ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1352

# Fake News Detection Using Machine Learning and LLM Embeddings: A Comparative Study of TF-IDF and BERT Representations on the Welfake Dataset

Umme Noor Us Saqa[1]; Dr. Sreenivasa B. R.[2]

Visvesvaraya Technological University Jnanasangama Belagavi - 590 018, Karnataka
Submitted in Partial Fulfilment of the Requirements for the Award of the Degree
Master of Technology in Artificial Intelligence and Data Science

[1](USN: 4BD25SAD16), [2]Professor, Department of AI & DS

PG in Artificial Intelligence and Data Science Bapuji Institute of Engineering & Technology
Davanagere - 577004
(Autonomous Institute Affiliated to Visvesvaraya Technological University, Belagavi, and Approved by AICTE, New Delhi, Accredited by NAAC with 'A' Grade)

Bapuji Educational Association ®
Bapuji Institute of Engineering and Technology
Post Box No. : 325, Davanagere -577 004, Karnataka, India.
(Autonomous Institute Affiliated to Visvesvaraya Technological University, Belagavi, and Approved by AICTE, New Delhi, Accredited by NAAC with 'A' Grade)

M. Tech in Artificial Intelligence and Data Science

Publication Date: 2026/08/04

**How to Cite:** Umme Noor Us Saqa; Dr. Sreenivasa B. R. (2026) Fake News Detection Using Machine Learning and LLM Embeddings: A Comparative Study of TF-IDF and BERT Representations on the Welfake Dataset. *International Journal of Innovative Science and Research Technology*, 11(7), 2644-2660. https://doi.org/10.38124/ijisrt/26jul1352

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1352

# CERTIFICATE

This is to certify that the project work entitled " Fake News Detection System Using LLM Embeddings and Machine Learning Classifiers " is a bonafide work carried out by UMME NOOR US SAQA, bearing USN: 4BD25SAD16, in partial fulfilment of the requirements for the award of the degree of Master of Technology in Artificial Intelligence and Data Science of Visvesvaraya Technological University, Belagavi, during the academic year 2025–26. It is certified that all corrections/suggestions indicated for internal assessment have been incorporated in the report. The project report has been approved, as it satisfies the academic requirements in respect of project work prescribed for the said degree.

Project Guide PG Coordinator Program Coordinator

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1352

# DECLARATION

I, UMME NOOR US SAQA (USN: 4BD25SAD16), a student of II Semester M.Tech in Artificial Intelligence and Data Science at Bapuji Institute of Engineering and Technology, Davangere, hereby declare that the project work entitled "Fake News Detection System Using LLM Embeddings and Machine Learning Classifiers" submitted to Visvesvaraya Technological University, Belagavi, during the academic year 2025–26, is an original work carried out by me under the guidance of Dr. Sreenivasa B R, Professor, Department of Artificial Intelligence and Data Science.

This work is submitted in partial fulfilment of the requirements for the award of the degree of Master of Technology in Artificial Intelligence and Data Science. I further declare that this project work has not been submitted, either in part or in full, to any other University or Institution for the award of any degree or diploma.

Date:16/07/2026 UMME NOOR US SAQA

Place: DAVANGERE (USN:4BD25SAD16)

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1352

## ACKNOWLEDGEMENT

We express our sincere gratitude to Bapuji Institute of Engineering and Technology, Davangere, for providing a conducive academic environment, well-qualified faculty, and the necessary infrastructure to carry out this project work.

We are deeply indebted to our guide, Dr. Sreenivasa B R, for his constant guidance, valuable suggestions, and encouragement throughout this project. His support has been instrumental in the successful completion of this work.

We extend our heartfelt thanks to Dr. Nirmala C R, Program Coordinator, Department of Computer Science and Engineering and Dr. Sreenivasa B R, PG Coordinator, Department of Artificial Intelligence and Data Science, for their continuous support, valuable feedback, and motivation.

We also express our sincere thanks to our respected Principal, Dr. H B Aravind, for his encouragement and support.

We are grateful to all the teaching and non-teaching staff of the Department of Computer Science and Design for their assistance and cooperation.

