A COMPARATIVE EVALUATION OF LIGHTWEIGHT TRANSFORMER MODELS AND TRADITIONAL ML APPROACHES FOR FAKE NEWS DETECTION

Ameed Othman
Faculty of Information Technology and AI
An-Najah National University
Nablus, Palestine
othman.ameed@gmail.com

Dr. Adnan Salman
Faculty of Information Technology and AI
An-Najah National University
Nablus, Palestine
aalshaikh@najah.edu

ABSTRACT

In today's digital age, the rise of large language models has made it incredibly easy to create and share fake news that looks convincing. With just a digital device and access to AI tools, anyone can generate fake content that mimics real news in style and format. This makes having good fake news detection (FND) methods more important than ever. While transformer-based deep learning models show excellent results for FND, they need a lot of computing power, making them hard to use in many real-world situations. In our study, we compared several models for detecting fake news: lightweight transformer models (DistilBERT, TinyBERT, MobileBERT, ALBERT) and traditional machine learning approaches (Logistic Regression, Random Forest). We tested these models on the WELFake dataset using standardized train/validation/test splits to ensure fair comparison. Our results show that while transformer models achieved excellent accuracy (99.28%-99.66%) compared to traditional ML models (95.61%-95.96%), the difference is relatively modest. However, when evaluating generalization to external datasets containing different types of misinformation, traditional ML models demonstrated superior robustness, with Random Forest achieving 97.79% accuracy compared to significant performance drops in transformer models. ALBERT achieved the best transformer performance with 99.66% accuracy while maintaining the smallest model size at 44.58 MB, while TinyBERT offered the best efficiency with 14.87-minute training time and competitive 99.28% accuracy. Our comprehensive analysis of performance, efficiency, and generalization capabilities provides practical guidance for choosing appropriate models based on specific deployment requirements and challenges the assumption that more complex models are always superior for fake news detection.

***K*eywords** Fake News Detection, Transformer Models, TinyBERT, Resource Efficiency, Machine Learning

1 Introduction

Fake news has become a major problem in our modern information ecosystem. When false information is presented as real news, it can shape public opinion [1], damage trust in legitimate media, and even affect democratic elections. The problem has gotten worse with recent advances in AI and large language models (LLMs), which can generate very convincing fake content [2] that's hard to distinguish from real news.

The research community has extensively studied fake news detection, with seminal works like [3] establishing foundational techniques and challenges in this domain. Early systems primarily relied on content-based features, but recent approaches leverage deep contextual understanding through transformer models like BERT [4]. However, there's a growing need to balance detection performance with computational efficiency, especially as the recent trend in moving towards edge AI.

Researchers have made good progress in developing fake news detections systems, especially using deep learning approaches like transformer-based models. These models can understand context and language nuances better than

earlier methods. However, the drawback is that they need a lot of computing power. This makes them difficult to use in many real-world situations, especially on mobile devices or when real-time detection is needed.

Our research addresses this problem by comparing different lightweight models for fake news detection. We look at both transformer-based models (DistilBERT, TinyBERT, MobileBERT, ALBERT) and traditional machine learning approaches (Logistic Regression, Random Forest). For each model, we measure both how well it detects fake news and how much computing power it needs.

The main questions we're trying to answer are:

1. Which lightweight model gives the best balance between FND accuracy and computational efficiency?
2. How do transformer models compare to traditional ML approaches in terms of performance and efficiency?

Most previous research has focused mainly on improving detection accuracy without paying much attention to the resources required. This creates a gap between research and practical applications, especially for resource-constrained environments like mobile devices or browser extensions. Our research aims to bridge this gap by providing a clear picture of the tradeoffs involved, helping developers and researchers choose the right model for their specific needs.

2 Literature Review

The proliferation of misinformation has intensified dramatically with sophisticated large language models capable of generating highly convincing fake content [5, 6]. Modern AI systems like GPT and its successors produce text so polished that distinguishing authentic from fabricated content has become increasingly challenging, even for experts [7]. This technological advancement has democratized fake news creation, enabling anyone with basic digital literacy to generate convincing misinformation, fundamentally altering the information landscape [8].

Existing fake news detection approaches can be classified into three main categories: traditional machine learning methods, deep learning approaches, and hybrid systems [9]. Traditional machine learning methods rely on manual feature engineering such as text length, readability scores, and lexical diversity [10]. These features train classifiers like Support Vector Machines and Random Forests, achieving 70-85% accuracy with high interpretability [11]. However, these approaches struggle with the nuanced semantic patterns of modern misinformation, failing to capture subtle linguistic manipulations employed by sophisticated campaigns [2].

