



AI-BASED FAKE NEWS DETECTION USING MACHINE LEARNING

Mrs. Dalavi M.T.¹, Jagruti Nawale²

¹ Faculty Guide & Supervisor | ² Dattakala Group of Institutions, Faculty of Engineering, Pune, Maharashtra, India, India

Corresponding author: jagrutinawale2@gmail.com

ARTICLE ID IJREK-2026-00113	ARTICLE TYPE Research Article	RESEARCH AREA Artificial Intelligence	PUBLICATION Volume 1, Issue Issue 1 (2026)
RECEIVED 19 October 2026	REVISED 21 October 2026	ACCEPTED 20 October 2026	PUBLISHED 24 October 2026

Abstract

Scholarly manuscript entitled 'AI-BASED FAKE NEWS DETECTION USING MACHINE LEARNING'.

Keywords: Research; Technology

I. INTRODUCTION

A. Background / Context Online platforms allow news to spread rapidly across websites, messaging services, and social networks. This improves access to information but also enables fabricated stories, misleading headlines, manipulated context, and unverified claims to reach large audiences. The scale and speed of publication make manual review difficult. Machine learning and natural-language processing can help screen collections of text by identifying linguistic and statistical patterns associated with labelled examples. B. Problem Statement Readers may have difficulty distinguishing credible reporting from fabricated or misleading content, particularly when a headline is emotionally framed or shared without context. Human fact-checking remains important but requires time, specialist knowledge, and access to reliable evidence. Automated classifiers can assist with initial screening, but they may incorrectly label legitimate reporting or fail on new topics. A useful system therefore needs a clear processing pipeline, careful evaluation, transparent limitations, and recommendations to verify important claims through trusted sources. C. Aim and Objectives Develop a text-classification workflow for identifying potentially misleading news content. Preprocess article text and convert it into numerical features suitable for machine learning. Train and compare suitable baseline classification algorithms. Evaluate performance using precision, recall, F1-score, accuracy, and a confusion matrix. Provide a simple interface that explains predictions as automated estimates, not proof of truth or falsehood. D. Significance and Scope The system focuses on text-based screening of news articles or claims using labelled data. It may help users decide which items need closer review, but it does not independently establish factual truth. The first version should document its dataset, class definitions, and evaluation conditions. A source, writing style, or political viewpoint alone should not be treated as evidence that a claim is false.

II. LITERATURE REVIEW / RELATED WORK

A. Overview of Existing Research Research on automated misinformation detection includes linguistic feature analysis, classical machine-learning classifiers, social-context analysis, propagation patterns, and deep-learning models. Early text-classification systems commonly represent documents using word counts or term-frequency and inverse-document-frequency (TF-IDF) features. These representations can be paired with Naive Bayes, Logistic Regression, or Support Vector Machines to establish reproducible baselines. B. Key Research Themes Natural-language processing provides methods for tokenization, normalization, n-gram extraction, and feature representation. Supervised learning uses labelled examples to learn associations between features and class labels. Neural networks and transformer-based language models can capture contextual relationships, but require careful dataset

design, computational resources, and testing for generalization. Evidence retrieval and claim verification are different tasks from classifying article text alone. C. Research Gap A classifier may perform well on a benchmark dataset yet fail on new topics, publishers, languages, or later events. Dataset leakage, duplicate articles, imbalanced classes, and source-specific writing patterns can inflate apparent performance. A practical student system should use a documented split strategy, avoid duplicate leakage, report multiple metrics, and communicate uncertainty instead of presenting its output as a final verdict. D. Contribution of the Proposed Work This paper presents a modular text-classification pipeline with preprocessing, feature extraction, model comparison, evaluation, and prediction display. It emphasizes reproducibility and responsible interpretation. Empirical claims should be added only after the implementation has been tested.

III. PROPOSED SYSTEM

TABLE I

Component	Mainresponsibility
Inputinterface	Accepts headline or article text and explains intended use.
Textvalidation	Checks empty input text length and unsupported formats. . ,
Preprocessing	Normalizes text and applies tokenization and optional stop-word handling.
Featureextraction	Converts text into numerical features such as TF-IDF vectors.
Classifier	Predicts a label using a trained and evaluated model .
Evaluationmodule	Reports metrics and compares models on held-out data.
Resultinterface	Displays prediction and limitations, encouraging independent verification.

Fig. 1. Proposed fake news detection system architecture. C. Main Functional Requirements The system should accept supported text, validate it, apply the same preprocessing used during training, load a trained classifier, and return a readable result. It should handle empty inputs and model failures gracefully. If confidence scores are shown, they should be explained accurately and must not be interpreted automatically as the probability that a claim is true. D. Non-Functional Requirements The system should be reproducible, responsive for typical input sizes, maintainable, and transparent about its dataset and limitations. Model files and dependencies should be versioned. Any stored user submissions should be minimized and protected, especially if the application is deployed publicly.

