STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **16**, July 2026, pp 116–131.
Published online in International Academic Press (www.IAPress.org)

Trade-offs between Accuracy and Efficiency in Fake News Detection: A Comprehensive Study of Lightweight and Deep Learning Techniques

Ahmed Abdelhafeez[1,2] Tareef S. Alkellezli[3], Moshira A. Ebrahim[4,*]

[1] *Faculty of Computers and Information Technology, Innovation University, Cairo, Egypt*
Ahmed.abdelhafeez@iu.edu.eg [2] *Applied Science Research Center. Applied Science Private University, Amman, Jordan*
[3] *Cyber Security Engineering Department, College of Engineering, Ashur University, Baghdad.Iraq*
tareef.alkellezli@au.edu.iq
[4] *Computer Engineering and Information Technology Department, Modern Academy for Engineering and Technology, Cairo, Egypt*

Abstract The fast spread of fake news on digital platforms needs effective and scalable fake news identification technologies. While deep learning models are very accurate, their computational cost restricts their use in limited-resource scenarios. Using FakeNewsNet dataset, this study provides an evaluation of a lightweight TF-IDF and logistic regression framework against SVM, LSTM, and BERT baseline models across headline-only and full-text scenarios. Experimental results show that the lightweight TF-IDF+LR model achieves competitive accuracy (78.8% on GossipCop and 83.7% on PolitiFact) with dramatic efficiency gains, with 20,000 times faster training than BERT, 0.55 ms inference time, and much smaller memory usage than BERT.
An error analysis demonstrates that lexical representations struggle to capture semantic and contextual nuances. It also highlights TF-IDF paired with logistic regression as an acceptable baseline for detecting fake news while outlining its performance limitations. Also, a practical decision matrix is provided for model selection based on the environment's primary constraints.

Keywords Fake news detection, TF-IDF, Logistic regression, computational efficiency, Resource-constrained systems

DOI: 10.19139/soic-2310-5070-3623

1. Introduction

Social media's explosive growth has completely changed how information is created, shared, and consumed. Instantaneous worldwide communication is made possible by these platforms, but they have also made it easier for false information and fake news to spread widely, compromising democratic processes, public confidence, and social stability [1]. Fake news poses significant risks to democratic processes, , influencing election results, public health, and social cohesion [2, 3].

Thus, one of the most important research challenges in the domains of machine learning and natural language processing is the ability to automatically identify fake news.

Deep learning (DL) algorithms, such as recurrent neural networks (RNNs)[4], gated recurrent units (GRUs)[5], long short-term memory networks (LSTMs)[6], and transformer-based models like BERT[7], have played a major role in recent developments in fake news identification. These methods frequently achieve outstanding performance across benchmark datasets as they represent deep semantic linkages, contextual dependencies, and long-range language patterns. Nevertheless, these performance improvements come at the expense of

*Correspondence to: Moshira A. Ebrahim (Email: mushira.ibrahim@eng.modern-academy.edu.eg). Computer Engineering and Information Technology Department, Modern Academy for Engineering and Technology, Cairo, Egypt.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

significant computational demands, such as higher inference delay, lengthy training cycles, and high memory usage. These limitations restrict the usefulness of deep models in large-scale pipelines, real-time systems, and resource-constrained settings like mobile devices and edge computing platforms.

Concurrently, a computationally efficient solution is provided by traditional machine learning techniques based on statistical text representations can perform competitively in structured text categorization challenges despite their simplicity [8]. However, their performance in complicated fake news detection scenarios—particularly in contrast to deep learning models under varied input conditions—remains little studied [9].
The optimized feature selection procedures can enhance text classification performance, so the precisely constructed feature spaces, such as Term Frequency–Inverse Document Frequency (TF-IDF), when paired with linear classifiers like Logistic Regression (LR). These models are far less computationally demanding, lightweight, comprehensive, and are more efficient and increase short-text classification accuracy [10] [11].

This paper offers a thorough comparison of a traditional TF-IDF with logistic regression (TF-IDF+LR) pipeline to typical deep learning architectures, with a particular emphasis on resource-constrained false news detection. The major goal is to quantify the conflicts between prediction performance, such as accuracy, precision, recall, F1-score and computational efficiency such as memory use, training time, and inference latency.

The assessment of models under various input configurations is a crucial component of this research. Due to latency and resource limitations, many real-world applications, including real-time news filtering, rely on succinct textual inputs like headlines, even though deep learning models are usually built to utilize full-text information. In order to address this, we examine two experimental scenarios: (1) a full-text scenario that permits a fair comparison by enabling deep models to use their entire representational potential, and (2) a lightweight situation where all models are trained and assessed using only headlines. The relationship between input richness and model complexity is better understood thanks to this dual evaluation paradigm.

Furthermore, this work presents a comprehensive empirical analysis into fake news detection approaches, with an emphasis on the trade-off between detection accuracy and computational efficiency. Rather than presenting a new algorithm, this paper seeks to provide helpful insights regarding selecting a detection model under different computing limitations. The main contributions of this work can be summarized as follows.

- A comprehensive comparison between traditional machine learning and deep learning models for detecting fake news.
- Systematic study of the trade-off between accuracy and computational efficiency.
- A practical decision approach to guide model selection based on resource availability.
- Statistical significance testing to ensure the reliability of the reported results.
- An in-depth error analysis to identify common failure cases and limitations of the evaluated models.

The rest of this paper is organized as follows. Section 2 examines related work in fake news detection. Section 3 describes the problem and discusses computation challenges in DL models. Section 4 provides the lightweight TF-IDF with logistic regression methodology. Section 5 shows experimental results under various evaluation settings. Section 6 provides a comprehensive error analysis and discussion of results. Finally, Section 7 summarizes the paper and outlines directions for future research.

2. Related Work

Fake news detection strategies are divided into three types: content-based analysis, propagation-based analysis, and source-based authenticity evaluation [12].

Early content-based techniques used handmade language characteristics and conventional classifiers. Dadgar et al. [13] used TF-IDF vectorization and Support Vector Machines for news categorization, with decent effectiveness on restricted datasets. Perez-Rosas et al. [14] created automated detection algorithms that incorporate lexical,

syntactic, and semantic aspects. Rubin et al. [15] divided fake news into three categories, large-scale hoaxes, hilarious fake news, and serious fabrication, and developed TF-IDF+LR specialized detection algorithms for each category.
Horne and Adali [16] revealed different linguistic features of fake news, such as greater article length, more usage of capitalized words, and lower stop word frequency than authentic journalism. Carvajal Builes [17] demonstrated that misleading tales have less semantic sophistication, shorter sentences, and a greater use of negative emotional and motion-related language.