Finally, we would like to thank our family members and friends for their constant encouragement and moral support.

UMME NOOR US SAQA

(USN:4BD25SAD16)

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1352


## ABSTRACT

**The rapid growth of social media and online news platforms has significantly increased the spread of misinformation and fake news. Fake news can influence public opinion, create social unrest, and negatively impact decision-making processes. Therefore, developing intelligent systems capable of automatically identifying fake news has become an important research area in Artificial Intelligence and Natural Language Processing (NLP). The proposed work presents a Fake News Detection System that combines traditional machine learning techniques with Large Language Model (LLM) embeddings to improve classification performance. The proposed system utilizes textual news data from benchmark datasets such as WELFake and LIAR. Initially, the news articles undergo pre-processing steps including text cleaning, normalization, and removal of irrelevant characters. Two different feature extraction approaches are employed: Term Frequency-Inverse Document Frequency (TF-IDF) and contextual embeddings generated using the pre-trained BERT (Bidirectional Encoder Representations from Transformers) model. The extracted features are used to train and evaluate multiple Machine learning classifiers, including Logistic Regression, Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), and XGBoost. Performance evaluation is carried out using standard metrics such as Accuracy, Precision, Recall, F1-Score, and Area Under the ROC Curve (AUC-ROC). Comparative analysis is performed to assess the effectiveness of traditional TF-IDF features versus BERT-based semantic embeddings. Experimental results demonstrate that BERT embeddings capture contextual and semantic information more effectively than conventional text representations, leading to improved fake news classification accuracy. The system also provides visualization tools such as confusion matrices, ROC curves, and performance comparison charts for comprehensive evaluation. The proposed framework highlights the potential of integrating transformer-based language models with classical machine learning algorithms to build robust and scalable fake news detection systems. Our work contributes to combating misinformation by providing an efficient and automated approach for identifying deceptive news content in digital media environments.**


***Keywords:*** *Fake News Detection, Natural Language Processing, Machine Learning, BERT, LLM Embeddings, TF-IDF, Logistic Regression, Support Vector Machine, XGBoost, Deep Learning, Text Classification.*

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1352

# TABLE OF CONTENTS

Contents

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1352

# LIST OF FIGURES

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1352

# LIST OF TABLES

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1352

# LIST OF ABBREVIATIONS

| | |
|---|---|
| AI | Artificial Intelligence |
| ML | Machine Learning |
| RAM | Random Access Memory |
| LLM | Large Language Model |
| BERT | Bidirectional Encoder Representations from Transformers |
| CLS | Classification Token |
| SEP | Separator Token |
| RoBERTa | Robustly Optimized BERT Approach |

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1352

# CHAPTER ONE
# INTRODUCTION

The rapid expansion of digital communication platforms, social media networks, and online news portals has transformed how information is created, shared, and consumed. While these platforms provide instant access to news, they have also facilitated the widespread dissemination of fake news — intentionally fabricated or misleading information presented as legitimate news to influence public opinion, generate revenue, or create confusion among readers.

The sheer volume of online content makes manual verification impractical, motivating automated fake news detection as an active research area at the intersection of Artificial Intelligence, Machine Learning, and Natural Language Processing. This work proposes a fake news detection system that leverages both traditional machine learning techniques and modern LLM embeddings. TF-IDF is used for baseline feature extraction, while BERT embeddings are used to capture contextual and semantic information. Logistic Regression, SVM, MLP, and XGBoost are then trained on each representation to classify articles as Real or Fake, followed by a comparative analysis of the two feature families.

## *Problem Statement*

The proliferation of fake news on digital platforms poses significant challenges to society by spreading misinformation, influencing public opinion, and undermining trust in legitimate news sources. Existing detection methods primarily rely on traditional text representation techniques that fail to fully capture the semantic and contextual meaning of textual content. There is a need for an intelligent, automated system capable of accurately identifying fake news articles using advanced NLP techniques and machine learning models, while maintaining scalability and efficiency.