Deep learning methods have demonstrated superior performance through automatic feature extraction [12]. Convolutional Neural Networks excel at capturing local semantic patterns, while Recurrent Neural Networks and LSTMs effectively model sequential dependencies in text [13]. Recent transformer models like BERT achieve 96-99% accuracy on benchmark datasets, fundamentally transforming detection capabilities through their ability to process entire documents holistically [9].

However, current research reveals critical limitations. While transformer models achieve exceptional accuracy, they require substantial computational infrastructure that exceeds typical deployment capabilities [14]. Recent analysis shows that only 23% of studies include computational efficiency metrics, and fewer than 15% evaluate generalization across different misinformation types [15]. More concerning, models trained on specific misinformation types show accuracy drops of up to 39% when encountering new patterns [16].

The interpretability challenge remains equally significant. State-of-the-art models often function as black boxes, making it difficult to understand decision-making processes—a critical limitation for high-stakes applications in journalism and legal contexts [17, 18]. This lack of explainability undermines trust and limits practical adoption in sensitive domains.

Recent developments in lightweight transformer models offer promising solutions to these challenges. Knowledge distillation techniques have produced efficient alternatives like DistilBERT (40% parameter reduction with 97% performance retention) [19], TinyBERT (7.5× compression with minimal accuracy loss) [20], and MobileBERT (optimized for mobile deployment) [21]. ALBERT demonstrates that parameter sharing can achieve dramatic size reductions while maintaining sophisticated language understanding [22].

Our study addresses the critical gap in comprehensive evaluation methodologies by conducting the first systematic comparison of traditional ML approaches with lightweight transformer models across three essential dimensions: accuracy on familiar data, computational efficiency, and generalization to new misinformation types. This evaluation framework provides practical guidance for model selection while advancing understanding of fundamental trade-offs in different architectural approaches to fake news detection, addressing the documented disconnect between research achievements and real-world deployment requirements.

3 Methodology

In this section, we explain our comprehensive experimental design for evaluating different models for fake news detection, covering our dataset preparation, model selection rationale, preprocessing steps, training configurations, and evaluation methodology.

3.1 Dataset and Preparation

We utilized the WELFake dataset, which is a comprehensive collection that combines real and fake news articles from four different sources: Politifact, GossipCop, Reuters, and BuzzFeed. The dataset contains 72,134 articles with a well-balanced distribution of 48.96% real news (35,028 articles) and 51.04% fake news (36,509 articles). After cleaning to handle missing values (558 missing titles and 39 missing text entries), the final dataset contains 71,537 articles. The WELFake dataset offers several advantages for fake news detection research. Its balanced nature reduces the risk of class bias during model training, and its diverse sources provide a variety of writing styles and topics, making it an excellent benchmark for comparing different approaches under realistic conditions. To ensure fair model comparison, we implemented standardized data splits: 70% for training (50,075 articles), 15% for validation (10,731 articles), and 15% for testing (10,731 articles). The validation set was used for hyperparameter tuning and early stopping, while the test set remained completely unseen until final evaluation.

3.2 Exploratory Data Analysis

Our exploratory data analysis revealed several critical insights that informed our preprocessing decisions and provided understanding of the underlying patterns that distinguish fake from real news. We conducted detailed examinations of content length patterns, linguistic characteristics, and stylistic differences between authentic and fabricated articles. Content length analysis revealed measurable differences between real and fake news. Fake news titles tend to be longer (mean: 85.13 characters) compared to real news (mean: 68.79 characters), suggesting that fabricated content often uses longer, more sensationalist headlines to attract attention. Conversely, real news articles are generally longer (mean: 3,495 characters) than fake news (mean: 3,098 characters), indicating that legitimate news sources typically provide more comprehensive coverage and detailed reporting. Our linguistic pattern analysis uncovered several distinguishing characteristics that help explain why machine learning models can successfully differentiate between real and fake news. Attribution words like "said" appear much more frequently in real news (122,295 times) compared to fake news (31,617 times), suggesting that authentic journalism more consistently attributes information to sources. Political references show distinct patterns, with fake news containing more direct references to specific political figures. Institutional focus differs significantly, with real news using more institutional terms like "government," "states," and "united," suggesting greater emphasis on official entities and formal reporting structures. Stylistic analysis revealed dramatic differences in presentation approaches. We found that 63.2% of fake news titles contain words in ALL CAPS, compared to only 23.5% of real news titles, indicating the use of dramatic formatting to create urgency or emotional response. Similarly, 10.8% of fake news titles contain exclamation marks versus just 0.2% of real news titles, representing a 54-fold difference. These patterns provide valuable insights into how fabricated content differs from legitimate reporting and offer potential features that classification models can leverage.

