Optimizing Fake News Detection in Resource-Constrained Devices Using Transformer Distillation

*Victor T. Odumuyiwa, Dayo O. Aderemi
Department of Computer Sciences, University of Lagos, Lagos, Nigeria
vodumuyiwa@unilag.edu.ng|aderemi.dayo.o@gmail.com

Received: 14-SEP-2024; Reviewed: 23-NOV-2024; Accepted: 12-DEC-2024
https://dx.doi.org/10.4314/fuoyejet.v9i4.6

ORIGINAL RESEARCH

Abstract—In today's digital era, stopping fake news from spreading, especially on social media is crucial, given its potential to undermine public trust, influence elections, and cause social unrest. The quick sharing of information, enabled by technological improvements and the decreasing size of computational devices, highlights the need for effective techniques for detecting fake news. This study looks into the possibility of using Transformer Distillation to create a small but precise model for detecting fake news. The research compares the performance of TinyBERT, a simplified version of the base BERT model, with other well-known BERT versions, such as BERT Base, DistillBERT, and MobileBERT. The feasibility of using a resized BERT model for fake news detection, especially on resource-constrained devices like mobile phones, is carefully examined by analysing key variables including model size, training time, and accuracy. The results show that TinyBERT performs admirably accurate given its 80% smaller size compare to BertBase Model, making it a viable option for preventing the spread of false information in the era of portable electronics. The present study enhances the continuous endeavours to curb the dissemination of false information by offering a proficient and effective detection system.

Keywords— BERT, DistillBERT, Fake News Detection, MobileBERT, TinyBERT, Transformer.

1 INTRODUCTION

The digital era has brought in a new idea of sharing information and communication. Although this transformation has brought in many profits, there has also been a concerning increase in the spread of false information and fake news. Fake news can sway public opinion, spread panic, and even interfere with political processes. The problem has been made worse by the widespread use of social media and instant messaging apps, which have allowed fake news to move at an increasing rate (Tandoc et al., 2018).

The prevalence of false news has emerged as a significant challenge, and people are finding it hard to identify it. The spread of fake news has serious repercussions for public opinion, society, and even democratic processes because of its false narratives that replace facts. The challenge of separating real from false news in textual and multimedia formats has become more complicated, pushing the limits of detection techniques that rely on humans or machines (Vosoughi et al., 2018). The distinctions between "fake news," "disinformation," and "misinformation" make the task of detecting fake news more difficult. These terms cover a wide range of false or misleading information, ranging from intentionally deceptive to unintentional errors. The problem lies in knowing the motives and intents behind the information. While misinformation frequently results from unintentional mistakes, disinformation is deliberately spread to manipulate (Tandoc et al., 2018). The boundaries between fact and falsehood are blurred by this complex continuum, making it difficult for automatic and human fact-checking methods to determine the veracity of the material.

The modern information environment is characterised by an alarming spread of disinformation, largely due to the ubiquitous presence of social media platforms. These online forums serve as rich environments for the spread of misleading stories, frequently confusing the distinction between reality and fiction. Enhanced by algorithms intended to optimise user interaction, false content spreads quickly across the internet, disguising itself as factual data and ensnaring unsuspecting users.

Fake news is particularly appealing because of its cleverly created appearance, which closely resembles the style and format of real journalism. When combined with the relentless pace at which information spreads online, it becomes more and more difficult for people to separate fact from falsehood when they are overloaded with information. Traditional defences, such as fact-checking and editorial control, against false information are routinely bypassed or rendered ineffective in the digital space. As a result, the public becomes more vulnerable to manipulation and exploitation since false information enters public discourse, biassed opinions and affect how decisions are made. Thus, in the era of digital communication and information overload, identifying fake news has become important. Researchers and technologists are investigating a range of strategies to tackle the issue of fake news spread, as there is a clear need for efficient tools and models to counter it (Vosoughi et al., 2018).

*Corresponding Author

Section B- ELECTRICAL/COMPUTER ENGINEERING & COMPUTING SCIENCES
Can be cited as:
Odunmuyiwa V.T. and Aderemi D.O (2024). Optimizing Fake News Detection in Resource-Constrained Devices Using Transformer Distillation. FUOYE Journal of Engineering and Technology (FUOYE**JET**), **9(4), 600-608. https://dx.doi.org/10.4314/fuoyejet.v9i4.6**

For the purpose of identifying and mitigating fake news, several types of deep-learning algorithms have been investigated and put into practice; however, they frequently encounter serious obstacles and constraints. Large-scale language models like BERT (Bidirectional Encoder Representations from Transformers), which are deep learning models employed for this purpose, are resource- and computationally-intensive. These models are not appropriate for deployment in situations with limited resources, such as mobile or edge devices, since they require a lot of processing power and memory. Furthermore, it can be unfeasible to train and fine-tune these big models in real-time or with low latency due to the vast datasets and lengthy training cycles involved. In addition, these models' enormous size presents deployment and scalability issues, particularly in situations where quick inference is essential.