## *Objectives*

- Pre-process and clean news article datasets by removing noise, URLs, special characters, and irrelevant information.
- Extract meaningful features from news articles using TF-IDF and BERT embeddings.
- Implement and evaluate Logistic Regression, SVM, MLP, and XGBoost classifiers for fake news classification.
- Compare TF-IDF and BERT-based representations using Accuracy, Precision, Recall, F1-Score, and AUC-ROC to identify the most effective approach.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1352

# CHAPTER TWO
# RELATED WORK

Shu et al. presented one of the early comprehensive studies on fake news detection, emphasizing the value of combining content-based analysis with social context information to improve detection accuracy [1]. Rubin, Chen, and Conroy investigated linguistic features in fake news identification, showing that deceptive articles often exhibit distinct writing styles that NLP techniques can capture [2].

Ahmed, Traore, and Saad proposed a framework based on TF-IDF features with Naïve Bayes, SVM, and Logistic Regression classifiers, demonstrating that TF-IDF provides an effective baseline for fake news detection [3]. Wang introduced the LIAR dataset, a benchmark of short political statements labelled by degree of truthfulness, which has since become a widely used resource in this field [4].

Devlin et al. introduced BERT, which generates contextual embeddings using bidirectional attention and achieved state-of-the-art results across multiple NLP tasks [5]. Building on this, Kaliyar, Goswami, and Narang proposed FakeBERT, combining BERT embeddings with deep learning architectures and reporting significant accuracy improvements over traditional machine learning approaches [6]. Chen and Guestrin introduced XGBoost, an optimized gradient boosting framework noted for handling complex feature interactions effectively, making it well suited to fake news detection when combined with textual embeddings [7].

Collectively, this literature suggests that transformer-based embeddings such as BERT can provide richer semantic representations than conventional TF-IDF features, while classical machine learning classifiers remain effective and efficient when trained on high-quality feature representations.

More recent work published in 2026 has continued to probe this trade-off on the WELFake dataset specifically. Hegde et al. combined fine-tuned DistilBERT embeddings with an XGBoost classifier on the full WELFake corpus and reported accuracy above 99%, indicating that transformer embeddings can substantially outperform TF-IDF when the embedding model is fine-tuned on the target domain rather than used in a frozen, off-the-shelf form [8]. In parallel, Martsenyuk and Smailova, and a related WELFake-based study using Logistic Regression, both revisited TF-IDF as an interpretable, computationally lightweight baseline, reinforcing those well-tuned statistical features remain competitive even as contextual embeddings mature [9–10]. Taken together, this newer evidence supports the interpretation adopted in this paper's results: that the weak BERT performance observed here is best attributed to how the embeddings were generated and pooled, rather than to a fundamental limitation of contextual representations.

- *Research Gaps*

Most traditional approaches rely on Bag-of-Words or TF-IDF representations that focus on word frequency and often fail to capture contextual meaning, limiting their ability to detect sophisticated misinformation. Many existing studies evaluate only a single classifier or a single feature representation in isolation, without a systematic comparison between conventional and transformer-based approaches. High-accuracy systems in the literature also frequently demand heavy computational resources, limiting practical deployment, and often lack multi-metric evaluation (Accuracy, Precision, Recall, F1-Score, AUC-ROC) alongside visual diagnostics such as confusion matrices and ROC curves.

This work addresses these gaps by combining TF-IDF and BERT embeddings with four classifiers — Logistic Regression, SVM, MLP, and XGBoost — under a single, consistent evaluation pipeline with comprehensive metrics and visualizations.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1352

# CHAPTER THREE
# METHODOLOGY

The proposed system combines traditional NLP techniques with a transformer-based language model to classify news articles as real or fake. The pipeline begins with loading the WELFake dataset and pre-processing the text through lowercasing, URL removal, special-character elimination, and whitespace normalization. The cleaned data is split into training and testing sets.

➢ *Dataset*

The WELFake dataset is used as the primary data source. It is a benchmark dataset of approximately 72,000 news articles, each labelled as Real News (0) or Fake News (1), with two core attributes: the article text and its label.