The development of deep learning (DL) has greatly improved detection skills. Ruchansky et al. [18] introduced CSI (Capture, Score, Integrate), a hybrid DL architecture that combines news material and user comments using LSTM-based document embeddings. Yang et al.[19] presented Hierarchical Attention Networks (HAN) for document categorization, which use word- and sentence-level attention methods. Long et al. [20] created HPA-BLSTM, which combines hypothalamic-pituitary-adrenocortical processes with bidirectional LSTM designs. Albahar [5] suggested a hybrid GRU-SVM model that encodes user comments and news content using bidirectional GRUs, followed by SVM classification. Despite its high performance, this technique necessitates significant processing resources for GRU state management and SVM kernel calculations.
Recent research has challenged traditional approaches for resource-constrained contexts. Ahmed et al. [21] showed that combining n-gram analysis with classification techniques yields competitive results for opinion spam identification. Gilda [22] tested several traditional techniques, including decision trees, random forests, and stochastic gradient descent, and demonstrated that lightweight approaches can match DL performance on selected datasets. Furthermore, Al-Heresh [23] emphasized the need of energy-efficient supervised learning models for real-time news verification, which are compatible with lightweight linear classifiers like logistic regression. From a larger technical standpoint, Al-Khalidi and Al-Sarayreh [24] emphasized the societal importance of accurate disinformation detection systems, particularly in regional digital ecosystems. Collectively, this research encourages the use of a scalable TF-IDF feature extraction framework paired with logistic regression, which provides a combination of classification performance, interpretability, and computational economy. Saragih et al. [25] suggested a hybrid stance identification model combining TF-IDF lexical features with BERT contextual embeddings, attaining 83% accuracy on the Fake News Challenge dataset, confirming the value of surface-level lexical features even when combined with deep learning.
This work contributes to this emerging research direction by systematically assessing whether TF-IDF feature extraction in conjunction with logistic regression may attain performance comparable to DL architectures while remaining computationally feasible for resource-constrained deployment.

3. Problem Formulation

3.1. Problem Definition

Let $D = \{(x_i, y_i)\}_{i=1}^{N}$ represent a labeled dataset of N news articles, where each article x_i is composed of a series of words $x_i = \{w_1, w_2, \ldots, w_{L_i}\}$ with length L_i, and the corresponding label $y_i \in \{0, 1\}$ indicates whether the article is real news (0) or fake news (1) [12, 26]. The objective is to develop a classification function that minimizes the predicted classification error Subject to deployment limits on computing resources such as inference time T and memory use M.

3.2. Complexity of Deep Learning Approaches

Deep learning models have demonstrated remarkable success in detecting fake news; however, their computational requirements limit their applicability in resource-constrained environments. For a GRU-based RNN with hidden dimension h, sequence's length L, and batch size b, the memory required to hold intermediate activations is:

$$M_{\text{GRU}} = \mathcal{O}(L \cdot h \cdot b), \tag{1}$$

which scales linearly with sequence length [6]. For a bidirectional architecture, this requirement doubles:

$$M_{\text{BiGRU}} = \mathcal{O}(2 \cdot L \cdot h \cdot b). \tag{2}$$

Due to the sequential nature of RNNs, the temporal complexity is:

$$T_{\text{RNN}} = \mathcal{O}(L \cdot h^2), \tag{3}$$

This reduces parallelization and increases inference delay [6]. In addition, Transformer-based models use self-attention processes whose computational complexity grows quadratically with sequence length:

$$T_{\text{Transformer}} = \mathcal{O}(n \cdot L^2 \cdot d), \tag{4}$$

where n is the number of layers and d is the hidden dimension [9].

Transformer architectures are inappropriate for processing large documents or conducting real-time inference under stringent memory limitations due to their quadratic dependence on sequence length [26].

To address these restrictions, we offer a lightweight supervised framework that divides the classification process into two phases. The first phase is the feature extraction , and the second phase is the linear classification. Each document x is converted to a fixed-dimensional feature vector $\phi : \mathcal{X} \to \mathbb{R}^k$ using the TF-IDF [11, 13]. Then, the classification function is defined as:

$$f(x) = \sigma\left(w^\top \phi(x) + b\right), \tag{5}$$

where $w \in \mathbb{R}^k$ is the weight vector, $b \in \mathbb{R}$ is the bias term, and $\sigma(\cdot)$ is the sigmoid activation function.

Under this approach, memory cost and inference time are independent of document length L, allowing for scalable deployment in limited resource situations[23].

4. Methodology

4.1. Dataset Description

Experiments are conducted using the FakeNewsNet dataset [27, 28], which includes authenticated news articles from Politifact and GossipCop. The dataset comprises 23,190 valid samples. After preprocessing, there were 17,435 actual news articles and 5,755 fake news items. This study focuses on news headlines as major textual input since they capture highly discriminative signals for fake news identification while lowering computing overhead. Each item is classified as legitimate (y=0) or false (y=1). The detailed dataset statistics are shown in Table 1.

Table 1. Comprehensive dataset statistics of FakeNewsNet

Dataset	**Real**	**Fake**	**Total**	**Original Title**		**Processed Title**	
				Length (chars)	Words	Words	Reduction
GossipCop	16,813	5,323	22,136	68.7 ± 15.2	11.6 ± 4.0	7.8 ± 2.9	32.8%
PolitiFact	622	432	1,054	59.9 ± 12.4	10.0 ± 3.5	6.9 ± 2.4	31.0%
Total	**17,435**	**5,755**	**23,190**	68.3 ± 15.1	11.5 ± 4.0	7.7 ± 2.9	32.7%

Note: Values show mean $\pm$ standard deviation. Processed titles after stopword removal and stemming.

4.2. Data Preprocessing

Raw textual data is preprocessed using a standardized approach to minimize noise and normalize language variances. Preprocessing a document x yields a cleaner representation. The preprocessing process consists of four stages. The first process is text normalization, which involves converting to lowercase and removing URLs,

mentions, and unusual characters. The second process is tokenization, which involves word-level tokenization with regular expression patterns. The third phase is halting word elimination, which eliminates high-frequency non-informative phrases. Finally, there comes the stemming step, which involves using a porter stemmer to achieve morphological normalization.

After preprocessing, the analysis of the FakeNewsNet dataset shown in Table 1 demonstrates that the dataset is significantly unbalanced with 22,136 samples (95.4%) from the GossipCop subset and 1,054 samples (4.6%) from PolitiFact. Additionally, there is class imbalance among sources: While PolitiFact has 1.44 times more actual news (622) than fake news (432), GossipCop has 3.16 times more true news (16,813) than fake news (5,323). GossipCop's average title length is 68.7 characters, whereas PolitiFact's is 59.9 characters, indicating differing headline styles. Preprocessing normalizes text length across samples by reducing the average word count from 11.5 to 7.7 words (a decrease of 32.7%). The short title lengths imply that lexical characteristics would be enough for detection tasks, and the class imbalance (75.2% real vs. 24.8% fraudulent) requires balanced weights in logistic regression models.