3.3 Data Preprocessing and Feature Engineering

Based on our exploratory analysis findings, we implemented a comprehensive preprocessing pipeline designed to optimize model performance while maintaining consistency across different architectural approaches. For all models, we combined article titles and bodies to provide complete contextual information, as our analysis showed that both components contain valuable distinguishing features. For traditional machine learning models, we implemented TF-IDF vectorization with carefully tuned parameters: max_features=10000 to capture a comprehensive vocabulary while maintaining computational efficiency, min_df=5 to eliminate extremely rare terms that might represent noise, and max_df=0.8 to remove overly common words that provide little discriminative power. This configuration resulted in feature matrices that effectively captured the linguistic patterns we identified in our exploratory analysis. For transformer models, we applied model-specific tokenization using the appropriate pre-trained tokenizers, with padding and truncation to a maximum sequence length of 512 tokens. This length was chosen to accommodate the vast majority of articles while maintaining computational efficiency. Each model required specific preprocessing considerations: DistilBERT used the standard BERT tokenization approach, ALBERT benefited from its factorized embedding approach, TinyBERT leveraged its knowledge distillation optimizations, and MobileBERT employed its mobile-optimized tokenization strategy.

3.4 Model Selection and Architectures

We selected models that represent different philosophical approaches to achieving efficient fake news detection, enabling us to evaluate the fundamental tradeoffs between complexity and performance in this domain. Our traditional machine learning approaches included Logistic Regression, chosen for its interpretability and effectiveness with high-dimensional text data, and Random Forest, selected for its ability to capture non-linear feature interactions while maintaining relative simplicity. These models serve as important baselines and represent approaches that many organizations can readily implement and understand. Our transformer-based models were carefully chosen to represent different efficiency optimization strategies. DistilBERT [19] (66.96 million parameters) uses knowledge distillation to compress BERT while retaining 97% of its performance, providing insights into how distillation affects fake news detection. ALBERT [22] (11.69 million parameters) employs parameter sharing and factorized embeddings to achieve dramatic size reduction while maintaining sophisticated language understanding. MobileBERT [21] (24.58 million parameters) focuses specifically on mobile deployment optimization, using architectural modifications designed for resource-constrained environments. TinyBERT [20] (14.35 million parameters) represents aggressive compression through knowledge distillation, achieving substantial efficiency gains while maintaining practical performance levels.

3.5 Model Training

We implemented model-specific training configurations that leverage each architecture's strengths while ensuring fair comparison through consistent evaluation steps. Our approach recognized that different models require different optimization strategies to achieve their best performance.

For traditional ML models, we conducted systematic hyperparameter tuning using the validation set to find optimal regularization parameters. Logistic Regression was tuned across different regularization strengths, while Random Forest was optimized for the number of estimators and tree depth parameters. This tuning process ensured that these baseline models achieve their best possible performance for fair comparison with more sophisticated models.

Transformer model training required more nuanced configuration approaches.

Model	Learning Rate	Warmup Steps	Batch Size	Epochs
DistilBERT	5e-5	500	16	3
ALBERT	2e-5	1000	16	5
MobileBERT	3e-5	500	16	4
TinyBERT	1e-4	500	16	3

Table 1: Training configurations for Transformer models.

Table 1 shows the hyperparameter configurations for the transformer models. All transformer models incorporated early stopping based on validation F1 score to prevent overfitting, weight decay of 0.01 for regularization, and gradient clipping to ensure training stability.

3.6 Evaluation Methods

Our evaluation methodology goes beyond traditional accuracy metrics to provide a comprehensive assessment of model suitability for real-world deployment. We developed a three-dimensional evaluation framework that considers performance accuracy, computational efficiency, and generalization capabilities.

Performance Metrics:

- **Accuracy:** The percentage of correctly classified articles (both real and fake)
- **F1 Score:** The harmonic mean of precision and recall
- **Precision:** The proportion of predicted fake news that were actually fake
- **Recall:** The proportion of actual fake news that were correctly identified

The performance evaluation was done on our held-out test set. However, we recognized that performance on familiar data represents only one aspect of model quality.

