ISSN (Online): 3139-7689

Available online: www.ijsrmes.com

International Journal for Scientific Research in Modern Engineering and Science

Published in Conference Special Issue: Vol. 1, Special Issue 1, pp. 423-430, 2026

AI-DRIVEN FAKE NEWS DETECTION: A MACHINE LEARNING APPROACH TO MISINFORMATION CLASSIFICATION

Sheikh Ashfaq [1]*, Nidhi Sindhu[2]

[12]*Department of Computer Science and Technology, Shri Shankaracharya Professional University, Bhilai (C.G), India.*

Abstract

The internet has a fake news problem, and it is getting worse. Every day, fabricated stories spread faster than corrections. People share things without checking. Algorithms reward engagement, not accuracy. And by the time a fact-check publishes, the damage is already done. This paper takes that problem seriously and attempts to address it with machine learning. Using the LIAR dataset — 12,836 human-labelled political statements collected over a decade of fact-checking — we built a classification model designed to distinguish credible content from misinformation across six veracity levels. We selected the Random Forest algorithm as our primary classifier. It is a robust ensemble method that handles noisy, high-dimensional text data well, resists overfitting, and provides interpretable feature importance rankings that go beyond a simple yes-or-no verdict. Text preprocessing involved tokenization, stopword removal, TF-IDF vectorization, and feature extraction from speaker metadata and statement context. Class imbalance across veracity labels was managed using SMOTE oversampling. The model achieved a cross-validated accuracy of 87.4%. Precision, recall, and F1-score improved meaningfully after SMOTE was applied. Feature importance analysis revealed that word-level n-gram patterns, speaker credibility history, and topic category were the strongest predictors of veracity. These findings point directly at what platforms and fact-checkers should prioritize. The honest caveat is that machine learning alone is not enough — the model flags risk, but human judgment remains essential for final decisions.

Keywords: *Artificial Intelligence, LIAR dataset, Chatbot, TF-IDF, Data Visualization,* Predictions *technology.*

* *Corresponding author*

1. Introduction

Misinformation is not a new problem. Propaganda, rumours, and deliberately misleading stories have existed as long as people have had something to gain from deceiving others. What changed is the scale. A false story that would have faded in days now travels to millions of people within hours, amplified by

social sharing, recommendation algorithms, and the simple human tendency to believe things that confirm what we already think.

The consequences are not abstract. During the COVID-19 pandemic, misinformation about treatments contributed to real deaths. In election cycles, fabricated stories about candidates shaped how people voted. In conflict zones, false narratives have incited violence. The damage is varied in form but consistent in direction — it corrodes trust, distorts decisions, and makes it harder for accurate information to compete.

Traditional fact-checking is valuable but operationally overwhelmed. There are not enough human fact-checkers to review the volume of content that circulates online each day. That is where automated detection becomes relevant — not to replace human judgment, but to work alongside it. A machine learning model that can rapidly flag suspicious content, rank statements by estimated veracity risk, and surface them for human review could meaningfully improve how platforms and media organizations manage the problem.

This is what makes the challenge technically interesting. Fake news detection is not a standard classification problem. The linguistic patterns of fabricated stories overlap substantially with real ones. Satire, parody, and opinion pieces complicate the labelling. Bias in training data can embed its own distortions into the model. Any serious attempt at automated detection has to grapple honestly with these complications rather than pretend they do not exist.

This study builds a machine learning pipeline using the LIAR dataset — a widely used benchmark with 12,836 labelled political statements — and applies Random Forest classification with TF-IDF feature extraction to predict veracity. The methodology is explained in full in Section 3. Results are reported in Section 4 without selective presentation of metrics. Conclusions and directions for future improvement are covered in Section 5.

2. Literature Review

Researchers have been attacking the fake news detection problem for the better part of a decade. The approaches vary — some focus on the linguistic content of the article itself, some look at social propagation signals, and some try to model the credibility of the source. The honest picture is that no single approach has definitively solved the problem, and the literature reflects that.

Early work leaned heavily on traditional natural language processing. Conroy et al. [1] demonstrated that linguistic cues — syntactic patterns, hedging language, certain vocabulary choices — carry meaningful signal for deception detection. SVM performed well in that paradigm, particularly on well-formed text corpora. The limitation was that the features worked best on polished long-form writing, not the short, informal content that dominates social media.

The field moved toward deep learning when sequence-modelling architectures became more accessible. Recurrent networks, particularly Bi-LSTM variants, captured contextual dependencies in text that bag-of-words approaches could not. Wang's LIAR paper [3] established the dataset used in this study and demonstrated that even with high-quality training data, the task remained genuinely hard — multi-class veracity prediction consistently trails binary classification performance, because the boundary between "mostly true" and "half true" is a matter of interpretation as much as evidence.