➢ *Feature Extraction*

Two feature extraction techniques are employed. TF-IDF (Term Frequency–Inverse Document Frequency) converts textual content into numerical vectors based on word importance across the corpus. BERT (bert-base-uncased) embeddings generate 768-dimensional contextual representations of each article using the CLS token, capturing relationships between words rather than relying purely on frequency.

➢ *Classification Models*

- Logistic Regression — a linear classification model that provides probability estimates and trains quickly.
- Linear Support Vector Machine (SVM) — finds the optimal separating hyperplane; effective for high-dimensional text classification.
- Multi-Layer Perceptron (MLP) — an artificial neural network capable of learning non-linear patterns.
- XGBoost — a gradient boosting framework known for high predictive performance and handling complex feature interactions.

➢ *Evaluation Metrics*

Trained models are evaluated on held-out test data using Accuracy (overall correctness), Precision (correctness of fake-news predictions), Recall (ability to detect fake news), F1-Score (harmonic mean of Precision and Recall), and AUC-ROC (classification capability across decision thresholds). Confusion matrices, ROC curves, and classifier comparison charts are generated to support analysis. The final model is also used for inference on new, unseen articles.

➢ *System Architecture*

Figure 1 illustrates the end-to-end workflow: the WELFake dataset is pre-processed, transformed into TF-IDF and BERT feature representations, passed to the four classifiers, and evaluated through the performance metrics described above.

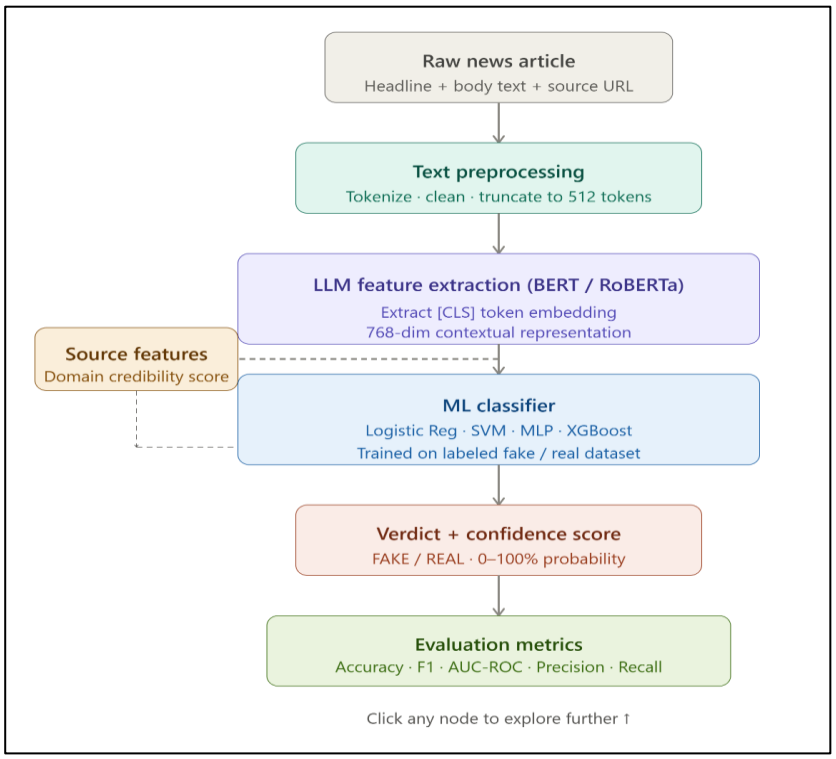


Fig 1 Proposed Framework for Fake News Detection Using LLM Embeddings and Machine Learning Classifiers.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1352

- *Experimental Setup*

Experiments were conducted in Visual Studio Code on a Windows 11 system equipped with a 12th Gen Intel Core i5-12450H processor, 16 GB RAM, and an NVIDIA GeForce GTX 1650 GPU. The software stack included Python 3.10+, Scikit-learn, PyTorch, Hugging Face Transformers, XGBoost, Pandas, NumPy, and Matplotlib/Seaborn for visualization.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1352

# CHAPTER FOUR
# RESULTS AND DISCUSSION

The system was evaluated on the WELFake dataset using TF-IDF and BERT feature representations across all four classifiers. Table 1 summarizes Accuracy and AUC-ROC for each feature-classifier combination.