Several transformer-based models have been proposed to address these limitations. DistillBERT, for example, is a smaller, faster, and lighter version of BERT that aims to retain much of BERT's performance while reducing its computational burden. However, despite its improvements, DistillBERT still takes significant amount of computational resources and may not be optimal for environments with stringent resource constraints. TinyBERT, on the other hand, is specifically designed to create lightweight versions of BERT by using techniques like knowledge distillation.

This study compares BERT, TinyBERT, and DistillBERT so as to examine how well Transformer-based models work in creating small, reliable models for fake news detection. The goal of this research is to develop tools that work well with widely used social media and information-sharing mobile apps like Facebook, WhatsApp, and Twitter. Users can gain real-time information verification and increase the trustworthiness of shared content by directly integrating these detection capabilities into these platforms. With its use of advanced AI technologies to maintain the integrity and reliability of information, this research represents a major breakthrough in the fight against disinformation in the digital age.

2 Literature Review

2.1 Approaches for Fake News Detection

Fake news travels quickly and extensively via social media. According to studies, polarising content, emotional appeal, novelty, and other characteristics cause false information to travel more quickly and reach a larger audience than correct information (Vosoughi et al., 2018). A broad variety of content can be classified as fake news, including edited photos and videos as well as made-up news articles. According to Vosoughi et al. (2018), fake news has far-reaching effects that include social unrest, political instability, and economic harm. Tandoc et al. (2017) claim that in recent years, there has been a strong correlation between false news and significant events including elections, public health emergencies, and societal challenges. False information is becoming a major social concern due to how simple it is to create and spread online.

Researchers from computer science, social science, and journalism are working together to address the problem of fake news identification. Numerous strategies have been investigated:

1. Rule-Based Approaches: These methods use pre-established rules to identify fake news based on features like sensationalism, attention-grabbing content, or well-known disinformation sources.
2. Machine Learning Models: A range of machine learning models have been implemented, frequently utilising task such as sentiment analysis, text analysis, and network analysis to detect trends linked to false information.
3. Deep Learning Models: According to Alotaibi & Alhammad (2022), deep learning models, especially those based on recurrent neural networks (RNNs) and convolutional neural networks (CNNs), have demonstrated potential in the detection of fake news.

2.1.1 Rule-Based Approach

These techniques mostly rely on established rules that can become inadequate or out-of-date as fake news strategies change (Broda and Strömbäck, 2024). Static rule sets are easily overcome by new types of misinformation, such as deepfakes or subtle manipulations, therefore these methods are not useful for dynamically identifying emerging trends in fake news. Because of this, it becomes more challenging to maintain and update rule-based systems, needing continuous human intervention to stay up to date with new disinformation strategies.

Furthermore, rule-based strategies frequently overlook the underlying context or purpose of the content in favour of concentrating on surface-level elements like sensationalism or clickbait. Due to the fact that fake news can often mimic authentic content in terms of style and presentation while spreading false or dangerous information, this constraint results in a limited scope for detection (Broda and Strömbäck 2024). Furthermore, more subtle kinds of deception, such as biassed reporting or selectively curated data, that do not show obvious red flags may go unnoticed by rule-based systems. As a result, relying too much on outward appearances may give rise to a false sense of security because sophisticated fake news operations can avoid detection by following preset guidelines without raising red flags based solely on content.

2.1.2 Machine Learning Approach

Shu et al. (2017) conducted a study titled "Fake News Detection on Social Media: A Data Mining Perspective" that focuses on using machine learning and data mining approaches to stop fake news from spreading on social media platforms. By extracting informative features from textual content shared on social media and using supervised learning algorithms like Support Vector Machines (SVM) and Random Forests for classification, the study seeks to address the problem of misinformation identification. The research aims to differentiate between authentic and fraudulent news stories by structural patterns in social media posts. This endeavour