Efficiency Metrics:

- **Training time:** Time required for model training in minutes
- **Inference time:** Time needed to process and classify each sample, measured in milliseconds per article

- **Model size:** Number of parameters and storage requirements in MB
- **Memory usage:** Peak memory consumption during inference in MB

Our generalization evaluation represents a novel contribution that addresses a critical gap in fake news detection research. We compiled external datasets containing verified real news articles and AI-generated fake news that differ significantly from the WELFake training distribution. This evaluation reveals how models perform when encountering new patterns of misinformation, providing crucial insights into their robustness and practical utility. The comprehensive evaluation framework enables us to identify not just which models perform best on standard benchmarks, but which approaches offer the most practical value across different deployment scenarios. This multi-dimensional perspective provides actionable insights for researchers and practitioners working to implement effective fake news detection systems.

3.7 Methodology Overview

Figure 1 provides a comprehensive overview of our research methodology, illustrating the complete workflow from dataset preparation through comparative analysis.

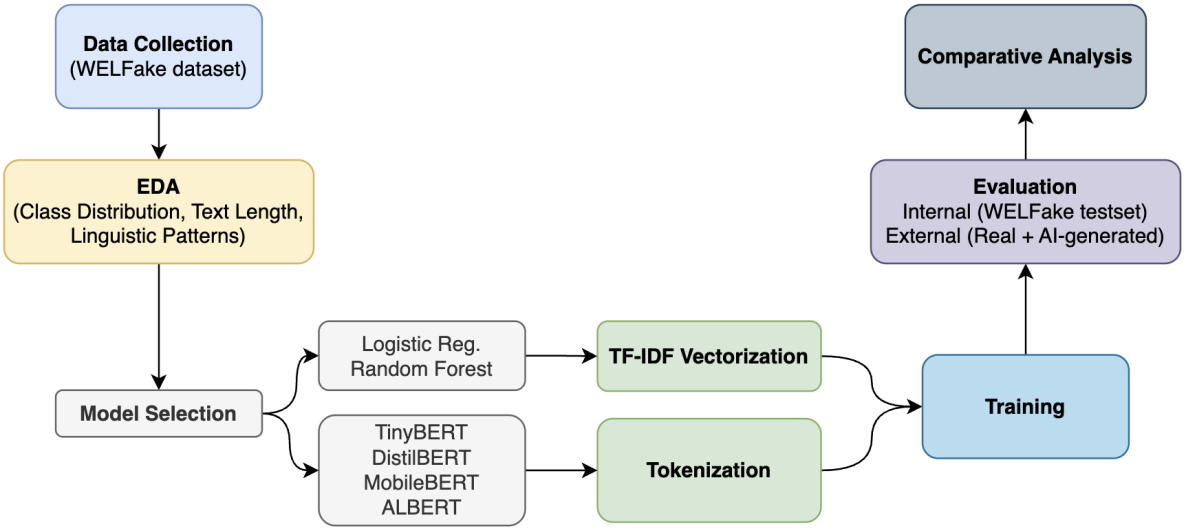

Figure 1: Methodology framework showing the complete workflow for comparing lightweight models in FND. The framework encompasses dataset preparation, model selection (traditional ML and transformer-based), training, and three-dimensional evaluation covering performance, resource consumption, and generalization capabilities.

4 Results and Discussion

In this section, we present a comprehensive analysis of both the performance and computational efficiency metrics for all evaluated models. We particularly focus on the trade-offs between accuracy and resource requirements, which is critical for real-world deployment scenarios.

4.1 Performance on WELFake Test Set

All evaluated models demonstrated strong performance on the WELFake test set, with transformer models achieving exceptional accuracy levels. ALBERT led the transformer models with 99.66% accuracy, followed closely by DistilBERT at 99.65%, MobileBERT at 99.61%, and TinyBERT at 99.28%. Traditional machine learning approaches also performed well, with Logistic Regression achieving 95.96% accuracy and Random Forest reaching 95.61%. The narrow margins between top-performing transformer models suggest that architectural optimizations have successfully addressed the core fake news detection challenge for in-domain data. The 99.66% accuracy achieved by ALBERT represents near-perfect classification performance, correctly identifying fake news in over 996 out of every 1000 articles. This level of performance demonstrates that parameter sharing and factorized embeddings can maintain sophisticated language understanding while dramatically reducing model complexity. Detailed error analysis reveals that transformer models exhibit balanced classification behavior across both classes. DistilBERT, for example, showed a false positive rate of 0.21% and false negative rate of 0.49%, indicating that the model does not systematically favor either real or

