoBERTa) have achieved state-of-the-art performance by learning semantic relationships via attention mechanisms [2]. However, they are very big power consumers, difficult to implement as lightweight real-time systems and they need large labelled datasets. For the purposes of this theoretical framework, classical ML is considered the foundational approach, with Transformer-based extension identified as a priority for future work.

3. Research Methodology

The methodology to be implemented in the future, with empirical testing, has a number of systematically integrated components. Fig. 1 shows the entire workflow.

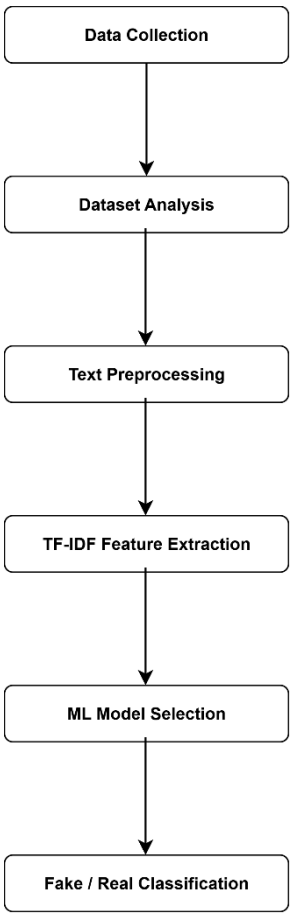

Figure 1: Methodology workflow for fake news detection.

Figure 1 shows a sequential pipeline that starts with a News Event triggering the workflow. In the Data Collection phase, raw articles are collected from the structured databases. This is followed by Data Preprocessing, which involves removing noise, tokenising text, removing stopwords, and TF-IDF vectorising the raw tokens into numerical feature representations. The Analysis of Data step includes exploratory analysis of distributions of features and linguistic patterns aimed at selecting a model. In the Training the Model stage, the classifiers theoretically learn the training partition. Once the decision logic has been trained, the Classification Algorithm uses the learned logic to classify unseen examples, while Testing the Model assesses performance using standard metrics such as accuracy, precision, recall, and F1 score. The last step is Fake News Detection and Analysed Result, where the model outputs a binary classification (real or fake) and an interpretive analysis of the model's decision patterns.

3.1 Dataset Description

The dataset selected contains 23,481 fake news articles and 21,417 real news articles. Table 1 provides a comparative overview of available fake news datasets.

Table 1: Comparison of Major Fake News Datasets

Dataset	Size	Labels	Source
LIAR [3]	12.8k	6-class	PolitiFact
FakeNewsNet	23k+	Binary	BuzzFeed, PolitiFact
FVC	45k	Binary	Multiple News Portals
FakeImages	30k	Binary	Social Media Posts
FakeText	20k	Binary	Twitter Threads
COVID-19 FN	10k	Binary	WHO, Fact-Checkers

A qualitative analysis of the texts shows fake news stories consistently follow urgent stylistic trends (all-caps, multiple exclamation points), contain emotionally charged or fear-inducing words and lack named sources and verifiable facts. Legitimate news articles consistently contain datelines, named sources and verifiable facts. These features guide the preprocessing and feature extraction [9].

3.2 Text Preprocessing

Raw textual information is untidy and unorganised. Fig. 2 depicts the preprocessing pipeline used to prepare text for numerical analysis.

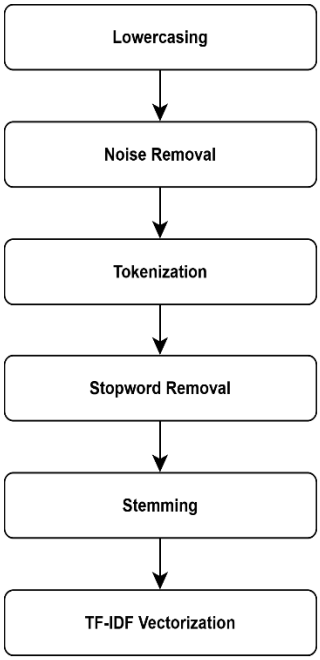

Figure 2: Text preprocessing steps.

The following are the stages of the preprocessing pipeline:

1) Lowercasing: standardises word forms by removing case redundancy. It is especially relevant in text that may contain fake news, as this often uses all-caps words to create sensational headlines, which will be treated the same as their lowercase equivalents.
2) Noise Removal: removes punctuation, URLs and symbols with no meaningful content. Some of the fake news articles have too much punctuation, and many irrelevant links to

other websites are embedded within the text and these are removed so that they do not bias the feature space.

3) Tokenisation: breaks down text into lexical units for processing. By breaking the articles down into their constituent parts, the model can pinpoint the words and phrases that lend language a deceptive quality, including emotionally charged words and hedging language.