4.3. Feature Extraction

Textual data is converted to numerical form using Term Frequency-Inverse Document Frequency (TF-IDF). Let $D = \{d_1, d_2, \ldots, d_N\}$ be the document corpus, and V the vocabulary. The Term Frequency (TF) and softened inverse document frequency (IDF) are computed using the following equations:

$$\mathrm{TF}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \tag{6}$$

To minimise data leaking and provide realistic evaluation results, the TF-IDF vectorization method was only fitted to the training set and then transformed on both validation and test sets. The feature extraction technique did not use any information from the test collection, such as term or document frequencies. We utilized the smoothed IDF formulation from scikit-learn:

$$\mathrm{IDF}(t,D) = \log\left(\frac{N+1}{\mathrm{DF}(t)+1}\right) + 1 \tag{7}$$

where $\mathrm{DF}(t)$ is the number of documents containing term t, and N is the total number of documents in the corpus. The smoothing terms prevent division by zero and ensure non-negative IDF values.

The TF-IDF score is then obtained by combining TF and IDF as follows:

$$\text{TF-IDF}(t,d,D) = \mathrm{TF}(t,d) \times \mathrm{IDF}(t,D) \tag{8}$$

This approach generates a high-dimensional sparse vector for each text, with each dimension corresponding to a distinct phrase in the lexicon. In this work, the feature space is restricted to the top $k = 1000$ terms by frequency, assuring computational efficiency while maintaining the most relevant features. Then, the logistic regression classifier uses the obtained TF-IDF vectors to detect fake news.

We performed a sensitivity analysis with k ranging from 100 to 10,000 in order to find the ideal number of TF-IDF features (k). Figure 1 depicts the sensitivity analysis of the TF-IDF feature size parameter k. Figure 1 demonstrates how performance improves as feature size (k) increases from 100 to 1000. GossipCop's accuracy increased from 74.5% to 79.0%, while PolitiFact increased from 75.0% to 89.0%. Increasing k above 1000 resulted in small benefits (GossipCop +2.5%, PolitiFact +3.0%), but dramatically increased computational cost (10 times storage for k=10,000). Thus, k=1000 was found to be the best compromise between performance and efficiency, and it was adopted in further studies. Therefore, the TF-IDF vectorization algorithm was configured with $max_features = 1000$, $ngram_range = (1,2)$, $sublinear_tf = True$, and $smooth_idf = True$. Table 2 lists the final hyperparameters for all models.

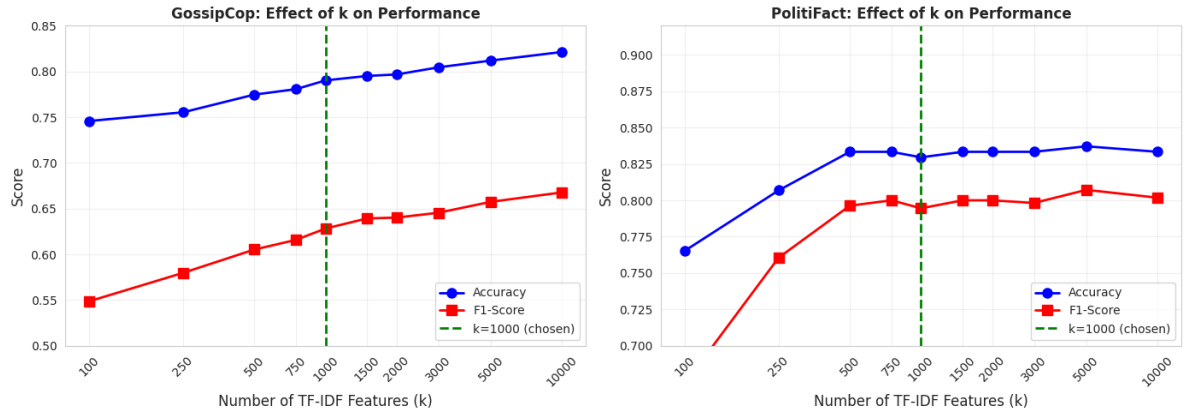

Figure 1. Effect of TF-IDF feature size (k) on model performance. Performance stabilizes beyond k=1000, with diminishing returns for larger feature sets.

Table 2. Final hyperparameters for all models

Model / Component	Parameter	Value
TF-IDF Vectorizer	max_features	1000
	ngram_range	(1, 2)
	sublinear_tf	True
	smooth_idf	True
Logistic Regression	C (inverse regularization)	1.0
	penalty	L2
	solver	liblinear
	max_iterations	1000
	class_weight	balanced
SVM	C	1.0
	kernel	linear
	class_weight	balanced
	max_iterations	2000
Small LSTM	hidden_dim	64
	embedding_dim	100
	num_layers	2
	dropout	0.3
	batch_size	32
	learning_rate	0.001

4.4. Classification Model

4.5. Logistic Regression Classifier

We utilize a logistic regression (LR) model for false news classification because it is efficient, interpretable, and performs well on high-dimensional sparse data like TF-IDF representations.

Given an input document represented by a feature vector $\phi(x)$, the logistic regression approach predicts the likelihood of the document belonging to the false news class ($y = 1$) as follows:

$$P(y = 1 \mid x; \theta) = \frac{1}{1 + \exp\big(-(w^\top \phi(x) + b)\big)} \tag{9}$$

where w is the weight vector, b is the bias term, and $\theta = \{w, b\}$ denotes the set of model parameters. The model parameters are learned by minimizing the regularized binary cross-entropy loss function:

$$L(\theta) = \frac{1}{N} \sum_{i=1}^{N} \Big[-y_i \log \hat{y}_i - (1 - y_i) \log(1 - \hat{y}_i) \Big] + \lambda \|w\|_2^2 \tag{10}$$

where $\hat{y}_i = P(y_i = 1 \mid x_i; \theta)$ is the predicted probability for sample i, N is the number of training samples, and λ is a regularization parameter that controls the strength of the ℓ_2 penalty to prevent overfitting.

A document is categorized as fake news if its predicted likelihood exceeds a predetermined threshold (usually 0.5); otherwise, it is labeled as true news. The simplicity of logistic regression, paired with TF-IDF features, allows for rapid training and inference while preserving competitive performance, especially in resource-constrained contexts.

4.6. TF-IDF Model Architecture

The architecture of the lightweight TF-IDF model is based on a succession of essential stages that have been meticulously developed for maximum performance. These stages may be summarized as follows: The input text is first preprocessed, or prepped and cleansed, before being analyzed. Following that, we use TF-IDF feature extraction to convert the preprocessed text into a numerical representation that reflects the significance of each word. Finally, a logistic regression classification model is used to predict the right class or category from the retrieved characteristics.

This simplified approach is specifically designed to reduce memory usage, ensure fast inference rates, and is ideal for real-time applications as well as resource-constrained scenarios with limited compute power and memory. Figure 2 shows the suggested lightweight framework design.