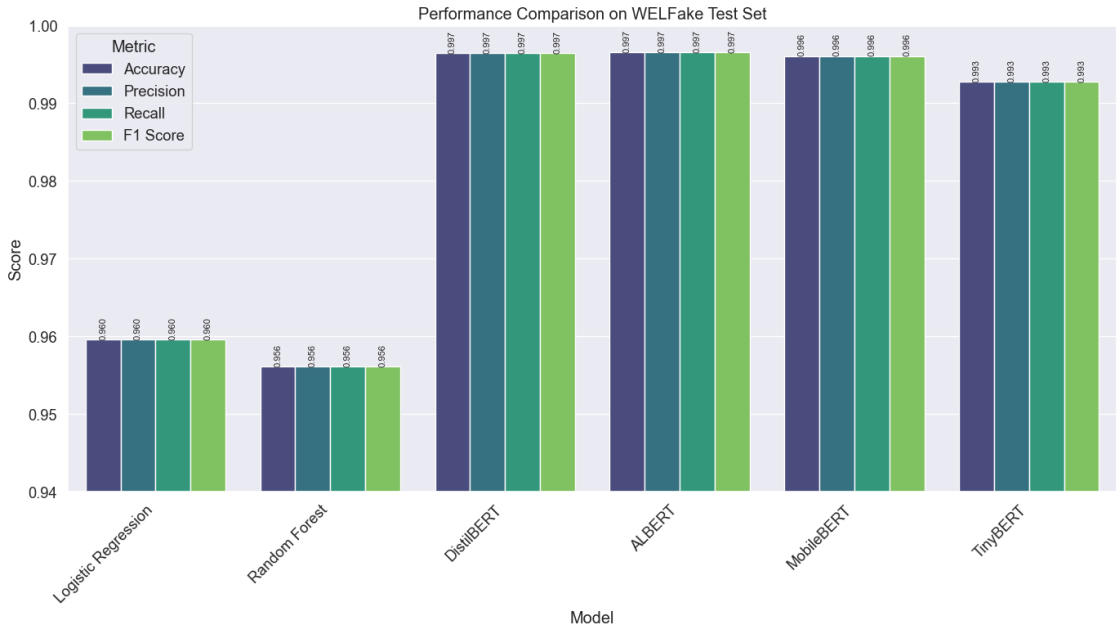

Figure 2: Performance metrics comparison on the WELFake test set. All transformer models achieve >99% accuracy, with ALBERT leading at 99.66%.

fake news classification. This balance is crucial for fair content moderation systems where both types of errors carry significant consequences. The performance gap between transformer and traditional ML models, while noticeable, is smaller than might be expected given the dramatic difference in model complexity. The approximately 4 percentage point difference between ALBERT (99.66%) and Logistic Regression (95.96%) raises important questions about the cost-benefit ratio of transformer deployment for this specific task, particularly when considering computational requirements.

4.2 Efficiency Analysis

Model	Parameters	Size (MB)	Inference (ms)	Training (min)
Logistic Regression	10,001	8.00	0.463	0.14
Random Forest	1,392,006 nodes	25.00	48.234	41.50
DistilBERT	66,955,010	255.41	9.20	98.83
ALBERT	11,685,122	44.58	17.65	230.91
MobileBERT	24,582,914	93.78	8.97	129.33
TinyBERT	14,350,874	54.74	14.49	14.87

Table 2: Comparison of model efficiency metrics across all evaluated models.

The efficiency metrics in Table 2 reveal significant differences between model families. Traditional machine learning models demonstrate remarkable computational efficiency, with Logistic Regression requiring just 0.463 milliseconds per inference—over 150 times faster than the most efficient transformer model. The training time difference is even more striking, with Logistic Regression completing hyperparameter tuning and final training in mere seconds compared to hours for transformer models. Among transformer models, TinyBERT emerges as the clear efficiency leader, requiring only 14.87 minutes for training and maintaining reasonable inference speed. This represents a substantial improvement over DistilBERT's 98.83-minute training time and ALBERT's 230.91-minute training requirement. The efficiency gains achieved by TinyBERT's knowledge distillation approach demonstrate that aggressive model compression can maintain practical performance while dramatically reducing resource requirements. Model size analysis shows that ALBERT achieves the smallest footprint among transformers at 44.58 MB through its parameter sharing architecture, representing an 89.4% reduction compared to comparable full-scale models. TinyBERT follows at 54.74 MB, while DistilBERT requires 255.41 MB. These size differences have significant implications for mobile deployment and edge computing scenarios where storage and memory constraints are critical factors. Memory usage patterns during inference reveal

additional efficiency considerations. Traditional ML models operate with minimal memory overhead, while transformer models require substantially more memory for their attention mechanisms and larger parameter sets. These differences affect the number of concurrent requests that can be processed and influence overall system scalability.

