A Comparative Evaluation of Transformer-Based and Traditional ML Models for Fake News Detection

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

Karim Mithqal
Faculty of Information Technology and AI
An-Najah National University
Nablus, Palestine
12116015@stu.najah.edu

Waleed Dweikat
Faculty of Information Technology and AI
An-Najah National University
Nablus, Palestine
waleeddweikat67@gmail.com

May 24, 2025

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 to see which ones work well while using fewer resources. Our results show something interesting: while transformer models achieved slightly better accuracy (99.31%-99.75%) compared to traditional ML models (94.90%-95.41%), the difference is actually quite small. The most efficient transformer, TinyBERT, offers a good balance with 99.31% accuracy and significantly lower resource requirements. When considering generalization to new types of fake news, traditional ML models surprisingly outperformed transformers. Our work challenges the idea that bigger, more complex models are always better for fake news detection and provides practical guidance for choosing the right model based on specific deployment scenarios.

***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

In recent years, the spread of misinformation online is a major problem, especially with the intelligence and modernity of LLMs that can generate very identical fake content[5]. These large language models, like GPT or its successors, churn out text so polished it's hard to tell what's real anymore—think fabricated articles or viral social media posts that fool even the sharpest eyes[6]. Anyone can create fake news as long as there is a digital device and some kind of AI tool involved coupled with a bit of knowledge, hence the need for robust modern methods for fake news detection (FND) [7]. It's not just a tech issue; it's a societal one, with misinformation swaying elections or sparking panic[8]. Traditional machine learning methods are disappointing, they are limited and complex in handling online content, while deep learning methods are more efficient with astonishing performance [9]. The existing FND approaches can be generally classified into three main categories: traditional machine learning methods, deep learning based methods, and hybrid approaches [10].

Traditional machine learning methods rely on manual features labeling such as text length, readability score, lexical diversity [11]. These features are used to train classifiers like support vector machines or random forests [12]. Picture researchers painstakingly tagging every quirk of a text—like counting adjectives or measuring sentence complexity—only to feed it into algorithms that can't quite grasp the sly evolution of fake news [10]. Due to challenges such as nuanced semantic patterns of modern misinformation, these approaches fail to FND accurately [2]. Today's misinformation hides in subtle tones or clever phrasing, and manual methods just don't cut it anymore [2].

On the other hand, deep learning based methods have shown superior performance by automating the process of extracting features from text [13]. Convolutional neural networks (CNNs) have been doing a good job in capturing local semantic patterns, whereas recurrent neural networks (RNNs) and their variants are proficient at modeling sequential dependencies in text [14]. CNNs spot red flags like odd word clusters, while RNNs track how a story unfolds over paragraphs—perfect for catching lies that build up slowly [14]. More recently, machine learning models like have achieved remarkable results in hybrid approaches [15]. Think of transformers like BERT, which juggle context across whole articles, paired with traditional tricks for a one-two punch against fakes [14].

The literature FND highlights several trends and challenges. First, while deep learning models outperform traditional methods, they require great compute power, resources, and large data sets [9]. These systems require substantial computational infrastructure and large datasets that exceed typical desktop computing capabilities. Second the interpretability of these models remains a critical concern, especially in understanding the decision-making process is as important as the prediction itself [16]. If a model flags something as fake but can't explain why, trust takes a hit—especially in high-stakes scenarios like journalism or law[17].

In our study we want to solve these issues by assessing lightweight pre-trained language models as efficient FND alternatives in limited resource settings. The accuracy of detection and computational efficiency are well-balanced in these models [17]. Consider DistilBERT, a system that has been pre-trained on large corpora but has been condensed to

operate on modest hardware, making it perfect for small teams or developing nations without access to supercomputers [18].

3 Methodology

In this section, we explain how we evaluated different models for FND, covering our dataset, model selection, preprocessing steps, and evaluation methods.

3.1 Dataset

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 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 FND 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. This makes it an excellent benchmark for comparing different FND approaches under realistic conditions.