http://.doi.org/10.46792/fuoyejet.v9i4.6 engineering.fuoye.edu.ng/journal

contributes to the advancement of automated technologies designed for the purpose of detecting fake news. Although the method used in this study is valuable for its creative application of machine learning and data mining to social media, there are a number of possible drawbacks and objections to take into account. First off, the calibre and representativeness of the training data have a major impact on how well supervised learning models work. It can be difficult to keep an updated and varied dataset given the dynamic nature of social media material and the changing strategies used by disinformation campaigns. Furthermore, focusing only on textual characteristics may ignore other crucial contextual cues that could improve the detection of false news, like user interaction trends, social network dynamics, and multimedia content. Although the method's attempt to automatically identify false information using easily accessible data is valuable, it might not be sufficient to handle the intricate and constantly changing nature of disinformation. Rule-based and feature-based methods may struggle to capture the nuanced linguistic and contextual cues indicative of fake news, leading to potential inaccuracies and limited generalizability.

The work of Amahan (2023) demonstrates significant strengths in leveraging traditional data mining techniques, such as meta-bagging algorithms, for detecting patterns in fake news datasets. One of its strengths that is worthy of attention is the simplicity and interpretability of the methods compared to more complex models like Transformers. By relying on linguistic attributes and statistical measures, it offers a clear and explainable framework for identifying misinformation, achieving an accuracy rate of 86%. This level of interpretability is often advantageous in applications requiring transparency, such as journalism or regulatory investigations.

However, compared to Transformer-based models (e.g., BERT), which excel in understanding context and semantics in natural language, the study's approach may be less robust for nuanced or sophisticated misinformation that requires deeper contextual understanding. Transformers leverage attention mechanisms to capture relationships between words and phrases over long sequences, outperforming traditional methods in tasks like fake news detection. The reliance on pre-determined linguistic attributes in data mining can lead to weaknesses in adaptability, as it may struggle to generalise across varied datasets or evolving fake news patterns.

2.1.3 Deep Learning Approach

Deep learning approaches have emerged as promising methodologies for tackling the pervasive issue of fake news detection. Leveraging the power of neural networks, these approaches delve deep into the textual content of news articles to uncover patterns and features indicative of misinformation. By utilizing architectures such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformer-based models like BERT, deep learning models can capture complex linguistic patterns and semantic relationships within textual data. Deep learning models can effectively identify subtle cues and characteristics associated with fake news through extensive training and fine-tuning processes. This provides an improved approach superior to traditional Machine Learning.

While non-transformer-based deep learning techniques, such as CNNs, have demonstrated potential in the identification of false news, they might be more constrained than transformer-based techniques, like BERT. Their reliance on fixed-size context windows, which could miss important semantic links and long-range dependencies found in complicated textual material, is one of their main limitations.

Because transformer-based models, such as BERT, can process and grasp text more contextually and semantically, they present a promising option for the detection of fake news. Transformers are particularly good at recognising long-distance dependencies and relationships between words and phrases, which is important for spotting minute manipulations and dishonest strategies used in fake news.

Zhang et al's (2020) research paper titled "BERT-Based Domain Adaptation Neural Network for Multi-Modal Fake News Detection" offers a useful investigation into the application of BERT for fake news identification tasks. Although the study shows that BERT can identify language clues and contextual details that are suggestive of fake news, its methodology has inherent drawbacks. BERT models need a lot of computing power and labelled training data, especially when they are directly refined on task-specific datasets. This dependence on massive processing capacity and annotated datasets could make it difficult to scale and make the data accessible, especially for researchers with limited access to large-scale computer infrastructure and tagged datasets or in locations with limited resources.

On the other hand, transformer models can use knowledge distillation techniques to provide a more scalable and resource-efficient method of detecting false news. Transferring knowledge from a large, pre-trained model such as BERT to a smaller, distilled model that minimises computational overhead while maintaining much of the original model's performance is known as knowledge distillation. It is possible for researchers to develop lightweight and effective models specifically designed for the identification of fake news by condensing the knowledge acquired by BERT into a smaller model architecture. These simplified models have reduced memory needs and quicker inference times, which makes them more suitable for use in real-world scenarios with constrained computational resources.

While many techniques, such as weight sharing, pruning, quantization, low-rank approximation, are available for compressing pre-trained language models, knowledge distillation has proven to be very effective. This study aims to build a compact false news detection model based

on TinyBERT by utilising knowledge distillation. To accurately evaluate the effectiveness of knowledge distillation in compressing models appropriate for the news domain, the final model will be tested against Base BERT and DistillBERT. By assessing these models' performance, the research hopes to clarify the potential of knowledge distillation as a crucial strategy for creating succinct, successful, and resource-constrained fake news detection systems.