4.3 Generalization to External Datasets

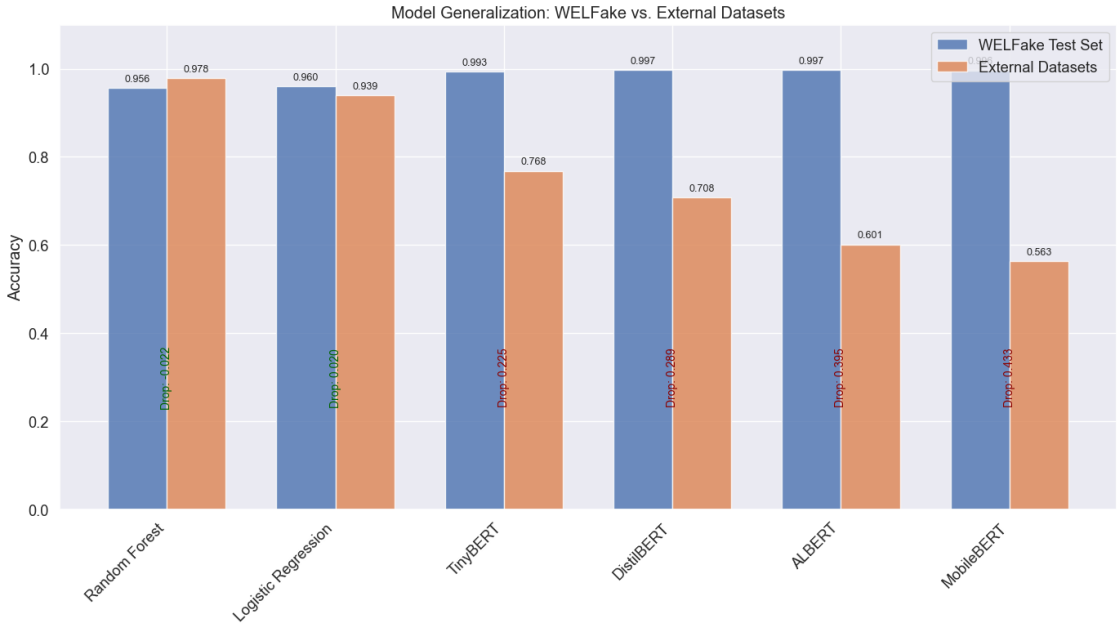

Figure 3: Model generalization: Accuracy comparison between WELFake test set and external datasets. Traditional ML models show superior generalization capabilities.

The generalization evaluation reveals perhaps the most surprising and significant finding of our study. When tested on external datasets containing different types of misinformation not represented in the training data, traditional machine learning models demonstrated dramatically superior robustness compared to transformer approaches. Figure 3 illustrates this disparity. Random Forest achieved 97.79% accuracy on external data, actually improving over its WELFake performance, while Logistic Regression maintained 93.94% accuracy, showing minimal performance degradation. This remarkable generalization capability suggests that TF-IDF features capture fundamental linguistic patterns that transfer effectively across different misinformation contexts and creation methods. In stark contrast, transformer models showed substantial performance deterioration when encountering unfamiliar misinformation patterns. MobileBERT's accuracy dropped to 56.29%, ALBERT fell to 60.14%, and DistilBERT decreased to 70.75%. TinyBERT demonstrated the best generalization among transformers but still declined to 76.81% accuracy. These performance drops represent failures to detect between 40-90% of fake news articles in the external datasets. The error pattern analysis reveals that transformer models exhibit extremely high false negative rates on external data, meaning they frequently fail to identify new patterns of misinformation as fake news. MobileBERT missed 87.4% of fake news articles, ALBERT missed 79.7%, and DistilBERT missed 58.5%. This systematic bias toward classifying unfamiliar content as real news represents a critical vulnerability for practical deployment scenarios. These generalization results suggest that transformer models may be learning specific stylistic and linguistic patterns present in the WELFake dataset rather than developing generalizable understanding of deception indicators. The sophisticated pattern recognition capabilities that enable excellent in-domain performance may paradoxically hinder adaptation to new misinformation strategies and content types.