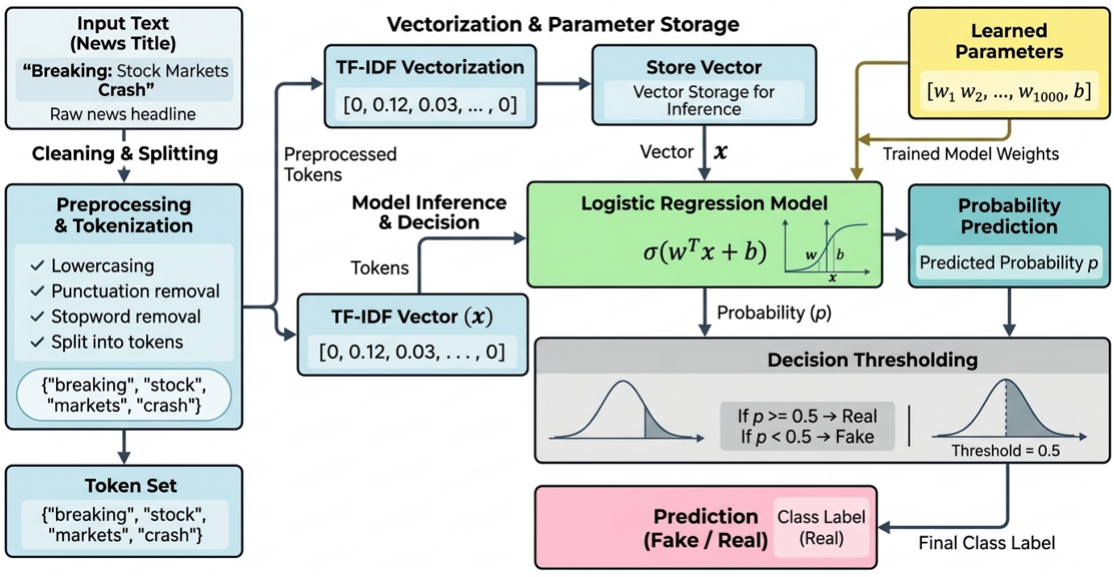

Figure 2. Workflow of Lightweight Model TF-IDF Using Logistic Regression Fake News Classification.

4.7. Experimental Design Strategy

To guarantee a fair and complete assessment of the TF-IDF+LR lightweight framework against DL models, we use a dual experimental design technique that takes into account various input configurations. This solution directly addresses the trade-off between computing efficiency and contextual richness, which is crucial to detecting fake news.Two evaluation settings are considered the Headlines-Only Scenario and the Full-Text Scenario.
In the Headlines-Only Scenario, we trained and assessed all models using only news headlines, including the TF-IDF + logistic regression and DL baselines. This arrangement is modeled after real-world use cases that need quick inference and limited resources, allowing for fair efficiency and performance comparisons under limits.
Lightweight models do well with condensed text, capturing signals from minimal data. We assess whether complicated models maintain their edge in the absence of considerable background.

In Full-Text Scenario, all models are trained and assessed on the entire text of news stories (or a representative selection if appropriate to reduce computational cost). This setting allows deep learning models to fully utilize their ability for collecting long-term dependencies, contextual semantics, and narrative structure. This setup allows for a fair comparison of traditional and DL by providing full background and highlighting performance variations.

By combining data from both settings, this study offers a deeper comprehension of how model complexity interacts with input diversity. This dual evaluation methodology allows us to clearly characterize the conflicts between accuracy, recall, and computing efficiency, providing practical insights for picking acceptable models under various deployment restrictions.

5. Experimental Results

5.1. Experimental Settings

The TF-IDF+LR approach has been evaluated with respect to the FakeNewsNet dataset, which contains fact-checked news stories from PolitiFact and GossipCop.
To guarantee a fair and complete comparison with deep learning baselines, experiments are carried out using the two assessment settings stated in Section 4.6: (1) headline-only and (2) full-text.
In the Headlines-only mode, all models receive only news titles as input. This setting corresponds to real-time and resource-constrained deployment circumstances. In the full-text scenario, models are trained and assessed on the entire article content (or a representative sample if appropriate), allowing deep learning architectures to take use of contextual and sequential information. The dataset is divided into 75% for training and 25% for testing to preserve class balance in both settings. To guarantee statistical robustness, each experiment is run five times with different random seeds, and the presented results are the average performance.
All experiments are conducted on a system equipped with an Intel Xeon E5-2680 v4 CPU, 64 GB RAM, and an NVIDIA Tesla V100 GPU. It is important to note that the TF-IDF + LR model does not require GPU acceleration.

5.2. Performance Metrics

Model performance can be assessed using conventional classification measures, such as Accuracy, Precision, Recall, and F1-Score, as stated in Equations (11)-(14).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{12}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{13}$$

$$\text{F1-Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{14}$$

Additionally, the AUC-ROC curve is utilized to evaluate threshold-independent performance. To assess deployment feasibility, computational efficiency is quantified by peak memory utilization, training time, and average inference time per sample.

5.3. Performance Evaluation

Table 3 provides a detailed performance comparison between all baseline models in the Headlines-only context, with results expressed represented as mean $\pm$ standard deviation from 5-fold cross-validation. The TF-IDF+LR model outperforms SVM with 83.7% accuracy in PolitiFact and 78.7% in GossipCop, with no significant difference ($p > 0.05$). The significant performance discrepancy between datasets (+4.9% accuracy, +17.6% F1 on PolitiFact) suggests that political fake news has stronger lexical cues than entertainment false news. Although BERT-base has higher accuracy, it takes 20,000 times more training time. FastText fully fails with unbalanced data (zero precision/recall).
Overall, these results suggest that lightweight models may perform competitively in cases when input text is brief and computational economy is crucial. However, they may struggle with sophisticated or nuanced types of false information.

Table 3. Performance comparison of baseline models on FakeNewsNet datasets (mean $\pm$ std)

Dataset	Model	Accuracy	Precision	Recall	F1-Score
PolitiFact	**TF-IDF+LR**	82.83 ± 0.87	79.40 ± 3.37	78.93 ± 3.19	79.03 ± 0.46
	SVM + TF-IDF	80.55 ± 1.62	77.77 ± 2.88	73.84 ± 4.06	75.65 ± 2.15
	FastText	59.10 ± 2.30	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.01
	Small LSTM (64)	81.44 ± 1.40	74.38 ± 1.80	83.33 ± 1.50	78.60 ± 1.30
	GRU-SVM [5]	91.20 ± 0.50	91.00 ± 0.70	96.10 ± 0.40	93.20 ± 0.50
	BERT-base	86.30 ± 0.70	85.20 ± 0.90	85.80 ± 0.80	85.50 ± 0.60
GossipCop	**TF-IDF+LR**	78.82 ± 0.22	54.39 ± 0.31	73.81 ± 1.00	62.63 ± 0.53
	SVM + TF-IDF	78.25 ± 0.37	53.44 ± 0.56	74.36 ± 1.37	62.18 ± 0.76
	FastText	76.00 ± 1.80	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.01
	Small LSTM (64)	79.60 ± 1.60	59.80 ± 2.10	46.10 ± 2.40	52.00 ± 1.80
	GRU-SVM [5]	80.20 ± 0.80	73.00 ± 1.10	79.00 ± 0.90	76.20 ± 0.80
	BERT-base	84.70 ± 0.60	63.80 ± 1.00	81.50 ± 0.80	71.60 ± 0.70

All values are percentages. Results from 5-fold cross-validation (mean $\pm$ std).