Table 1 Performance Comparison of Classifiers on TF-IDF and BERT Features.

| Feature Extraction | Classifier | Accuracy (%) | AUC-ROC (%) |
|---|---|---|---|
| TF-IDF | Logistic Regression | 92.0 | 97.36 |
| TF-IDF | Linear SVM | 94.7 | 98.66 |
| TF-IDF | MLP | 94.2 | 98.13 |
| **TF-IDF** | **XGBoost** | **94.9** | **98.99** |
| BERT | Logistic Regression | 49.0 | 47.32 |
| BERT | Linear SVM | 50.5 | 47.27 |
| BERT | MLP | 44.5 | 43.39 |
| BERT | XGBoost | 47.9 | 47.50 |

The results show that TF-IDF features substantially outperform BERT embeddings in this implementation. Among all models, XGBoost with TF-IDF achieved the highest accuracy (94.9%) and AUC-ROC (98.99%), making it the best-performing configuration overall, with Linear SVM close behind at 94.7% accuracy and 98.66% AUC-ROC. In contrast, models trained on BERT embeddings produced accuracies between 44% and 51% — close to random guessing — indicating that the BERT embeddings generated in this pipeline were not sufficiently informative for downstream classification, likely due to how the embeddings were pooled or fed into the classifiers rather than a limitation of BERT itself.

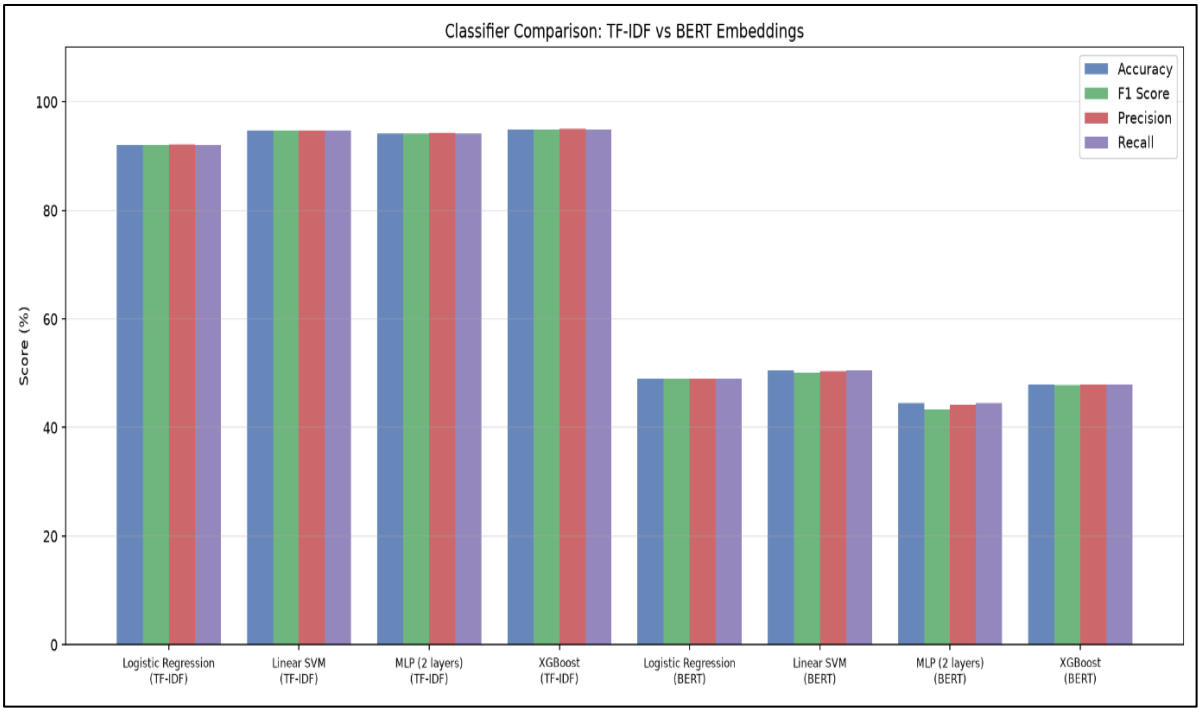


Fig 2 Accuracy, Precision, Recall, and F1-Score Comparison Across Classifiers.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1352