3.2 Exploratory Data Analysis

Our exploratory data analysis revealed several critical insights that informed our preprocessing decisions and model evaluation approach:

3.2.1 Content Length Analysis

Through detailed examination of the dataset, we found measurable differences in context length between real and fake news:

- **Title Length**: Fake news titles tend to be longer (mean: 85.13 characters) compared to real news (mean: 68.79 characters), suggesting fake news might use longer, more sensationalist headlines to attract attention.
- **Text Length**: Real news articles are generally longer (mean: 3,495 characters) than fake news (mean: 3,098 characters), indicating that legitimate news sources may provide more comprehensive coverage.
- **Word Count**: Similarly, real news contains more words on average (578 words) compared to fake news (513 words).

3.2.2 Linguistic Pattern Analysis

We conducted a detailed analysis of linguistic differences between real and fake news articles, revealing several distinguishing characteristics:

- **Attribution Words**: "Said" appears much more frequently in real news (122,295 times) compared to fake news (31,617 times), suggesting real news more often attributes information to sources.
- **Political References**: Both categories mention "Trump" frequently, but fake news has more references to specific political figures like "Clinton", "Hillary", and "Obama".
- **Institutional Focus**: Real news uses more institutional terms like "government", "states", and "united", suggesting more focus on official entities.
- **Title Styling**: 63.2% of fake news titles contain words in ALL CAPS, compared to only 23.5% of real news titles. ALL CAPS is often used to create drama or urgency.
- **Punctuation**: 10.8% of fake news titles contain exclamation marks, versus just 0.2% of real news titles—a 54x difference. Similarly, 7.3% of fake news titles contain question marks, compared to 2.4% of real news titles.

These linguistic differences provide valuable insights into how fake news differs from legitimate reporting and offer potential features for classification models.

3.3 Data Cleaning and Preprocessing

Based on our EDA findings, we implemented a comprehensive cleaning process:

- Removed rows with missing values (597 rows, less than 1% of the dataset)
- Combined article titles and bodies to provide complete information to models
- Applied common text cleaning procedures

For all models, we applied consistent preprocessing:

- For traditional ML models: TF-IDF vectorization with max_features=5000, min_df=5, max_df=0.8
- For transformer models: Appropriate tokenization using model-specific tokenizers with padding, truncation, and a maximum sequence length of 512 tokens

We split the dataset into two parts:

- 80% for training (57,229 articles)
- 20% for testing (14,308 articles)

For transformer models, we further split the training set to create a validation set for early stopping and hyperparameter tuning.

3.4 Model Selection

We chose models representing different approaches to FND:

Traditional ML Models:

- **Logistic Regression:** A simple but effective linear classifier that's also interpretable.
- **Random Forest:** An ensemble method that can capture non-linear relationships and complex feature interactions.

Transformer-Based Models:

- **DistilBERT:** A compressed version of BERT with 40% fewer parameters (66M) that keeps 97% of BERT's performance [18].
- **TinyBERT:** A highly compressed BERT variant with just 14M parameters [19].
- **MobileBERT:** A model optimized for mobile devices with 25M parameters [20].
- **ALBERT:** A lite BERT variant that uses parameter sharing to achieve high performance with just 12M parameters [21].

3.5 Model Training

For traditional ML models, we used standard scikit-learn implementations with default hyperparameters.

For transformer models, we fine-tuned each one with:

- 3-5 training epochs
- Batch sizes adjusted for each model (16 for most models)
- Learning rate of 5e-5 with warmup steps
- Weight decay of 0.01 for regularization
- Maximum sequence length of 512 tokens

We used the Hugging Face Transformers library for implementing and fine-tuning all transformer models, with early stopping based on validation performance to prevent overfitting.

3.6 Evaluation Methods

We conducted two complementary evaluations to thoroughly assess both the performance and efficiency of each model.