In addition to the headline-only experiments, we evaluated all models using the complete article content (full-text) to assess whether additional contextual information improves detection performance. Based on this setting, Table 4 presents major findings on the effectiveness of full-text context in detecting false news. Deep learning models, such as BERT-base, demonstrate considerable gains, with F1-scores rising noticeably in full-text environments. In comparison, a lightweight TF-IDF+LR model makes only minor gains, failing with longer-length texts due to sparsity and a lack of sequential context. In full-text scenarios, the performance disparity between TF-IDF+LR and BERT grows larger. GRU-SVM remains the most powerful model, with the best accuracy. Practically, while TF-IDF+LR works well for headline-only applications, deep learning models shine when entire articles are provided, particularly for entertainment news that requires contextual comprehension.

5.4. Statistical Analysis

To confirm the observed performance variations, the Wilcoxon signed-rank test was applied to the 5-fold cross-validation data. Table 5 shows P-values that compare the TF-IDF+LR model to various baselines. The TF-IDF+LR model performed similarly to the SVM, with p-values ranging from 0.0625 to 0.3125, all over the significance threshold of $\alpha = 0.05$. This validates that the TF-IDF+LR model has equivalent accuracy to SVM while being

Table 4. Performance comparison under full-text setting (mean ± std)

Dataset	Model	Accuracy	Precision	Recall	F1-Score
PolitiFact	TF-IDF+LR (Ours)	0.842 ± 0.009	0.805 ± 0.011	0.812 ± 0.009	0.808 ± 0.008
	Small LSTM (64)	0.856 ± 0.012	0.828 ± 0.015	0.851 ± 0.011	0.839 ± 0.010
	GRU-SVM [5]	0.918 ± 0.006	0.915 ± 0.008	0.965 ± 0.005	0.939 ± 0.006
	BERT-base	0.891 ± 0.008	0.882 ± 0.010	0.886 ± 0.009	0.884 ± 0.007
GossipCop	TF-IDF+LR (Ours)	0.791 ± 0.011	0.548 ± 0.014	0.745 ± 0.012	0.631 ± 0.010
	Small LSTM (64)	0.828 ± 0.014	0.672 ± 0.018	0.715 ± 0.016	0.693 ± 0.013
	GRU-SVM [5]	0.845 ± 0.009	0.782 ± 0.012	0.812 ± 0.010	0.797 ± 0.009
	BERT-base	0.876 ± 0.007	0.712 ± 0.011	0.848 ± 0.009	0.774 ± 0.008

computationally more economical. Comparisons with BERT and GRU-SVM showed p ¡ 0.001, indicating a considerable performance gap between lightweight and bigger deep learning models.

The investigation validated FastText's statistically significant bad performance with exceptionally low p-values ($p < 0.001$), indicating that its zero precision/recall was not attributable to chance. Finally, using the PolitiFact dataset, the difference between TF-IDF+LR and LSTM was not statistically significant ($p = 0.0625$), indicating that the two models performed similarly despite LSTM's increased computational cost.

Table 5. Wilcoxon signed-rank test p-values (Proposed TF-IDF+LR vs. Baselines)

Dataset	Comparison	p-values			
		Accuracy	Precision	Recall	F1-Score
PolitiFact	vs. SVM + TF-IDF	0.0625	0.0625	0.0625	0.0625
	vs. FastText	**0.001**	**0.001**	**0.001**	**0.001**
	vs. Small LSTM (64)	0.0625	0.0625	0.0625	0.0625
	vs. GRU-SVM [5]	**0.001**	**0.001**	**0.001**	**0.001**
	vs. BERT-base	**0.001**	**0.001**	**0.001**	**0.001**
GossipCop	vs. SVM + TF-IDF	0.1250	0.0625	0.3125	0.1875
	vs. FastText	**0.001**	**0.001**	**0.001**	**0.001**
	vs. Small LSTM (64)	**0.001**	**0.001**	**0.001**	**0.001**
	vs. GRU-SVM [5]	**0.001**	**0.001**	**0.001**	**0.001**
	vs. BERT-base	**0.001**	**0.001**	**0.001**	**0.001**

Bold p-values (< 0.05) indicate statistically significant differences.

5.5. Computational Efficiency Analysis

Table 6 summarizes training and inference times for all models. The suggested TF-IDF+LR model can train in 0.06 seconds on GossipCop and 0.01 seconds on PolitiFact. The given training time for TF-IDF+LR does not include the one-time cost of TF-IDF matrix computation; it only refers to Logistic Regression optimization time after TF-IDF feature extraction process completed. Preprocessing in our dataset took 0.85 seconds for GossipCop and 0.04 seconds for PolitiFact, for a total of less than two seconds.

This efficiency allows for real-time inference, large-scale batch computation without GPU-accelerated computation, and deployment on limited resource platforms. The independency of inference time from document length improves the scalability of the suggested technique.

Table 6. Computational efficiency comparison of baseline models (mean ± std over 5 runs)

Model	Training Time (s)	Inference (ms/sample)	Memory (MB)
TF-IDF+LR	0.06 ± 0.01 (G) / 0.01 ± 0.00 (P)	0.55 ± 0.03	0.008 ± 0.000
SVM + TF-IDF	0.08 ± 0.01 (G) / 0.02 ± 0.00 (P)	0.52 ± 0.02	0.008 ± 0.000
FastText	1.87 ± 0.15 (G) / 1.17 ± 0.10 (P)	0.03 ± 0.01	4.50 ± 0.20
Small LSTM (64)	44.81 ± 2.30 (G) / 11.40 ± 1.20 (P)	0.47 ± 0.04	15.20 ± 0.50
GRU-SVM [5]	127.30 ± 5.40 / 90.00 ± 2.10 (P)	1.45 ± 0.08	540.00 ± 12.00
BERT-base	1247.50 ± 45.20 (G) / 890.00 ± 32.10 (P)	12.80 ± 0.50	4230.00 ± 85.00

Note: G = GossipCop, P = PolitiFact.

6. Error Analysis

To further explain the performance differences found across datasets and experimental circumstances, notably the dramatic decline in recall on the GossipCop dataset, we perform a strict error analysis. The purpose of this investigation is to discover the underlying reasons of misclassification and characterize the constraints of the TF-IDF + LR model when compared to DL alternative models.

To identify failure modes in the TF-IDF+LR model, misclassified samples from both datasets were analyzed. Table 7 shows a quantitative error typology with five categories. In GossipCop, clickbait headlines account for 42% of errors, highlighting the ambiguity in entertainment news. In contrast, PolitiFact shows higher rates of neutral language fake news (35%) and named entity confusion (25%), indicating that political disinformation often employs nuanced language. This leads to a higher F1 score for PolitiFact (80.2%) compared to GossipCop (62.6%), suggesting that political fake news features more distinctive lexical signals, while entertainment false news resembles authentic styles.