4.4 Performance-Efficiency Trade-offs

Figure 4 visualizes the critical trade-off between performance on external datasets and computational efficiency. The ideal model would appear in the upper-left corner (high accuracy, low inference time). Traditional ML models occupy this desirable position, offering both excellent generalization and efficiency. TinyBERT represents the best compromise among transformer models, with reasonably good external performance and significantly better efficiency than other transformers.

Figure 4: Performance vs. efficiency trade-off showing accuracy on external datasets plotted against inference time (log scale), with bubble size representing model size.

This analysis challenges the common assumption that more complex models always deliver better results. For fake news detection, our findings suggest that simpler models may provide more robust and practical solutions in many real-world scenarios, especially when generalization to new patterns of misinformation is crucial.

4.5 Comprehensive Model Comparison

Model	WELFake Accuracy	External Accuracy	Accuracy Drop	Training Time (min)	Inference Time (ms)	Model Size (MB)
Logistic Regression	95.96%	93.94%	2.02%	0.14	0.463	8.0
Random Forest	95.61%	97.79%	-2.18%	41.5	48.234	25.0
ALBERT	99.66%	60.14%	39.52%	230.91	17.65	44.58
DistilBERT	99.65%	70.75%	28.90%	98.83	9.20	255.41
MobileBERT	99.61%	56.29%	43.32%	129.33	8.97	93.78
TinyBERT	99.28%	76.81%	22.47%	14.87	14.49	54.74

Table 3: Comprehensive Comparison

Table 3 shows a comprehensive comparison of all evaluated models across performance, generalization, and efficiency dimensions. Negative accuracy drop indicates improved performance on external data.

5 Conclusion, Limitations, and Future Work

Our comprehensive evaluation of lightweight transformer models and traditional machine learning approaches for fake news detection reveals a nuanced performance landscape that challenges conventional assumptions about model selection. While transformer models achieve exceptional accuracy on familiar data patterns, their significant generalization limitations and computational requirements suggest that traditional approaches may offer superior practical value for many real-world deployments. The key finding that traditional ML models demonstrate superior generalization to new misinformation patterns represents a fundamental insight for the field. Random Forest's ability to maintain 97.79%

accuracy on external datasets while transformer models experienced dramatic performance drops suggests that simpler feature representations may capture more transferable patterns of deception. This finding has significant implications for developing robust fake news detection systems that must adapt to evolving misinformation tactics. Among transformer models, our results identify clear efficiency leaders and optimal use cases. TinyBERT emerges as the most balanced option, offering 99.28% accuracy with 14.87-minute training time and reasonable inference speed. ALBERT demonstrates that parameter sharing can achieve the smallest model size (44.58 MB) without performance compromise, while maintaining top-tier accuracy (99.66%). These efficiency improvements make transformer deployment feasible in previously prohibitive scenarios.

5.1 Limitations

Our study has several important limitations that should be considered when interpreting results. The external dataset used for generalization evaluation, while carefully selected to represent different misinformation patterns, is relatively small compared to WELFake and may not capture the full spectrum of real-world misinformation diversity. Additionally, our efficiency measurements were conducted on specific hardware configurations and may vary significantly in different computational environments. The evaluation focuses on English-language news articles from specific sources and time periods, which may limit generalizability to other languages, cultural contexts, or emerging misinformation formats like multimedia content or social media posts. We also did not explore hybrid approaches that might combine the complementary strengths of different model families, which could potentially address some of the identified limitations. Our generalization evaluation, while novel and important, represents only one approach to assessing model robustness. Other evaluation methods, such as adversarial testing or domain adaptation scenarios, might reveal different patterns of model behavior and provide additional insights into practical deployment considerations.