2.2 TRANSFORMER

The transformer model uses self-attention techniques in contrast to RNN or CNN, which rely on sequential or local processing. This makes it possible for the model to evaluate the complete input sequence at once, improving the way it captures semantic connections and global dependencies.

2.2.1 TRANSFORMER LAYER

Transformer layers are a characteristic of most modern pre-trained language models, including BERT, DistillBERT, and RoBERTa. They use a self-attention mechanism to capture long-term dependencies between input tokens. The two main sub-layers of these Transformer layers are usually a fully linked feed-forward network (FFN) and multi-head attention (MHA).

Multi-Head Attention: The calculation of the attention function relies on three key components: queries, keys, and values, represented as matrices Q, K, and V, respectively. This attention function can be expressed as follows:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (1)$$

Here, d_k represents the dimension of the keys, serving as a scaling factor. The attention matrix is represented by the parametric expression in the softmax function and is derived by a dot-product operation from the compatibility of Q and K. A weighted sum of the values V, with the weights decided by running the softmax() operation over each column, yields the output of the function as a whole. According to Clark (2019), BERT's attention matrices are capable of capturing a substantial amount of linguistic knowledge, which makes them an essential component of the distillation technique that this study suggests.

Multi-head attention is defined by concatenating the attention heads from different representation subspaces as follows:

$$MHA(Q,K,V) = Concat(h_1,\dots,h_k)W, \quad (2)$$

In this context, k represents the number of attention heads and h_i refers to the i-th attention head, inputs originating from distinct representation subspaces are used to compute the Attention() function. The matrix W serves as a linear transformation within this process.

Positional Feed-Forward Network (FFN): The Transformer layer additionally comprises a fully linked feed-forward network (FFN), expressed by the following equation:

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2. \quad (3)$$

In this formulation, the FFN involves two linear transformations, denoted by $xW_1 + b_1$ and W_2, and a Rectified Linear Unit (ReLU) activation function. It contributes to the model's ability to capture complex patterns and relationships within sequences.

To capture the concept of FFN further:

1. The input x is transformed by the first linear layer with weights W_1 and bias b_1, then the ReLU activation function is applied to introduce non-linearity.
2. The output of the ReLU activation function undergoes another linear transformation with weights W_2 and bias b_2., yielding the final output of the FFN.

The ReLU activation function $max(0, x)$ ensures that only positive values are passed through, effectively introducing non-linearity to the network.

2.2.2 KNOWLEDGE DISTILLATION

Knowledge distillation (KD) tries to transfer information from a large teacher network, denoted as T, to a smaller student network, represented as S. The objective of KD is for the student network to adopt the behaviours demonstrated by the teacher network during training. Let f^T and f^S denote the behaviour functions of the teacher and student networks, respectively. These behaviour functions, which are the output of any network layer, are in charge of converting network inputs into instructive representations. Several elements can act as behaviour functions in the context of Transformer distillation, including the Feed-Forward Network (FFN) layer, the Multi-Head Attention (MHA) layer, and intermediate representations like the attention matrix A. Formally, KD can be conceptualized as minimizing the following objective function:

$$L_{KD} = \sum_{x\epsilon\chi} \quad L(f^S(x), f^T(x)), \quad (4)$$

L_{KD} denotes the Knowledge Distillation Loss, expressed as $L_{KD} = L(f^S(x), f^T(x))$, where L(·) represents a loss function that quantifies the disparity between the outputs of the student and teacher networks for a given text input x, which belongs to the training dataset χ.

2.3 RELATED WORKS

Existing works in fake news detection have seen a proliferation of approaches, ranging from traditional rule-based methods to sophisticated deep learning techniques. The research by Shu, Wang & Liu (2017) explores rule-based methods for detecting fake news on social media platforms. While the study provides valuable insights, its reliance on manual rule creation limits its scalability and effectiveness in handling diverse types of fake news. Authors like Ruchansky et al. (2017) explored the application of deep learning methods for the detection of false news, such as long short-term memory (LSTM) networks and recurrent neural networks (RNNs).

http://.doi.org/10.46792/fuoyejet.v9i4.6 engineering.fuoye.edu.ng/journal

Although deep learning models show great promise, their practical utility in real-world contexts may be limited by issues such as overfitting and lack of interpretability.

