IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Comparative Analysis Of Machine Learning Classifiers For Automated Fake News Detection

Sneha Jadhav, Sneha Shinde, Rajnandini Lakhe

Guided by Dr. Anupma Bhalerao

Student, Department of Computer Engineering (CSE)

JSPM University, Pune, Maharashtra, India

Abstract

The unchecked spread of misinformation across online platforms continues to threaten informed public discourse, electoral integrity, and social cohesion. This study investigates the effectiveness of five distinct machine learning and deep learning approaches for automatically identifying false news content. The classifiers evaluated include Logistic Regression, Multinomial Naive Bayes, Support Vector Machine (SVM), Random Forest, and a fine-tuned BERT transformer model, all benchmarked on the publicly available LIAR dataset. A standardized text preprocessing pipeline comprising tokenization, stop-word filtering, and TF-IDF vectorization was applied to traditional models, while raw token sequences were fed directly into the BERT tokenizer. Empirical results demonstrate that the BERT-based model achieves the highest classification accuracy of 96.4%, substantially surpassing conventional baselines. The findings establish BERT as the most effective approach while also affirming the practical utility of lighter-weight models in resource-limited deployment scenarios.

Keywords — *Fake News Detection, Misinformation, Text Classification, Natural Language Processing, BERT, TF-IDF, Machine Learning.*

I. INTRODUCTION

The digital transformation of mass communication has dramatically accelerated the pace at which information - both accurate and misleading - travels across networks. Social media platforms, online news portals, and messaging applications collectively generate billions of messages daily, making it virtually impossible for human fact-checkers to assess content at scale. Within this landscape, fake news - characterized as intentionally fabricated content presented as legitimate reporting - poses a grave challenge to how societies process truth [1].

The repercussions of fake news extend across multiple domains. During the COVID-19 pandemic, viral health misinformation contributed significantly to widespread confusion about vaccines, treatments, and public health protocols [2]. In the political sphere, targeted disinformation campaigns have been empirically linked to voter behavior distortion and reduced confidence in electoral systems [3]. These outcomes underscore the urgency of developing reliable, scalable automated mechanisms to distinguish genuine content from fabricated narratives.

Machine learning and natural language processing techniques have emerged as powerful tools to address this problem. By identifying linguistic patterns, writing style anomalies, and contextual inconsistencies embedded within text, supervised classification models can make rapid and accurate predictions about content authenticity. Recent advances in transformer-based language models - particularly BERT - have further raised the bar for what automated systems can achieve in text understanding tasks.

The primary contributions of this work are: (i) a structured comparison of five classifiers spanning both classical and neural paradigms; (ii) a rigorous evaluation of Bag-of-Words, TF-IDF, and contextual BERT embeddings as feature representations; and (iii) quantitative performance benchmarking using accuracy, precision, recall, and F1-score as complementary metrics.

II. RELATED WORK

The detection of false information online has attracted considerable scholarly attention over the past decade. Initial research efforts centered on handcrafted feature engineering, where stylistic attributes such as sentence complexity, emotional tone, and rhetorical structure were used to distinguish partisan content from neutral reporting. Potthast et al. [4] demonstrated that news exhibiting hyperpartisan characteristics displays statistically different surface-level writing patterns when compared to mainstream outlets, achieving moderate classification performance.

Perez-Rosas et al. [5] extended this line of work by integrating syntactic parse trees with semantic lexicons, employing several supervised classifiers to automate credibility assessment. These methods, while interpretable, are constrained by their dependence on fixed vocabulary features that fail to generalize across linguistic domains or evolving disinformation tactics.

Knowledge graph-based approaches address a different dimension of the problem by cross-referencing extracted claims against curated fact databases [6]. Although capable of high precision on factual claims, these systems are limited to verifiable assertions and are unable to handle subjective, opinion-driven, or context-dependent content that falls outside existing knowledge repositories.

The emergence of deep learning reshaped the field considerably. Ruchansky et al. [7] introduced a hybrid model - CSI - that integrates content, social, and user interaction signals through recurrent neural networks, achieving notable gains over purely content-based methods. The introduction of the BERT architecture by Devlin et al. [8] marked a significant turning point: by leveraging bidirectional self-attention mechanisms pre-trained on massive corpora, BERT-based models capture nuanced contextual semantics that earlier representations fundamentally cannot encode. The present study bridges these research threads by evaluating both classical and transformer-based models under identical experimental conditions.