4) Stopword Removal: eliminates common, non-informative words to reduce the feature space. The removal of high-frequency function words helps to ensure that the TF-IDF weights are based on terms that are meaningful for the content that have deceptive or credible meaning rather than grammatical fillers.

5) Stemming: reduces feature sparsity by reducing morphemes to their root. Stemming is particularly important for TF-IDF feature vectors in text classification problems because it collapses inflected forms into a single stem, thereby reducing dimensionality and semantically linking related terms.

6) TF-IDF Vectorisation: converts preprocessed tokens to a numerical feature matrix. A major advantage of TF-IDF over just word counts is that it downweights common terms that occur in many documents and upweights terms that are unique to certain articles, which can help identify unusual vocabulary patterns that characterise fake news content.

3.3 Mathematical Foundation of TF-IDF

TF-IDF assigns each term a weight that reflects its importance within a specific document relative to the broader corpus.

Term Frequency:

$$\mathrm{TF}(t,d) = \frac{f_{t,d}}{\sum_k f_{k,d}}$$

Inverse Document Frequency:

$$IDF(t) = \log\left(\frac{N}{1+n_t}\right)$$

Combined TF-IDF Score:

$$TF\text{-}IDF(t,d) = TF(t,d) \times IDF(t)$$

where f(t,d) is the frequency of term t in document d, N is the total number of documents, and n_t is the number of documents containing term t.

3.4 Planned Train-Test Split

During the empirical evaluation, an 80-20 train-test split will be used to ensure the class ratios in the train and test sets are similar. Cross-validation techniques will be applied to validate the results.

4. Theory and Calculation: Machine Learning Models

This section presents the theoretical basis and mathematical formulation of the four classifiers selected for this framework.

4.1 Logistic Regression

Logistic Regression is a discriminative linear classifier which fits the conditional probability of a document being fake using the sigmoid function:

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$$

Model parameters are optimised by minimising the binary cross-entropy loss:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log h_\theta\left(x^{(i)}\right) + \left(1-y^{(i)}\right)\log\left(1-h_\theta\left(x^{(i)}\right)\right)\right]$$

Logistic Regression is highly efficient with sparse and high-dimensional TF-IDF features. Its weight vector is directly interpretable, revealing the terms that contribute to the classification decision, and it provides a strong linear baseline for comparison [8].

4.2 Decision Tree Classifier

Decision Trees divide the feature space by selecting the attribute that maximises information gain. An example of binary classification of fake news is shown in Figure 3. The impurity of a node is measured by Shannon's entropy:

$$H(S) = -\sum_i p_i \log_2(p_i)$$

Information gain for attribute A is:

$$IG(S,A) = H(S) - \sum_v \frac{|S_v|}{|S|} H(S_v)$$

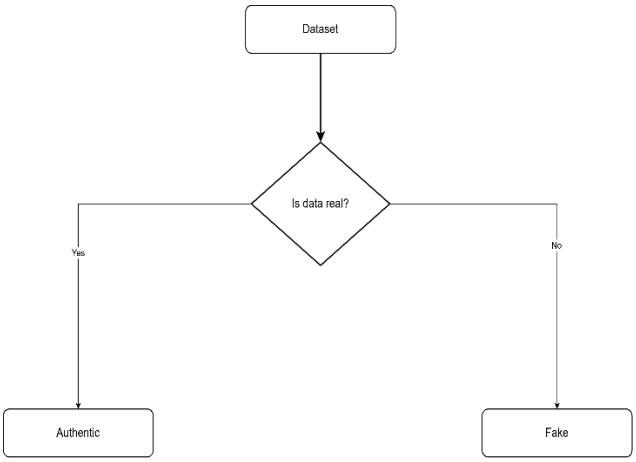

Figure 3: Decision Tree structure for fake news classification.

Decision Trees can capture complex patterns between language features and provide interpretable rules. They are prone to overfitting in high-dimensional spaces; hence, they are commonly used in ensemble methods.

4.3 Gradient Boosting

Gradient Boosting builds a strong classifier by successively adding weak learners, each trained to improve the ensemble's performance. The architecture of parallel sub-data and the decision tree is illustrated in Figure 4:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

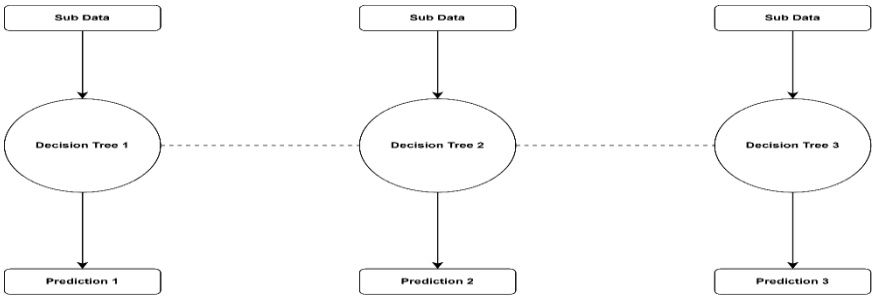

Figure 4: Gradient Boosting architecture.