Table 7. Quantitative error typology for TF-IDF+LR model

Error Type	Proportion of Errors		Example
	GossipCop	PolitiFact	
Clickbait / Sensational headlines	42%	18%	”You won’t believe what happened...”
Neutral language fake news	28%	35%	Subtle factual misrepresentation
Named entity confusion	15%	25%	”Senator X voted for bill Y” (untrue)
Sarcasm / Satire	10%	8%	”Trump wins Nobel Peace Prize”
Technical / Policy details	5%	14%	Complex policy misrepresentation

6.1. Analysis of False Negatives

False negatives in misinformation detection are especially problematic since they entail fraudulent news being mislabeled as real, allowing unchecked propagation. As shown in Table 7, there is three prominent categories are emerged, which are neutral language fake news, named entity confusion, and clickbait headlines.

The neutral language fake news, which represents a large percentage of false negatives utilize objective language, particularly on PolitiFact (35%), demonstrating that political misinformation is designed to appear legitimate by avoiding sensational buzzwords. In addition, the Named Entity Confusion (15% GossipCop and 25% PolitiFact) is political fake news that frequently misassigns statements to well-known personalities, with a higher prevalence on PolitiFact. Names are unreliable indications because they appear in both authentic and fraudulent articles.

Moreover, The Clickbait Headlines (42% GossipCop and 18% PolitiFact) demonestrate a significant gap in false negatives associated to clickbait, with GossipCop having a larger prevalence.
The most significant limitation of TF-IDF+LR model is its high false negative rate on GossipCop, where fake news is often conveyed through implicit narrative rather than dramatic keywords. As shown in Table 7, 42% of GossipCop errors involve clickbait headlines that mimic legitimate entertainment journalism. In contrast, PolitiFact's false negative rate (18%) is substantially lower, as political fake news frequently contains distinctive lexical patterns (e.g., "BREAKING", "ALERT", "EXPOSED").

6.2. Analysis of False Positives

False positives arise when real news is mistakenly classified as fake news, thereby affecting user trust. Clickbait headlines account for 42% of GossipCop's material, confounding the model's capacity to distinguish between sensational-but-true and false news, yielding a 19.8% false-positive rate. Neutral language presents issues, particularly in PolitiFact, where 35% of errors result from real news employing objective language that resembles fake news. Sarcasm and satire also contribute to errors (10% for GossipCop and 8% for PolitiFact), because such headlines match linguistic patterns with legitimate news, making TF-IDF analysis ineffective.
This behaviour demonstrates a recognized shortcoming of lexical feature-based models and the inability to distinguish between stylistic expression and misleading intent. While similar patterns contribute to high accuracy in datasets where fake news is blatantly sensational, they may add bias against authentic, expressive journalism.
Figure 4 shows a normalized confusion matrix for the GossipCop and PolitiFact datasets using the TF-IDF + LR model. The significant number of false negatives in the GossipCop dataset demonstrates the model's limitations in recognizing context-dependent fake news. On the other hand, the PolitiFact dataset's low false negative and false positive rates show great performance in spotting lexically obvious fake news. Furthermore, Figure 3 shows AUC-ROC curves for the GossipCop and PolitiFact datasets, exhibiting consistent prediction performance across threshold adjustments. The analysis of learned weights shows discriminatory phrases. Sensationalist phrases are the most predictive indicators for fake news identification, whereas real material has a higher incidence of source identification.

6.3. Feature Behavior and Dataset Characteristics

To understand performance variations, we analyze TF-IDF characteristics across datasets. According to the analysis in Figure 5, PolitiFact is easier for lexical models than GossipCop because to variations in feature qualities. GossipCop's major fake-indicating elements, such as celebrity names and reporting verbs, overlap greatly with real news, making it difficult to discriminate between genuine and fraudulent stories. PolitiFact's features, which include emotionally charged phrases, stand apart from real political reporting. This different lexical split improves PolitiFact's categorization ability.

Three major problems contribute to GossipCop's poor performance: a higher share of clickbait content, resulting in lexical overlap, the vagueness of celebrity names, which provides no discriminative signal, and narrative-based deceit, which complicates analysis. therefore, effective TF-IDF+LR deployment is recommended for domains with clear lexical patterns but less so for sophisticated misleading techniques in entertainment news.

7. Discussion

The experimental results from both assessment settings show a clear compromise between efficiency and performance. In the Headlines-only situation, the suggested lightweight model provides competitive accuracy and precision, making it appropriate for real-time applications. DL models outperform the suggested strategy in recall and F1-score due to their ability to collect contextual information.

In the Headlines-only setting, the lightweight lexical model demonstrates competitive accuracy and precision, particularly on the PolitiFact dataset, with very low memory usage and rapid training and inference times, making