5.2 Future Work

Several promising research directions emerge from our findings. Investigating continual learning approaches could address transformer models' generalization limitations by enabling adaptation to evolving misinformation patterns without catastrophic forgetting of previous knowledge. Such approaches might combine the sophisticated pattern recognition of transformers with the robust generalization of traditional methods. Exploring ensemble methods that systematically combine traditional ML and transformer models represents another valuable direction. Our results suggest that these approaches capture different aspects of the fake news detection problem, and carefully designed ensemble strategies might leverage their complementary strengths to achieve both high accuracy and robust generalization. Advanced optimization techniques including quantization, pruning, and specialized hardware deployment could further improve transformer efficiency while maintaining performance levels. Additionally, investigating domain adaptation techniques specifically for fake news detection could help transformer models better generalize across different types of misinformation sources and creation methods. The development of more sophisticated evaluation protocols for assessing generalization capabilities would benefit the entire field. This might include standardized external datasets representing different misinformation creation methods, cultural contexts, and temporal periods, enabling more comprehensive assessment of model robustness across the research community.

5.3 Practical Implications

Our findings have immediate practical implications for organizations developing fake news detection systems. The superior generalization capabilities of traditional ML approaches suggest they should be strongly considered for production systems. The efficiency advantage of these approaches also makes them attractive for resource-constrained deployments. For organizations with sufficienct computational resources and well-defined content domains, transformer models can provide exceptional accuracy when regular retraining is feasable. However, the generalization gap identified in our study suggest that such systems should incorporate robust fallback mechanisms and continuous monitoring for performance degredation when encountering new misinformation patterns.

References

[1] Sergio Muñoz and Carlos Á. Iglesias. Exploiting content characteristics for explainable detection of fake news. *Big Data and Cognitive Computing*, 8(10):129, 2024.

[2] Despoina Mouratidis, Andreas Kanavos, and Katia Kermanidis. From misinformation to insight: Machine learning strategies for fake news detection. *Information (2078-2489)*, 16(3), 2025.

[3] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.

[5] Linmei Hu, Siqi Wei, Ziwang Zhao, and Bin Wu. Deep learning for fake news detection: A comprehensive survey. *AI open*, 3:133–155, 2022.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[7] Jawaher Alghamdi, Yuqing Lin, and Suhuai Luo. Unveiling the hidden patterns: A novel semantic deep learning approach to fake news detection on social media. *Engineering Applications of Artificial Intelligence*, 137:109240, 2024.

[8] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *science*, 359(6380):1146–1151, 2018.

[9] Konstantinos I Roumeliotis, Nikolaos D Tselikas, and Dimitrios K Nasiopoulos. Fake news detection and classification: A comparative study of convolutional neural networks, large language models, and natural language processing models. *Future Internet*, 17(1), 2025.

[10] Mohammed E. Almandouh, Mohammed F Alrahmawy, Mohamed Eisa, Mohamed Elhoseny, and AS Tolba. Ensemble based high performance deep learning models for fake news detection. *Scientific Reports*, 14(1):26591, 2024.

[11] Abu Sarwar Zamani, Aisha Hassan Abdalla Hashim, Sara Saadeldeen Ibrahim Mohamed, and Md Nasre Alam. Optimized deep learning techniques to identify rumors and fake news in online social networks. *Journal of Computational and Cognitive Engineering*, 2022.

[12] Aswini Thota, Priyanka Tilak, Simrat Ahluwalia, and Nibrat Lohia. Fake news detection: a deep learning approach. *SMU Data Science Review*, 1(3):10, 2018.

[13] Mohammed Al-Alshaqi, Danda B Rawat, and Chunmei Liu. Ensemble techniques for robust fake news detection: Integrating transformers, natural language processing, and machine learning. *Sensors*, 24(18):6062, 2024.

[14] Kayato Soga, Soh Yoshida, and Mitsuji Muneyasu. Exploiting stance similarity and graph neural networks for fake news detection. *Pattern Recognition Letters*, 177:26–32, 2024.

[15] Xinyi Zhou and Reza Zafarani. A comprehensive survey on fake news detection with deep learning. *IEEE Transactions on Computational Social Science*, 8(2):1043–1060, 2021.

[16] Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification*, pages 85–90, 2018.

[17] Junxiao Xue, Yabo Wang, Yichen Tian, Yafei Li, Lei Shi, and Lin Wei. Detecting fake news by exploring the consistency of multimodal data. *Information Processing & Management*, 58(5):102610, 2021.

[18] Ajay Kumar and James W Taylor. Feature importance in the age of explainable ai: Case study of detecting fake news & misinformation via a multi-modal framework. *European Journal of Operational Research*, 317(2):401–413, 2024.

[19] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[20] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, 2020.

[21] Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. Mobilebert: a compact task-agnostic bert for resource-limited devices. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2158–2170, 2020.

[22] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations, 2020.