Such iterative bias-correction process makes Gradient Boosting a good choice of model for capturing subtle and context-sensitive language cues of fake news. The main drawback is the computational complexity, which increases with the number of trees.

4.4 Random Forest

Random Forest overcomes the problem of overfitting of individual Decision Trees, by training many trees, each on a bootstrap sample of the data, and a random subset of the features. The data splitting and voting is demonstrated in Figure 5. Finally, the mode function is used to make predictions:

$$\hat{y} = \text{mode}\,(h_1(x), h_2(x), \ldots, h_K(x))$$

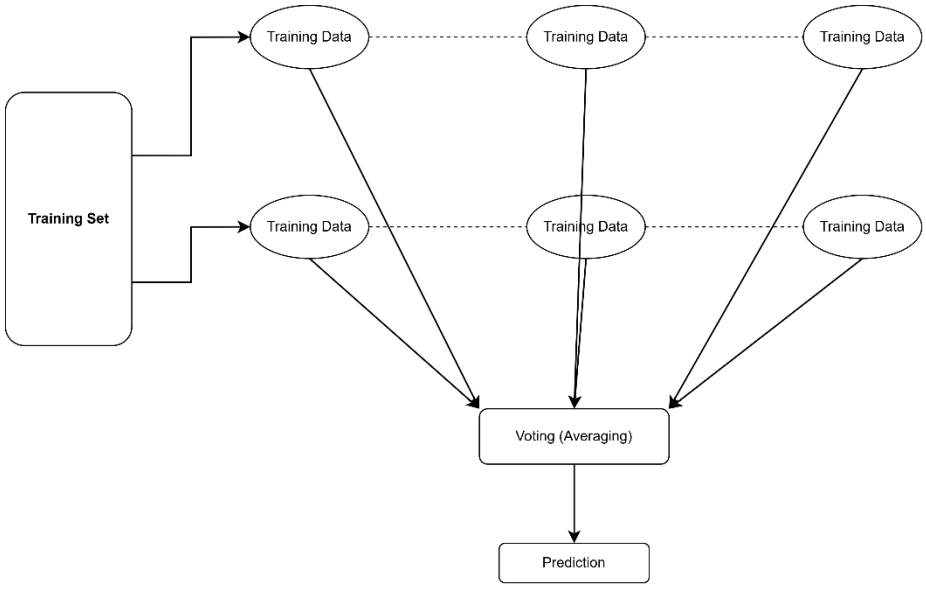

Figure 5: Random Forest workflow.

The ensemble averaging mechanism substantially reduces variance while preserving the capacity to model nonlinear feature interactions. Random Forest handles noisy TF-IDF features effectively and resists overfitting without requiring careful pruning.

5. Results and Discussion

This paper is theoretical; no empirical results are reported here. Here, provides an overview of the evaluation framework, the benchmark scenario, and the anticipated results.

5.1 Performance Metrics

The following standard classification metrics will be recorded upon empirical implementation:

- Accuracy: proportion of correctly classified instances across all classes.
- Precision: proportion of predicted fake articles that are genuinely fake.
- Recall: proportion of all fake articles that the model correctly predicted.
- F1 Score: weighted average of precision and recall that offers a single metric.

5.2 Comparative Benchmark

Table 2 summarises the reported benchmark performance metrics from recent deep learning models, providing a reference point for evaluating the classical ML framework.

Table 2: Reported Benchmark Metrics from DL Models

Metric	Reported Value
Single Prediction Latency (GPU)	0.2 – 0.5 seconds

Single Prediction Latency (CPU)	1.0 – 2.0 seconds
Batch Inference – 10 samples (GPU)	2.0 – 3.0 seconds
Memory Usage	~500 MB model + 2 GB inference
Fake Text Accuracy (FVC dataset)	94.1%
Fake Image Accuracy (FakeSV dataset)	92.3%
Fake URL Accuracy (FakeTT dataset)	89.7%

5.3 Technique Comparison

Table 3 presents a structured comparison of the detection techniques considered in this framework.