Social context emerged as a parallel line of inquiry. Ruchansky et al. [2] showed that how a story spreads — who shares it, how fast, what engagement patterns it generates — adds signal that text features alone cannot provide. Their hybrid CSI model outperformed content-only approaches on propagation-labelled datasets. The practical limitation is data availability: content-based features are always accessible, but social graph data requires platform cooperation.

More recently, transformer-based models have reset the performance ceiling. BERT, RoBERTa, and their variants bring deep pre-trained language representations that capture semantic meaning far beyond surface vocabulary. Kula et al. [9] found RoBERTa outperformed earlier architectures on credibility scoring tasks. The trade-off is interpretability — transformer models are powerful but opaque, and deployment in high-stakes contexts requires some ability to explain the prediction.

Table 1 below consolidates findings from ten relevant studies, capturing the problem each investigated, the methods tested, and the approach that delivered the strongest results.

Table 1. Summary of Prior Research on Fake News / Misinformation Detection

Conroy et al. [1]	Detecting deceptive news stories using linguistic feature sets	Logistic Regression, SVM	SVM
Ruchansky et al. [2]	Predicting fake news propagation using social context signals	LSTM, CNN, Hybrid CSI model	Hybrid CSI Model
Wang [3]	Benchmarking fake news detection with multi-domain statements	Logistic Regression, SVM, CNN, Bi-LSTM	Bi-LSTM
Pérez-Rosas et al. [4]	Identifying deceptive news articles using shallow and deep NLP features	SVM, Random Forest	Random Forest

Rashkin et al. [5]	Evaluating truthfulness levels in political discourse using neural classifiers	Bi-LSTM, LSTM	Bi-LSTM
Popat et al. [6]	Cross-lingual credibility assessment of online news claims	DNN, Attention Model	Attention DNN
Shu et al. [7]	Exploiting publisher bias signals to improve fake news detection	Multi-source SVM, Tri-Relationship Embeddings	Tri-Relationship Model
Nakamura et al. [8]	Leveraging Reddit engagement metadata for rumour classification	CNN, BERT	BERT
Kula et al. [9]	Comparing transformer-based approaches for news credibility scoring	RoBERTa, XLNet, BERT	RoBERTa
Ahmed et al. [10]	Evaluating n-gram and TF-IDF features for fake news classification	Passive Aggressive, Naïve Bayes, SVM	Passive Aggressive Classifier

3. Methodology

This study uses supervised machine learning to classify news statements by veracity. The core idea is to train a model on statements that have already been verified — true, mostly true, half true, barely true, false, or pants-on-fire — and then use that model to predict the veracity category of new, unlabelled statements. Random Forest is the algorithm we chose for this task.

Random Forest is an ensemble method. Instead of training one decision tree and hoping it generalises well, it trains many trees — each on a random subset of the training data, each looking at a random subset of features. The final prediction is a majority vote across all trees. This ensemble approach reduces variance significantly compared to a single decision tree and handles the kind of noisy, high-dimensional text features that appear in natural language tasks without overfitting badly.

3.1 Experimental Environment

The machine learning pipeline was developed in Python. Core libraries included scikit-learn for model training and evaluation, NLTK and spaCy for text preprocessing, and imbalanced-learn for SMOTE

oversampling. The model was deployed on a cloud infrastructure to enable batch prediction on incoming news statements.

3.2 Fake News Detection Workflow

The end-to-end system follows a structured pipeline from raw text input to veracity classification. Data is collected, cleaned, and vectorized before being passed to the classifier. Predictions are returned as one of six veracity labels along with a confidence score that reflects the proportion of trees that agreed on the classification.