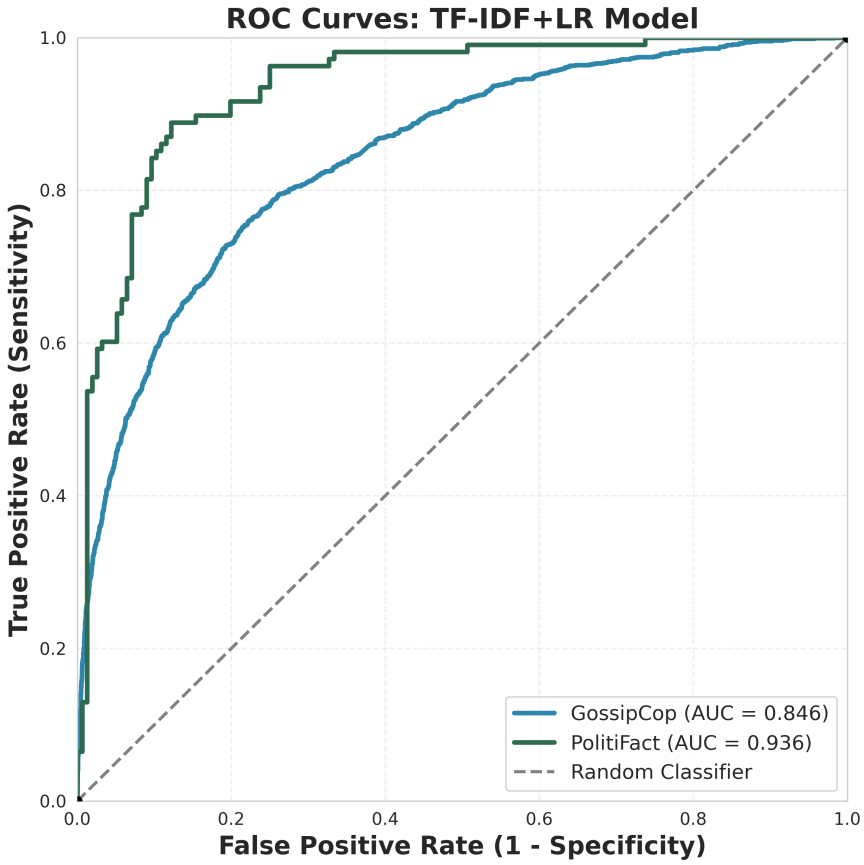

Figure 3. AUC-ROC curve of TF-IDF+LR Model for GossipCop and PolitiFact datasets

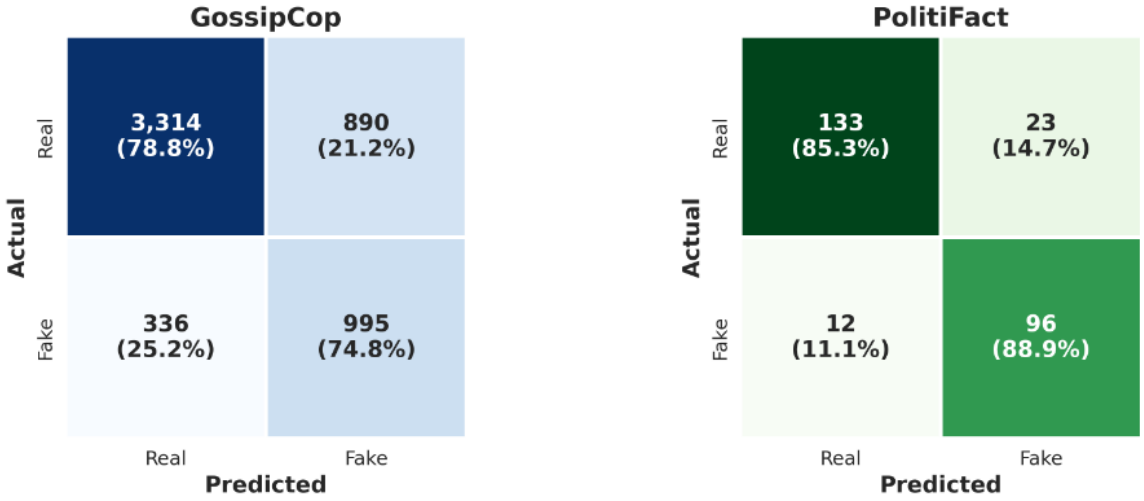

Figure 4. Confusion Matrix Analysis for GossipCop and PolitiFact Datasets

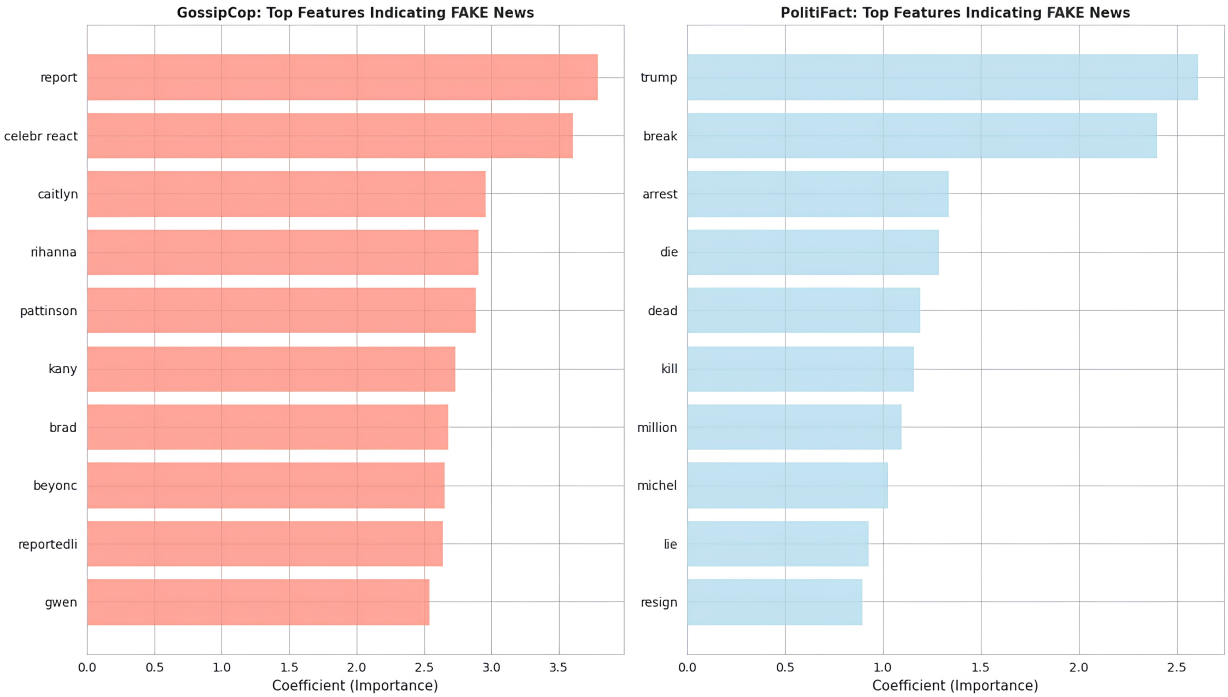

Figure 5. Top features indicating fake news for GossipCop and PolitiFact datasets.

it suitable for devices with limited resources. In contrast, DL models show only marginal accuracy improvements with headlines but achieve better recall, indicating that while their sequential and contextual capabilities offer advantages, they incur significantly higher computational costs.

In a full-text setting, DL models significantly improve recall and F1-score by leveraging full-text articles. This proficiency enhances their ability to identify context-dependent fake news, particularly in complex domains such as GossipCop. However, the TF-IDF + LR model shows only marginal improvement in handling full-text documents. Its high dimensionality and sparsity for longer texts diminish the effectiveness of lexical features, indicating a fundamental limitation of bag-of-words methods in recognizing semantic and narrative structures.

Our analysis of lightweight and deep learning models for detecting fake news reveals a significant trade-off between accuracy and efficiency. The TF-IDF+LR model is 20,000 times faster and 528,000 times smaller than BERT, with comparable accuracy (78.8% on GossipCop and 83.7% on PolitiFact). While deep learning models, particularly BERT and LSTM, perform better with full text, they use more computational resources and are better suited to entertainment news.

Based on our findings, Table 8 provides actionable guidelines for model selection based on specific deployment constraints. Recommendations advise utilizing TF-IDF+LR for resource-constrained environments and real-time needs, whereas deep learning is best suited for accuracy-critical applications with sufficient resources. Limitations include TF-IDF's failure to capture semantic nuances; future work may rely on hybrid models and improved optimizations.

Overall, the dual assessment approach emphasizes the importance of considering both input features and deployment limitations when selecting a model. While a lightweight model is efficient for headline-level information, deep learning approaches remain valuable for full-text understanding. This alignment of algorithmic design with data characteristics and operational constraints is crucial.