Table 3: Strengths and Weaknesses of Different Fake News Detection Techniques

Technique	Strengths	Weaknesses	Use-case Suitability
Bag-of-Words	Fast, Simple	Ignores Order / Context	Rapid prototyping on small corpora
TF-IDF	Good Term Weighting	Sparse, High Dimensional	Classical NLP pipelines with sparse text
LSTM	Captures Sequence	Requires More Data	Sequence-aware detection with sufficient labelled data
CNN for Text	Captures Local Patterns	Weak Global Context	Local pattern detection in article structure
Transformers	State-of-the-Art Accuracy	High Computation Cost	High-accuracy production systems
Ensembles	Robust, Powerful	Less Interpretable	Robust detection across heterogeneous datasets

5.4 Classical ML Model Comparison

Table 4 presents a direct comparison of the classical ML classifiers evaluated in this framework.

Model	Strengths	Limitations	Use-case Suitability
Logistic Regression	Interpretable, Fast	Linear Boundaries Only	Baseline systems & interpretability focused pipelines
SVM	High Accuracy	Expensive on Large Data	High dimensional text with large training sets
Decision Tree	Interpretable Rules	Easily Overfits	Transparent rule extraction for audit trails

Random Forest	Robust to Noise	Less Interpretable	General purpose fake news detection with noisy data
Gradient Boosting	High Predictive Power	Computationally Heavy	Performance critical production deployments
Naive Bayes	Fast, Simple	Assumes Feature Independence	Low resource settings requiring fast inference

Table 4: Comparison of Classical ML Models for Fake News Detection

5.5 Expected Theoretical Outcomes

Based on the theoretical analysis and the literature, we expect to see:

- Logistic Regression will provide a good linear baseline.
- Decision Trees will uncover explicit linguistic patterns in deceptive content.
- Random Forest will be more generalisable than individual trees due to variance reduction.
- Gradient Boosting will likely be most discriminatory due to the use of subtle interactions between features.

6. Conclusions

This paper has offered a holistic theoretical perspective on the computer-aided detection of fake news using Natural Language Processing and classical Machine Learning methods. The study shows that classical ML algorithms, especially ensemble methods such as Random Forests and Gradient Boosting, are theoretically well-suited to high-dimensional, sparse feature vectors arising from text vectorisation and can identify subtle linguistic signatures that separate fake from genuine news. The framework recognises the potential limitations to its approach: TF-IDF feature vectors do not carry semantic information and the linguistic patterns of fake news exhibit temporal drift, which can affect model generalisation using historical data records. The next steps will be the empirical realisation of the proposed framework, benchmarking the performance of the four classifiers, and augmenting the system with Transformer models such as BERT. The long-term aim is to develop the system as a lightweight, real-time browser extension that can warn users about potentially misleading information as they encounter it.

Acknowledgements

The authors wish to thank the Department of Computer Science at Galgotias College of Engineering and Technology, Greater Noida, UP, India for providing the academic resources and support that made this research possible.

References

[1] K. 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] X. Zhou and R. Zafarani, “A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities,” ACM Computing Surveys, vol. 53, no. 5, pp. 1–40, 2020.

[3] W. Y. Wang, “‘Liar, Liar Pants on Fire’: A New Benchmark Dataset for Fake News Detection,” in Proc. 55th Annual Meeting of the Association for Computational Linguistics (ACL), Vancouver, Canada, 2017, pp. 422–426.

[4] M. L. Della Vedova, E. Tacchini, S. Moret, G. Ballarin, M. DiPierro, and L. de Alfaro, “Automatic Online Fake News Detection Combining Content and Social Signals,” in Proc.

22nd Conference of Open Innovations Association (FRUCT), Jyvaskyla, Finland, 2018, pp. 272–279.

[5] A. P. S. Bali, P. Bhatt, A. Ahmad, S. Ranka, and P. Rai, “Comparative Performance of Various Machine Learning Algorithms for Fake News Detection,” in Proc. ICACDS, Ghaziabad, India, 2019, pp. 279–289.

[6] K. 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, 2018.

[7] E. Monti, F. Frasca, D. Eynard, A. Mannion, and M. Bronstein, “Fake news detection on social media using geometric deep learning,” arXiv preprint arXiv:1902.06673, 2019.

[8] G. Shrivastava, P. Kumar, R. P. Ojha, P. K. Srivastava, S. Mohan, and G. Srivastava, "Defensive modeling of fake news through online social networks." IEEE Transactions on Computational Social Systems 7.5 (2020): 1159-1167.

[9] V.K. Mishra, K. Sharma, V. Sharma, and G. Shrivastava, "A Machine Learning based approach to detect fake news in social media." In 2023 5th International conference on advances in computing, communication control and networking (ICAC3N), pp. 1029-1036. IEEE, 2023.