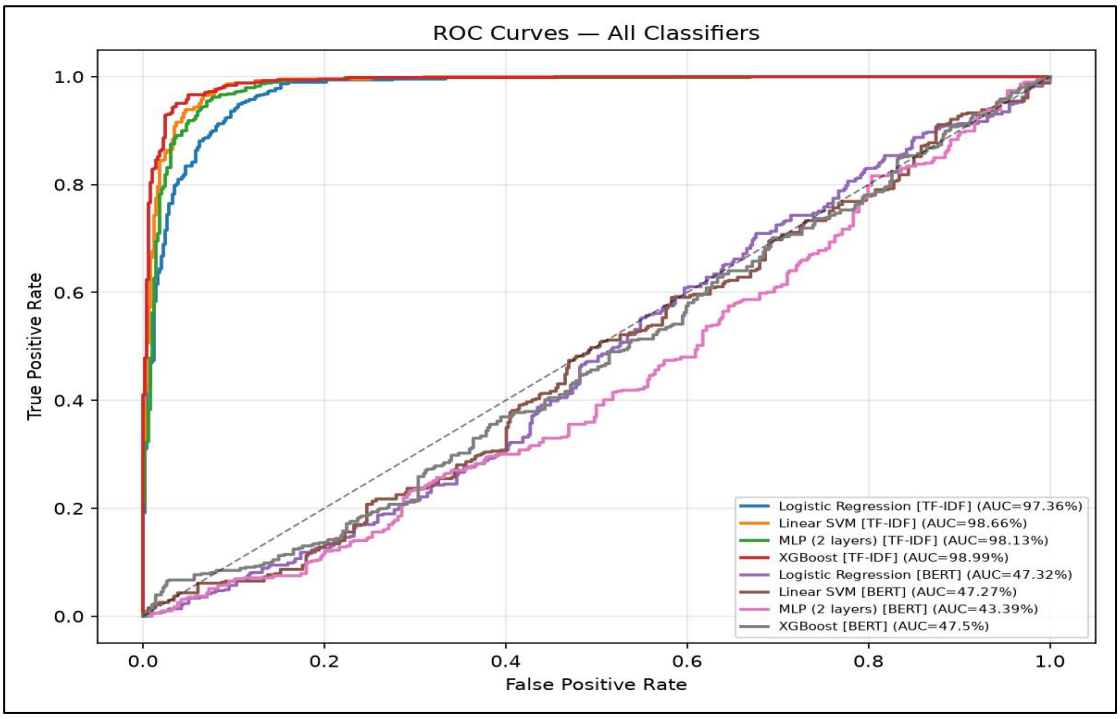


Fig 3 ROC Curves of Classifiers on TF-IDF and BERT Features.

The ROC curves for TF-IDF-based classifiers sit near the upper-left corner of the plot, indicating a high true-positive rate combined with a low false-positive rate. XGBoost again leads with an AUC of 98.99%, followed by Linear SVM (98.66%), MLP (98.13%), and Logistic Regression (97.36%) — confirming that TF-IDF features capture strongly discriminative textual patterns in the WELFake dataset. The BERT-based curves, by comparison, lie close to or below the diagonal reference line, with AUC values between 43.39% and 47.50%, below the 50% random-classifier threshold, reinforcing that these models were unable to meaningfully separate real from fake articles under the current embedding setup.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1352

# CHAPTER FIVE
# CONCLUSION AND FUTURE SCOPE

This work developed a fake news detection system on the WELFake dataset using a complete pipeline of pre-processing, feature extraction, classification, and evaluation, comparing TF-IDF and BERT embeddings across four classifiers. TF-IDF-based models significantly outperformed BERT-based models in this implementation: XGBoost with TF-IDF achieved the best results, with roughly 94.9% accuracy and a 98.99% AUC-ROC, confirmed by confusion matrix and ROC curve analysis. The findings show that well-engineered traditional text representations, combined with strong classifiers, can deliver highly accurate, computationally efficient fake news detection, and can serve as a practical tool for reducing the spread of misinformation on digital platforms.