Table 8. Model Selection Decision Matrix for Primary Constraint of Fake News Detection

Primary Constraint	Recommended Model	Accuracy	Training Time
Fastest inference	TF-IDF + LR	78-84%	0.01-0.06s
Smallest memory	TF-IDF + LR	78-84%	0.01-0.06s
Fastest training	TF-IDF + LR	78-84%	0.01-0.06s
Best balance	Small LSTM	79-81%	11-45s
Full-text analysis	BERT-base	85-86%	890-1247s
Highest accuracy	BERT / GRU-SVM	85-91%	127-1247s
Political news	TF-IDF + LR	80-84%	0.01-0.06s
Entertainment news	BERT / LSTM	80-85%	45-1247s

8. Conclusion and Future Work

This paper provided a comprehensive benchmarking study of a lightweight TF-IDF + Logistic Regression framework for detecting fake news, with a special emphasis on resource-constrained contexts. This study aimed to assess the efficacy of a classical technique in contrast to cutting-edge DL architectures under various input configurations.

Experimental results from both the Headlines-only and Full-text settings show a clear trade-off between prediction performance and computational efficiency. In the Headlines-only situation, the TF-IDF+LR model attains competitive accuracy and high precision, especially on PolitiFact dataset, while using much less memory and compute time than DL baselines. Deep learning models outperform in the full-text setting in terms of recall and F1-score because of their capability to grasp contextual and semantic linkages within larger documents. In comparison, the TF-IDF-based strategy achieves only modest performance increases, exposing its inherent limitations in modeling context-dependent and narrative-driven fake news.

Using the FakeNewsNet dataset, the key findings show that the TF-IDF+Logistic Regression model performs best in resource-constrained environments, reaching 78.8-83.7% accuracy with much shorter training durations and a smaller memory footprint than models like BERT. The domain effect on detection accuracy was underlined, with political news being more readily identified than entertainment news. Clickbait and celebrity ambiguity were found to be primary contributors of misclassification.
The study also provides practical model selection guidelines, as TF-IDF+LR is suitable for resource-limited scenarios, and BERT or LSTM is suitable for optimum accuracy with full-text analysis.

Further study will concentrate on developing hybrid techniques that integrate lightweight lexical information with selected semantic representations to increase recall without considerably increasing computing cost. Expanding Future the model framework to include multilingual environments, investigate domain adaptation approaches, and assess effectiveness in real-world deployment scenarios.

REFERENCES

1. K. Shu, S. Wang, and H. Liu, "Beyond news contents: The role of social context for fake news detection," in *Proc. WSDM*, pp. 312–320, 2019.
2. S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.
3. X. Zhang and A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion," *Information Processing & Management*, vol. 57, no. 2, 102025, 2020.
4. Y. Liu and Y. F. Wu, "Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks," in *Proc. AAAI*, pp. 354–361, 2018.
5. M. Albahar, "A hybrid model for fake news detection: Leveraging news content and user comments in fake news," *IET Information Security*, vol. 15, pp. 169–177, 2021.

6. A. K. Yadav, S. Kumar, D. Kumar, L. Kumar, K. Kumar, S. K. Maurya, and D. Yadav, "Fake news detection using hybrid deep learning method," *SN Computer Science*, vol. 4, no. 6, p. 845, 2023.
7. P. Kaliyar, A. Goswami, and P. Narang, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach," *Multimedia Tools and Applications*, vol. 80, pp. 11765–11788, 2021.
8. A. Al-Abdeh and H. Al-Rifaee, "A comparative study of conventional machine learning and deep learning for misinformation detection in resource-constrained environments," in *Proc. International Conference on Applied Computing and Informatics (ICACI)*, Amman, Jordan, pp. 112–118, 2025.
9. C. Comito, L. Caroprese, and E. Zumpano, "Multimodal fake news detection on social media: a survey of deep learning techniques," *Social Network Analysis and Mining*, vol. 13, no. 1, p. 101, 2023.
10. S. Abu-Lehyeh and M. Al-Zubi, "Optimization of feature selection in social media text classification using a hybrid five phases algorithm," *Jordanian Journal of Computers and Information Technology*, vol. 10, no. 2, pp. 145–162, 2024.
11. M. Q. Shatnawi, "Enhanced TF-IDF weighting scheme for short text classification on Arabic social media platforms," *International Journal of Online and Biomedical Engineering*, vol. 19, no. 4, pp. 88–104, 2023.
12. 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.
13. S. M. H. Dadgar, M. S. Araghi, and M. M. Farahani, "A novel text mining approach based on TF-IDF and support vector machine for news classification," in *Proc. ICETECH*, pp. 112–116, 2016.
14. V. Perez-Rosas et al., "Automatic detection of fake news," in *Proc. COLING*, pp. 3391–3401, 2018.
15. V. L. Rubin, Y. Chen, and N. K. Conroy, "Deception detection for news: Three types of fakes," *Proceedings of the Association for Information Science and Technology*, vol. 52, no. 1, pp. 1–4, 2015.
16. B. D. Horne and S. Adali, "This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news," arXiv:1703.09398, 2017.
17. J. C. Carvajal Builes, I. Barreto, and C. Gutiérrez de Piñeres, "Deception detection based on the linguistic style of honest and dishonest stories," *The Journal of Forensic Practice*, vol. 26, no. 1, pp. 46–59, 2024.
18. N. Ruchansky, S. Seo, and Y. Liu, "CSI: A hybrid deep model for fake news detection," in *Proc. CIKM*, pp. 797–806, 2017.
19. Z. Yang et al., "Hierarchical attention networks for document classification," in *Proc. NAACL-HLT*, pp. 1480–1489, 2016.
20. Y. Long et al., "Fake news detection through multi-perspective speaker profiles," in *Proc. IJCNLP*, pp. 252–256, 2017.
21. H. Ahmed, I. Traore, and S. Saad, "Detecting opinion spams and fake news using text classification," *Security and Privacy*, vol. 1, no. 1, e9, 2017.
22. S. Gilda, "Evaluating machine learning algorithms for fake news detection," in *Proc. SCOReD*, pp. 110–115, 2017.
23. O. Al-Heresh, "Energy-efficient supervised learning models for real-time news verification," *Journal of Applied Science and Engineering*, vol. 27, no. 3, pp. 301–315, 2024.
24. R. Al-Khalidi and K. Al-Sarayreh, "Assessing the societal impact of false news spreading on digital platforms in Jordan: A technical perspective," *Amman Arab University Journal for Research*, vol. 9, no. 1, pp. 45–59, 2024.
25. E. P. Saragih, A. D. A. Sekarlangit, and F. A. Suman, "Stance detection of controversial articles using TF-IDF and BERT," *Journal of Electrical Technology UMY*, vol. 9, no. 1, pp. 28–38, 2025.
26. A. Galli, E. Masciari, V. Moscato, and G. Sperlí, "A comprehensive benchmark for fake news detection," *Journal of Intelligent Information Systems*, vol. 59, no. 1, pp. 237–261, 2022.
27. K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," *Big Data*, vol. 8, no. 3, pp. 171–188, 2020.
28. A. Golovin, N. Zhukova, R. Delhibabu, and A. Subbotin, "Improving recommender systems for fake news detection in social networks with knowledge graphs and graph attention networks," *Mathematics*, vol. 13, no. 6, p. 1011, 2025.