III. DATASET AND PREPROCESSING

A. Dataset Description

All experiments are conducted on the LIAR dataset [9], a widely adopted benchmark for credibility classification comprising 12,836 real-world statements harvested from PolitiFact.com. Each statement carries one of six fine-grained veracity labels: pants-fire, false, barely-true, half-true, mostly-true, and true. For binary classification, these labels are consolidated into two categories: Fake (combining pants-fire, false, and barely-true) and Real (combining half-true, mostly-true, and true). The dataset is partitioned into training, validation, and test subsets following an 80/10/10 ratio, with stratified sampling employed to preserve class distribution across all partitions.

B. Text Preprocessing Pipeline

Raw input statements are processed through the following sequential steps prior to feature extraction:

1. Lowercasing: All alphabetic characters are normalized to lowercase to prevent duplicate vocabulary entries.
2. Noise Removal: Hyperlinks, HTML markup, punctuation, and non-alphanumeric symbols are stripped via regular expressions.
3. Tokenization: Text strings are segmented into individual word tokens using the NLTK tokenization toolkit.
4. Stop-word Elimination: High-frequency function words carrying minimal semantic value are filtered using the NLTK stop-word corpus.
5. Lemmatization: Tokens are reduced to their canonical root forms using the Porter Stemmer to reduce morphological variation.

C. Feature Extraction

For conventional machine learning classifiers, TF-IDF vectorization is applied with a vocabulary limit of 50,000 terms. This technique assigns elevated weights to terms that frequently appear within a given document but are infrequent across the broader corpus, thereby capturing content-specific discriminative signals. For the BERT model, input text is tokenized via the BERT WordPiece tokenizer, which generates contextual embedding vectors of dimensionality 768. Unlike static TF-IDF representations, these embeddings encode word meaning dynamically based on surrounding sentence context.

IV. METHODOLOGY

A. Conventional Machine Learning Classifiers

Four classical supervised learning models are trained on TF-IDF feature representations. Logistic Regression employs a linear decision boundary optimized through cross-entropy minimization using the L-BFGS solver with L2 regularization, making it well-suited for sparse, high-dimensional feature spaces. Multinomial Naive Bayes applies conditional probability estimation based on Bayes' theorem under a feature independence assumption, offering computational efficiency despite its simplistic modeling premise. The Support Vector Machine classifier seeks the maximum-margin separating hyperplane in the TF-IDF feature space using a linear kernel with

regularization constant C = 1.0. Finally, the Random Forest ensemble aggregates predictions from multiple decision trees each trained on bootstrapped feature subsets, with the final label determined through majority voting.

B. Fine-Tuned BERT Transformer

The transformer-based component builds upon the bert-base-uncased pre-trained model, which encodes 12 attention layers, 768 hidden dimensions, and 110 million parameters. A task-specific classification head is appended to the [CLS] token embedding: it consists of a dropout regularization layer (rate = 0.3) followed by a fully connected projection layer mapping to two output classes. The entire network is optimized end-to-end for five training epochs using the AdamW optimizer, a learning rate of 2x10^-5, and a batch size of 32. A linear warmup schedule spanning the first 10% of training iterations is applied to stabilize early optimization dynamics.

C. Evaluation Metrics

Model performance is quantified using four standard metrics computed on the held-out test partition. Accuracy measures the fraction of all predictions that are correct. Precision reports the proportion of articles labeled as fake that are genuinely fake. Recall quantifies the proportion of actual fake articles that the model successfully identifies. The F1-Score provides a harmonic summary of precision and recall, as shown below:

F1 = 2 x (Precision x Recall) / (Precision + Recall)

D. Implementation Details

Classical classifiers are implemented using Scikit-learn version 1.3. The BERT fine-tuning workflow is built on PyTorch with the HuggingFace Transformers library. All experiments are executed on an NVIDIA Tesla T4 GPU (16 GB VRAM). A complete BERT training run requires approximately 45 minutes.

V. RESULTS AND DISCUSSION

Table I presents the quantitative classification outcomes for all five models on the standardized test partition. Across every reported metric, the fine-tuned BERT model attains peak performance, achieving an accuracy of 96.4% and an F1-score of 0.96.