IV. METHODOLOGY

TABLE II

Field	Exampleformat	Purpose
Title	Textstring	Capture headline language.
Articletext	Text	Provide the main classification input .
Label	Fake / real or dataset-defined classes	Provide the supervised target .
Source/date	Metadata, if available	Support dataset analysis and time-aware evaluation.

V. SYSTEM DESIGN AND IMPLEMENTATION PLAN

TABLE III

Layer	Possible student-level choice
Frontend	HTML , CSS , JavaScript , or a simple React interface.
Backend	Python Flask or FastAPI for model inference.
Textprocessing	scikit-learn TF-IDF vectorizer or equivalent .
Models	Naive Bayes, Logistic Regression, Linear SVM .
Storage	Local files or database for permitted records and model metadata.
Deployment	Local demonstration or approved hosting service.

TABLE IV

Input	Processing	Output
Headlineorarticle	Validationandpreprocessing	Cleantext
Cleantext	TF-IDF transformation	Featurevector
Featurevector	Trainedclassifier	Predicteddatssetclass
Prediction	Displayandlimitationmessage	Screeningresultwithverificationadvice

VI. RESULTS AND EVALUATION PLAN

TABLE V

TestID	Testcase	Expectedresult
T1	Valid news text is submitted .	System returns a prediction using the trained pipeline.
T2	Input is empty.	System requests text instead of crashing.
T3	Input contains unusual punctuation or encoding.	Preprocessing handles text or shows a clear error.
T4	Model file is unavailable.	System reports a safe inference error.
T5	Text differs substantially from training data.	Prediction is shown with limitations.
T6	Duplicate or near-duplicate articles exist .	Duplicates are prevented from leaking across splits.

TABLE VI

Parameter	Manualreview	Keywordrules	MLtextclassifier
Processingscale	Limitedbyreviewertime	Fastbutbrittle	Canprocessmanytextsateftraining
Contexthandling	Humanjudgementandevideince	Limitedpatternmatching	Learnspatternsfromlabelledexamples
Adaptation	Dependsonreviewerexpertise	Rulesneedmanualupdates	Requiresnewdataandretaining
Mainrisk	Humaninconsistencyordelay	Missesparaphrasesandcontext	Datssetbiasandgeneralizationerrors

VII. ADVANTAGES AND LIMITATIONS

TABLE VII

Advantages	Limitations
Can screen text at scale.	Cannot independently prove a claim true or false.
Provides repeatable predictions for a fixed model .	Performance depends on dataset quality and labels.
Baseline models can be computationally efficient .	May learn publisher or topic shortcuts.
Metrics enable comparison between models.	New events and language shifts can reduce accuracy.
Can be integrated into a simple web interface.	Human review and external verification remain important .

VIII. APPLICATIONS

Preliminary screening of news articles for closer review. Media-literacy demonstrations and educational projects. Research into NLP-based misinformation classification. Content moderation support when combined with human review. Newsroom tools that prioritize items for independent verification.

IX. FUTURE SCOPE

Add evidence retrieval and source comparison rather than relying on text style alone. Evaluate multilingual datasets and language-specific preprocessing. Use time-aware splits to test performance on later news events. Explore transformer models when data, compute, and evaluation resources permit. Add calibrated confidence estimates and an uncertain/abstain option if validated. Conduct bias and robustness analysis across topics and publishers.

X. CONCLUSION

This paper proposes an AI-Based Fake News Detection system using machine learning and natural-language processing. The workflow includes dataset preparation, text preprocessing, feature extraction, model training, evaluation, and a user-facing prediction interface. Classical classifiers provide useful baselines, while rigorous splitting and multi-metric evaluation are necessary to avoid misleading performance claims. The system should be treated as an aid for screening, not as an authority that determines truth. Future work can improve generalization through evidence-based verification, time-aware evaluation, multilingual support, and transparent uncertainty handling. The final study should report only measurements obtained from the actual implementation.

REFERENCES

[1] W. Y. Wang, “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, 2017.

[2] 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.

[3] H. Allcott and M. Gentzkow, “Social Media and Fake News in the 2016 Election,” Journal of Economic Perspectives, vol. 31, no. 2, pp. 211–236, 2017.

[4] N. L. 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, pp. 1–6, 2015.

[5] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” arXiv:1301.3781, 2013.

Publication archival note: Official scholarly contribution published under CC BY 4.0 International License. Permanent record on <https://www.ijrek.org> | Article ID: IJREK-2026-00113