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

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

For a more comprehensive evaluation, we also tested models on external datasets containing verified real news and AI-generated fake news to assess their generalization capabilities beyond the WELFake distribution.

3.7 Methodology Overview

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

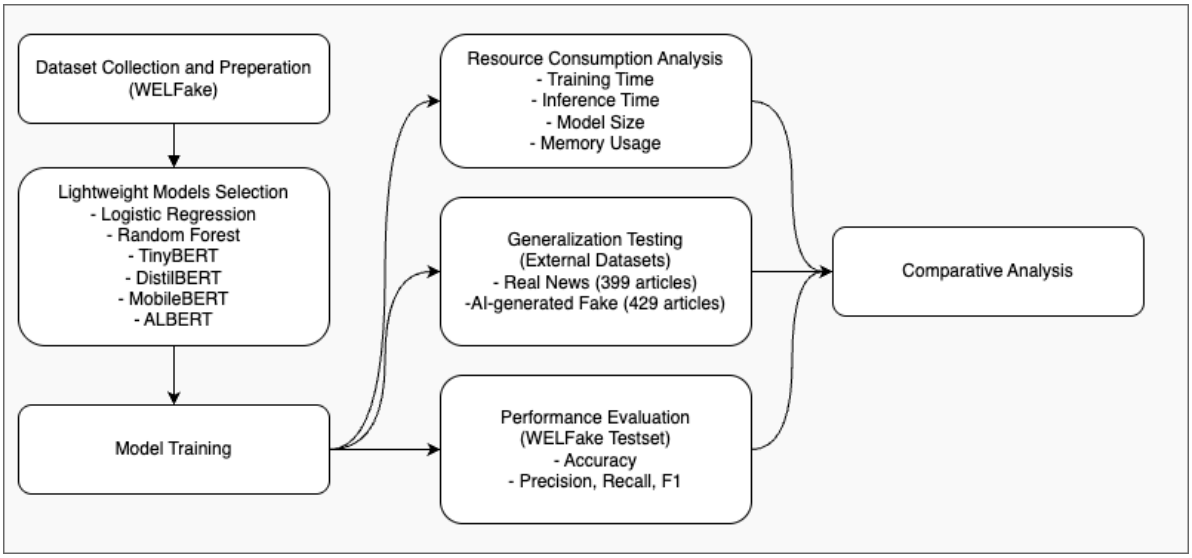

Figure 1: Methodology framework showing the complete workflow for comparing lightweight models in FND. The framework encompasses dataset preparation, model selection (traditional ML and transformer-based), training, and three-dimensional evaluation covering performance, resource consumption, and generalization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