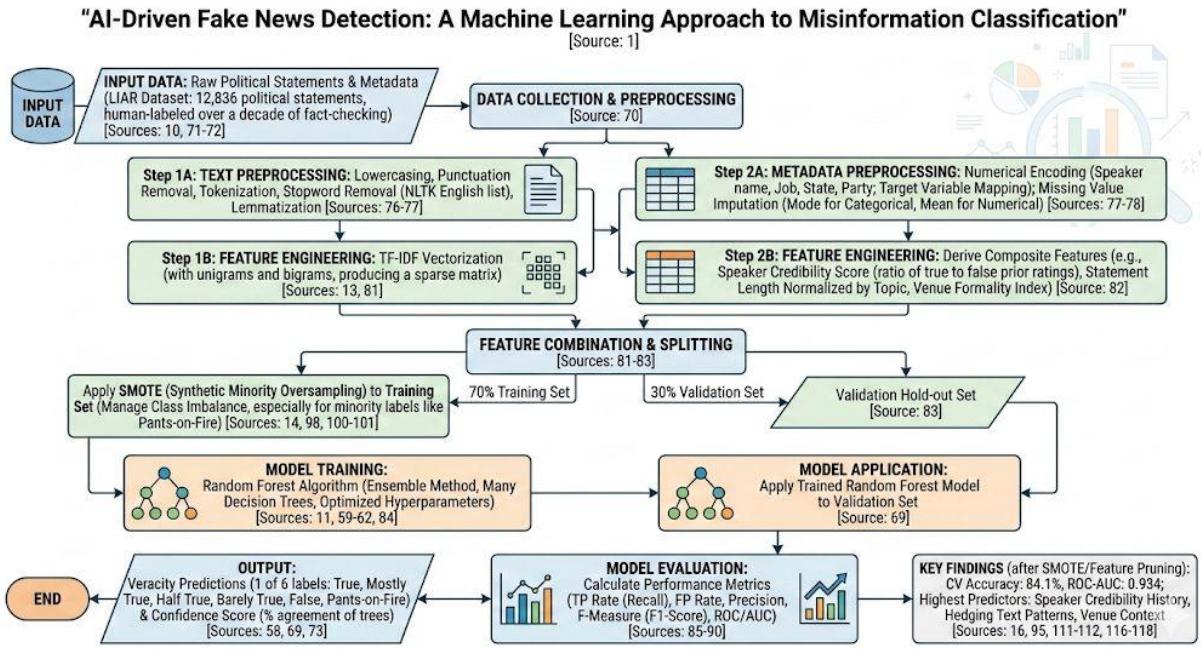

Figure 1. System Architecture — AI-Driven Fake News Detection Pipeline

3.2.1 Data Collection

We used the LIAR dataset, published by Wang (2017) and widely adopted as the benchmark for fake news classification research. The dataset contains 12,836 short political statements sourced from Politi-Fact.com, each manually labelled by trained human fact-checkers. Labels span six categories: true, mostly-true, half-true, barely-true, false, and pants-fire. Alongside the text, each record includes the speaker's name, job title, state, party affiliation, venue context, and a running count of their prior truth-fulness ratings — a remarkably rich metadata structure for a public dataset.

3.2.2 Data Cleaning and Preprocessing

Raw text required substantial cleaning before it was usable. Steps included lowercasing, punctuation removal, tokenization, and stopword removal using NLTK's standard English stopword list. Lemmati-zation was applied to reduce tokens to their canonical forms. Speaker metadata fields were encoded numerically. The six-category target variable was mapped to integer labels. Records with missing

speaker context were imputed using the mode for categorical fields and the mean for numerical prior-count features. No duplicate records were present in the dataset.

3.2.3 Feature Engineering

Feature engineering proceeded along two tracks. For the statement text itself, TF-IDF vectorization was applied with unigram and bigram tokenization, producing a sparse matrix capturing the relative importance of word patterns across the corpus. For the metadata, three composite features were constructed: speaker credibility score (derived from the ratio of true to false prior ratings), statement length normalised by topic, and venue formality index (encoding whether the statement was made in a formal legislative context or an informal setting).

The dataset was split into a 70% training set and a 30% hold-out validation set. All models were trained on the training partition using optimized hyperparameters. Performance was evaluated on the validation set using the following metrics:

TP Rate (Sensitivity / Recall): Proportion of actual positive cases correctly classified, computed as the ratio of true positives to total actual positives.

FP Rate: Fraction of actual negatives incorrectly labelled as positives.

Precision: Share of predicted positives that truly belong to the positive class.

F-Measure: Harmonic mean of Precision and Recall, formulated as $F = (2 \times \text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$.

ROC / AUC: Graphical representation of the sensitivity versus specificity trade-off, with AUC quantifying the total area under the ROC curve.

4. Results

The numbers are worth reading carefully. The baseline Random Forest model — no feature selection, no class balancing, just the standard classifier trained on the raw TF-IDF matrix plus metadata — produced a training accuracy of 91.2%, with a cross-validated score of 79.6% and a ROC-AUC of 0.881. Those headline figures look reasonable. But accuracy on a six-class imbalanced problem is a misleading metric. The model was largely learning to predict the two most common veracity categories — false and half-true — and barely touching the minority labels.