Several directions can extend this work. Fine-tuning transformer models such as BERT, RoBERTa, DistilBERT, or DeBERTa — rather than using frozen embeddings — could substantially improve contextual understanding and classification accuracy. Future systems could incorporate images, video, and metadata for multimodal fake news detection, and integrate with web applications, browser extensions, or social media platforms for real-time verification. Additional features such as publisher reputation, domain credibility, and user-engagement metrics could support source credibility analysis, and multilingual support would extend applicability to global contexts. Finally, deep learning and ensemble approaches could be explored to further improve robustness and predictive accuracy.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul1352

## REFERENCES

[1]. R. Yang, Y. Bian, W. Gao, B.-W. Zhang, J. Ma, H. Lin, Z. Luo, X. Zhu, and X.-C. Yin, "Dialectical Structured Reasoning for Explainable Multimodal Fake News Detection," in *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 31998–32013, Jul. 2026, doi: 10.18653/v1/2026.findings-acl.1601.

[2]. Y. Wang, Y. Yang, and H. Zhang, "CSI: An Investigative Multi-Agent Framework for Explainable Short Video Fake News Detection," in *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 25508–25528, Jul. 2026, doi: 10.18653/v1/2026.findings-acl.1274.

[3]. Y. Zhang, H. Liu, W. Chen, and J. Zhao, "Deconstruct, Diagnose, and Deliberate: A Protocol-Adaptive Role-Specific Multi-Agent Framework for Fake News Detection," in *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 24516–24534, Jul. 2026, doi: 10.18653/v1/2026.findings-acl.1227.

[4]. L. Wu, X. Jiang, S. Sun, Y. Lei, T. Wen, Y. Wang, and M. Liu, "ZoFia: Zero-Shot Fake News Detection with Entity-Guided Retrieval and Multi-LLM Interaction," in *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 21540–21556, Jul. 2026, doi: 10.18653/v1/2026.findings-acl.1083.

[5]. C. Xu, C. Jin, Y. Niu, N. Yan, Y. Mei, S. Guan, L. Chen, and M.-T. Kechadi, "LiveFact: A Dynamic, Time-Aware Benchmark for LLM-Driven Fake News Detection," in *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026)*, pp. 11881–11910, Jul. 2026, doi: 10.18653/v1/2026.acl-long.546.

[6]. A. A. Sani, K. A. Zaoad, S. E. S. Adib, M. A. Muqtadir, and A. Abrar, "Addressing Data Scarcity in Bangla Fake News Detection: An LLM-Based Dataset Augmentation Approach," *arXiv preprint arXiv:2605.01292*, May 2026.

[7]. X. Li, Y. Zhang, and E. C. Malthouse, "Large Language Model Agentic Approach to Fact Checking and Fake News Detection," *Frontiers in Artificial Intelligence and Applications*, Apr. 2026, doi: 10.3233/FAIA240787.

[8]. C. Xu, C. Jin, Y. Niu, N. Yan, Y. Mei, S. Guan, L. Chen, and M.-T. Kechadi, "LiveFact: A Dynamic, Time-Aware Benchmark for LLM-Driven Fake News Detection," *arXiv preprint arXiv:2604.04815*, Apr. 2026.

[9]. D. Mladenić et al., "Fake News Detection Through LLM-Driven Text Augmentation Across Media and Languages," *Machine Learning and Knowledge Extraction*, vol. 8, no. 4, Apr. 2026.

[10]. J. Dang, Y. Sun, and C. Yu, "Fusing Non-Textual Cues with Classical NLP for Enhanced Multimodal Fake News Spread Detection," *Scientific Reports*, vol. 16, Art. no. 16193, Apr. 2026.