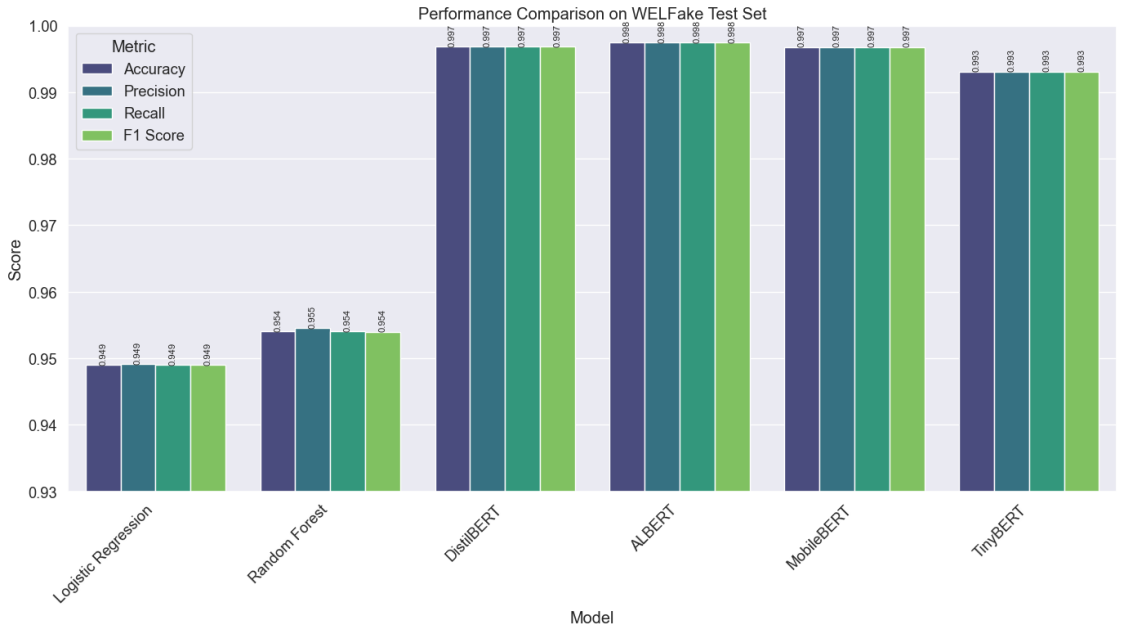

Figure 2: Performance metrics comparison on the WELFake test set. All transformer models achieve >99% accuracy, with ALBERT leading at 99.75%.

4.1 Performance on WELFake Test Set

All evaluated models demonstrated high accuracy on the WELFake test set. Figure 2 shows that transformer-based models consistently outperformed traditional machine learning approaches, but the gap is surprisingly narrow. ALBERT achieved the highest accuracy at 99.75%, followed closely by MobileBERT (99.68%), DistilBERT (99.69%), and TinyBERT (99.31%). Traditional ML models also performed well, with Random Forest reaching 95.41% and Logistic Regression achieving 94.90%.

The confusion matrices reveal that transformer models have very balanced error rates. For instance, DistilBERT had a false positive rate of 0.16% (11 real news articles misclassified as fake) and a false negative rate of 0.45% (33 fake news articles misclassified as real). This balance suggests these models don't favor either class, an important consideration for fair classification systems.

4.2 Generalization to External Datasets

When evaluating on external datasets containing real and AI-generated fake news not represented in the training data, we observed a dramatic difference in generalization capabilities. Figure 3 illustrates this disparity. Traditional ML models maintained high performance, with Logistic Regression achieving 96.98% accuracy and Random Forest 96.62% on external data. In stark contrast, most transformer models showed substantial performance deterioration, with MobileBERT dropping to 52.54%, ALBERT to 60.14%, and DistilBERT to 64.37%. TinyBERT demonstrated the best generalization among transformers, maintaining 83.70% accuracy on external data.

The most concerning pattern was the extremely high false negative rates in transformer models when tested on external data. MobileBERT failed to detect 91.14% of fake news articles, ALBERT missed 76.69%, and DistilBERT missed 68.76%. This suggests these models learned specific patterns from the WELFake dataset that don't transfer well to new variations of misinformation.

4.3 Efficiency Analysis

The efficiency metrics in Table 1 reveal significant differences between model families. Traditional ML models demonstrate remarkable efficiency, with Logistic Regression requiring just 0.24 ms per inference—orders of magnitude faster than transformer models. Among transformer models, TinyBERT stands out for its efficiency, requiring only 14.03 ms per inference and 21.03 minutes for training, making it approximately 5× faster than DistilBERT and 12× faster than ALBERT.