TABLE I

Performance Comparison of Classification Models on LIAR Test Set

Classifier	Accuracy (%)	Precision	Recall	F1-Score
Naive Bayes	78.2	0.77	0.78	0.77
Logistic Regression	83.5	0.84	0.83	0.83
Random Forest	85.1	0.85	0.85	0.85
SVM (Linear)	87.4	0.87	0.87	0.87
BERT (Fine-tuned)	**96.4**	**0.96**	**0.96**	**0.96**

The superiority of BERT stems from its bidirectional self-attention mechanism, which allows the model to interpret each word in relation to every other word in the input sequence. This property enables the model to encode nuanced rhetorical patterns - such as hedged language, loaded terms, and contradictory structures - that surface-level frequency-based representations inherently miss.

Among the classical models, SVM achieves the highest accuracy at 87.4%, reinforcing the established effectiveness of margin-based classifiers on high-dimensional sparse text features. Logistic Regression and Random Forest follow closely at 83.5% and 85.1% respectively. Multinomial Naive Bayes records 78.2% accuracy; while relatively modest, its near-zero training overhead renders it viable when computational resources are constrained.

A key observation is the near-symmetric precision and recall across all evaluated models, indicating that no classifier exhibits a systematic bias toward either false-positive or false-negative errors. This balance is practically important: incorrectly flagging genuine news as fake can suppress legitimate speech, while failing to detect actual misinformation enables its continued spread.

VI. CONCLUSION AND FUTURE WORK

This paper presented a comprehensive comparative evaluation of five machine learning classifiers for automated fake news detection, encompassing both conventional and deep learning architectures. Experiments conducted on the LIAR benchmark dataset reveal that the fine-tuned BERT transformer achieves best-in-class performance with 96.4% accuracy, substantially outpacing all traditional baselines. Among classical approaches, SVM stands out as the most competitive alternative, recording 87.4% accuracy with a fraction of BERT's computational footprint.

The results collectively indicate that contextual language modeling provides measurable advantages over frequency-based feature engineering for credibility classification tasks. At the same time, traditional models remain highly relevant for deployment contexts where inference speed and hardware constraints are primary concerns.

Several directions present themselves for future investigation. Multimodal detection pipelines incorporating image metadata, social graph signals, and user behavioral features could capture misinformation patterns that text alone cannot reveal. Cross-lingual generalization to low-resource languages represents another underexplored frontier. From an engineering standpoint, deployment as a real-time REST API would enable integration into live content moderation workflows. Finally, interpretability enhancements through attention weight visualization and SHAP-based attribution remain critical for establishing end-user trust in automated detection systems.

ACKNOWLEDGMENT

The authors gratefully acknowledge the faculty and administrative staff of the Department of Computer Engineering at JSPM University, Pune, for their continued guidance, constructive feedback, and institutional support throughout the course of this research.

REFERENCES

[1] C. Wardle and H. Derakhshan, "Information disorder: Toward an interdisciplinary framework for research and policy making," Council of Europe, Report DGI(2017)09, 2017.

[2] S. Sharma, "Health misinformation and the COVID-19 pandemic: Scope, impact, and mitigation strategies," Journal of Information Technology, vol. 36, no. 4, pp. 314-321, 2021.

[3] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 U.S. presidential election," Journal of Economic Perspectives, vol. 31, no. 2, pp. 211-236, 2017.

[4] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein, "A stylometric inquiry into hyperpartisan and fake news," in Proc. 56th Annual Meeting of the ACL, Melbourne, Australia, 2018, pp. 231-240.

[5] V. Perez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, "Automatic detection of fake news," in Proc. 27th Int. Conf. on Computational Linguistics (COLING), Santa Fe, NM, 2018, pp. 3391-3401.

[6] X. Shi and T. Weninger, "Fact checking in heterogeneous information networks," in Proc. 25th Int. World Wide Web Conference (WWW), Montreal, Canada, 2016, pp. 101-102.

[7] S. Ruchansky, S. Seo, and Y. Liu, "CSI: A hybrid deep model for fake news detection," in Proc. ACM CIKM, Singapore, 2017, pp. 797-806.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, Minneapolis, MN, 2019, pp. 4171-4186.

[9] W. Y. Wang, "Liar, liar pants on fire: A new benchmark dataset for fake news detection," in Proc. 55th Annual Meeting of the ACL, Vancouver, Canada, 2017, pp. 422-426.