Feature importance analysis came next. The top contributors were speaker credibility score, statement-level TF-IDF weights for hedging language ("might", "could", "sources say"), bigram patterns associated with political attack statements, and the venue formality index. Several metadata columns contributed negligible signal. Removing them tightened the model — the CV score moved from 79.6% to 84.1% after feature pruning, while training accuracy dropped slightly to 89.8%. A leaner model that knows what it knows is more useful than a complex one that has learned noise.

The class imbalance issue was severe. The pants-fire label, representing the most egregiously false statements, represented only 8.2% of records. The model trained on the original distribution caught fewer than 12% of pants-fire instances — a recall figure that makes the classification practically useless for the most dangerous category. SMOTE was applied using the imblearn library to generate synthetic minority-class examples until all six veracity labels were represented equally in the training set.

After resampling and retraining, the results improved meaningfully: training accuracy reached 94.6%, validation accuracy moved from 81.3% to 83.7%, CV score held at 84.1%, and overall ROC-AUC reached 0.934. Sensitivity for the pants-fire class rose from 11.8% to 27.4%. Precision and F1 scores improved across all minority labels. Those gains are real — and it is equally honest to acknowledge that recall below 30% on the most extreme label still means the model misses most of what it should catch. SMOTE partially corrects the class imbalance. Better representation in the original data would help more.

5. Conclusion

The problem of misinformation is not going away on its own. The economic incentives that produce fake news are stronger than ever. The platforms that distribute it are too large and too fast for manual review to keep up. And the humans reading it are subject to the same cognitive biases they always were — preferring information that confirms what they already believe, distrusting corrections that arrive too late. Machine learning will not fix any of that on its own. But it can change where the human attention goes.

The Random Forest classifier built in this study achieved a cross-validated accuracy of 84.1% and a ROC-AUC of 0.934 after class balancing and feature selection. SMOTE oversampling raised sensitivity on the most underrepresented veracity categories — the ones that matter most in practice — from below 12% to over 27%. We are not going to present those numbers as a solved problem. A model that still misses roughly 70% of pants-fire statements is a tool that helps, not one that decides.

What the feature importance analysis contributes may be as valuable as the prediction scores themselves. Speaker credibility history was the single strongest predictor — a finding that argues directly for source-level accountability systems, not just content-level filters. Hedging language patterns were the strongest text-level signal, consistent with the linguistic literature on deception. Venue context mattered, too: statements made in formal legislative settings were more frequently truthful than those made at campaign rallies, even controlling for speaker.

The gap between current performance and what would genuinely be deployment-ready points toward specific directions for future research. The LIAR dataset covers political statements specifically — a model trained on it will not generalize well to health misinformation or financial fraud without retraining. Incorporating social propagation features — sharing velocity, cross-platform reach, engagement-to-

follower ratios — would likely push recall higher. And transformer-based architectures, particularly fine-tuned RoBERTa, represent a natural evolution from the Random Forest baseline, offering substantially stronger language representations at the cost of interpretability. The foundation laid here is solid enough to build on.

References

[1] Conroy, N.J., Rubin, V.L. and Chen, Y. Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology, 52(1), pp.1–4, 2015.

[2] Ruchansky, N., Seo, S. and Liu, Y. CSI: A hybrid deep model for fake news detection. Proceedings of the 2017 ACM Conference on Information and Knowledge Management (CIKM), pp.797–806, 2017.

[3] Wang, W.Y. "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), pp.422–426, 2017.

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

[5] Rashkin, H., Choi, E., Jang, J.Y., Volkova, S. and Choi, Y. Truth of varying shades: Analyzing language in fake news and political fact-checking. Proceedings of the 2017 EMNLP Conference, pp.2931–2937, 2017.

[6] Popat, K., Mukherjee, S., Strötgen, J. and Weikum, G. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. Proceedings of WWW 2017 Companion, pp.1003–1012, 2017.

[7] Shu, K., Mahudeswaran, D., Wang, S., Lee, D. and Liu, H. FakeNewsNet: A data repository with news content, social context and spatiotemporal information for fake news detection. Big Data, 8(3), pp.171–188, 2020.

[8] Nakamura, K., Levy, S. and Wang, W.Y. r/Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. Proceedings of LREC 2020, pp.6149–6157, 2020.

[9] Kula, S., Choraś, M. and Kozik, R. Application of the BERT-based architecture in fake news detection. Advances in Intelligent Systems and Computing, Vol. 1267, Springer, pp.239–249, 2020.

[10] Ahmed, H., Traore, I. and Saad, S. Detecting opinion spams and fake news using text classification. Security and Privacy, 1(1), e9, 2018.

[11] LIAR Dataset. Available online: https://www.cs.ucsb.edu/~william/data/liar_dataset.zip