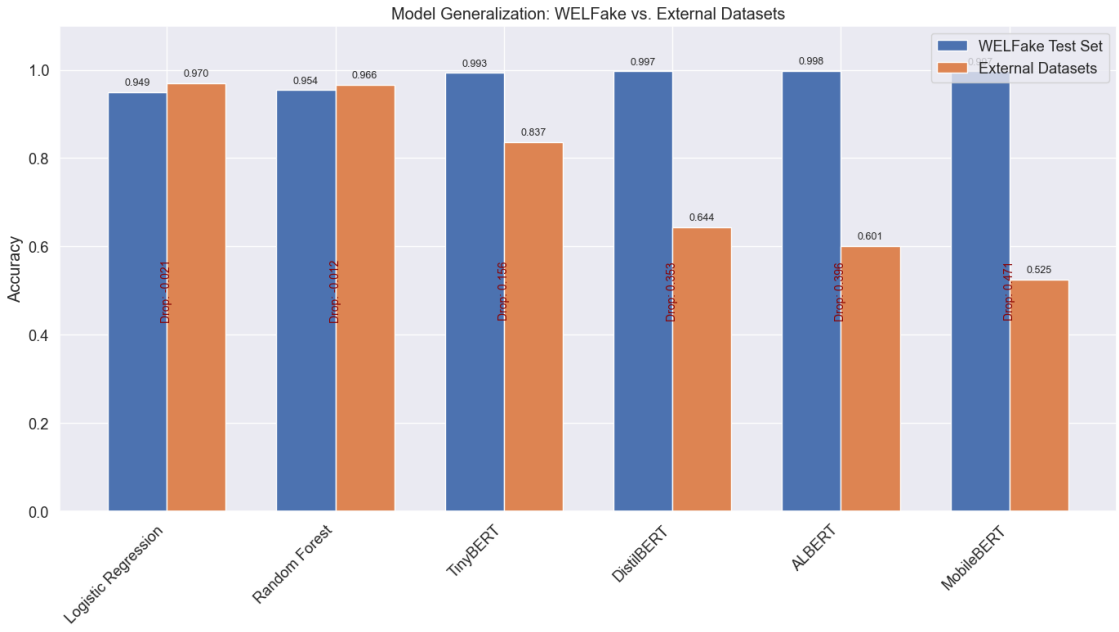

Figure 3: Model generalization: Accuracy comparison between WELFake test set and external datasets. Traditional ML models show superior generalization capabilities.

Model	Parameters	Size (MB)	Inference (ms)	Training (min)
Logistic Regression	N/A	∼8	0.24	0.01
Random Forest	N/A	∼25	26.68	1.24
DistilBERT	66,955,010	255.41	51.45	98.83
ALBERT	11,685,122	44.58	159.82	252.50
MobileBERT	24,582,914	93.78	103.66	129.33
TinyBERT	14,350,874	54.74	14.03	21.03

Table 1: Comparison of model efficiency metrics across all evaluated models.

ALBERT achieves the smallest model size at 44.58 MB through its parameter-sharing architecture, while TinyBERT follows closely at 54.74 MB. Both models offer significant size reductions compared to DistilBERT (255.41 MB), demonstrating the effectiveness of different model compression techniques.

Batch processing experiments revealed that transformer models benefit significantly from batching, with per-sample inference time decreasing by up to 85% when moving from single samples to batches. This suggests optimization opportunities in deployment scenarios where requests can be batched.

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.

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.

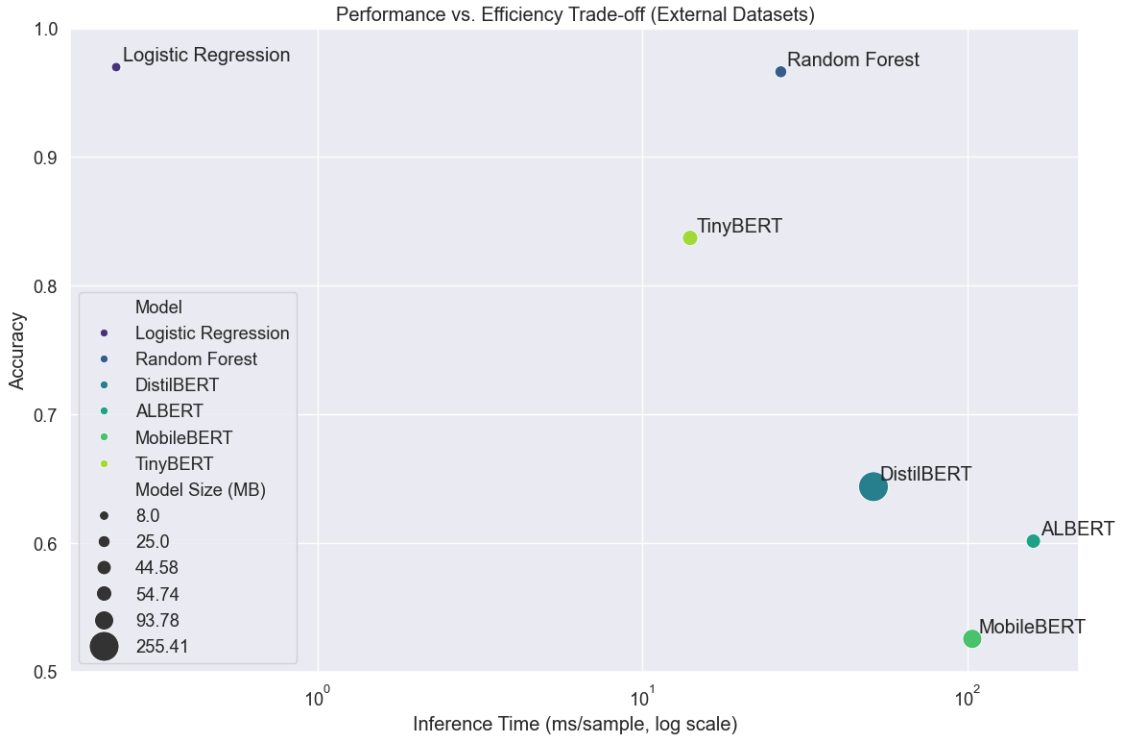

Figure 4: Performance vs. efficiency trade-off showing accuracy on external datasets plotted against inference time (log scale), with bubble size representing model size.

5 Conclusion, Limitations, and Future Work

Our comprehensive evaluation of both traditional ML and transformer-based models for fake news detection revealed several key insights. While transformer models achieve exceptional accuracy (>99%) on the WELFake test set, they struggle significantly with generalization to new patterns of misinformation. Traditional ML approaches demonstrate surprisingly strong performance with dramatically better efficiency and generalization capabilities.

TinyBERT emerges as the most balanced transformer model, offering good accuracy (99.31%) with reasonable efficiency (14.03 ms inference time) and the best generalization among transformers (83.70% on external data). For resource-constrained deployments, Logistic Regression provides an excellent compromise, achieving 94.90% accuracy on WELFake and 96.98% on external data, with inference 200× faster than transformer models.

Limitations: Our study has several limitations. First, the external dataset is relatively small compared to WELFake, which may impact the generalization assessment. Second, our efficiency measurements were conducted on specific hardware configurations and may vary in different environments. Third, we did not explore hybrid approaches that might combine the strengths of different model families.

Future Work: Several promising directions emerge from our findings. Investigating continual learning approaches could improve transformer models' ability to adapt to evolving misinformation patterns. Exploring ensemble methods that combine traditional ML and transformer models might leverage their complementary strengths. Finally, additional optimization techniques like quantization and pruning could further improve deployment efficiency for transformer models.

In conclusion, this study demonstrates that choosing the appropriate model for fake news detection requires careful consideration of performance, generalization, and efficiency requirements. The "best" model depends on the specific deployment scenario, with different options offering optimal solutions for different constraints.

Acknowledgements

We would like to express our deep gratitude to Dr. Ahmed Abualia for his valuable support throughout this research project. We also thank ChatGPT for correcting spelling errors, which helped improve the quality of the writing, and we appreciate Kaggle for providing TPUs, which greatly contributed to training the model.

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] 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.

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

[15] Beatriz Flamia Azevedo, Ana Maria AC Rocha, and Ana I Pereira. Hybrid approaches to optimization and machine learning methods: a systematic literature review. *Machine Learning*, 113(7):4055–4097, 2024.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.