An increasing volume of recent research has concentrated on transformer-based models, like BERT, for the purpose of detecting fake news. These models take advantage of knowledge distillation techniques to compress big pre-trained models into smaller, more efficient forms, and they also harness the power of transformer architectures to capture contextual information and complex language patterns within text. For instance, a distilled transformer model for fake news detection is proposed in the work of Alghamdi et al. (2020), which achieves competitive performance while drastically lowering the amount of memory and processing resources needed. The method, which is well-suited for deployment in resource-constrained contexts, strikes a balance between accuracy and efficiency by condensing knowledge from large-scale pre-trained models like BERT into compact variations like TinyBERT.

Additionally, Liu et al.'s (2019) investigates how well distilled transformer models work to identify false information in a variety of languages and domains. The work highlights the resilience and generalizability of distilled transformer models by showing them to regularly outperform conventional techniques and non-transformer-based deep learning approaches. Distilled transformer models, by reducing complexity, make it easier to apply interpretability techniques which can enhance explainability and help users identify potential biases or mistakes in the model's decision-making process. In particular, in vital applications like news verification and media literacy instruction, this transparency is essential for fostering confidence and trust in fake news detection systems.

Guo et al (2024) used multiscale transformer to detect fake news in mixed languages while Praseed et al (2023) explored transformer ensembles for fake news detection in low-resource languages (Odumuyiwa et.al. 2024). Bahmanyar & Miyaneh (2024) used a hybrid approach to detect fake news through the application of a transformer model serving as the backbone to generate three components: sentiment analysis, hate speech detection, and topic modeling. The outputs of these components are combined through layer concatenation, which is subsequently passed into the final classifier for fake news prediction.

For this paper, our interest is to experiment with different BERT models for fake news detection emphasising the increasing importance of transformer-based models, especially distilled transformer models like TinyBERT, as practical and efficient means of battling false information in resource-constrained devices.

3 METHODOLOGY

3.1 WORKFLOW

This section outlines the step-by-step process followed while building the proposed model from TinyBert. The workflow is divided into data preparation and model training. Data preparation encompasses dataset selection, cleaning and preparation and feature engineering. These processes are further illustrated in Figure 1.

Data preparation processes pass prepared data to model training processes. The pre-processed data are first tokenized using the BERT tokenizer with padding and truncation to maintain uniform sequence length.

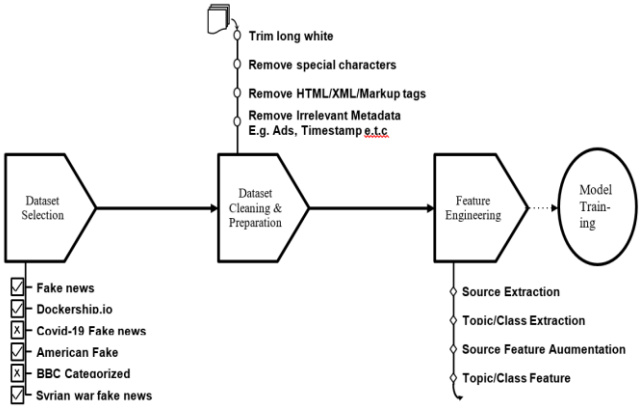

Fig. 1: Dataset preparation

The model training processes comprise model building, model evaluation and metric scoring. This setup was deliberately designed to facilitate the abstraction of both the dataset and the model. This means that we used the same code base to quickly test multiple datasets and models without needing to restructure the code, ensuring that the quality of the results remain unaffected. Figure 2 illustrates the interaction of the processes involved in training the model.

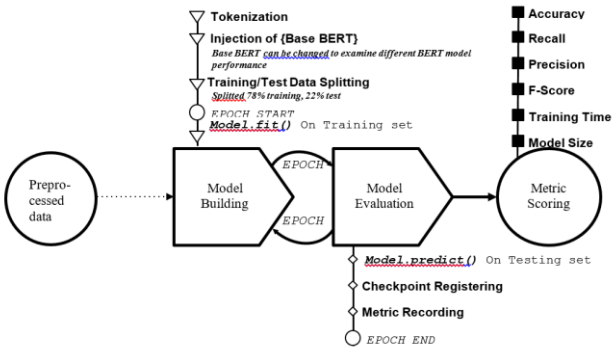

Fig. 2: Model training

3.2 DATASET

Four distinct datasets were carefully selected to facilitate comprehensive analysis and evaluation of fake news detection models. The datasets chosen were the Fake News Network dataset, Dockership.io dataset, American Fake News dataset, and Syrian War Fake News dataset. All dataset were sourced from Kaggle dataset repositories. These datasets were selected based on their diversity in terms of sources, topics, and linguistic characteristics, aiming to provide a well-rounded representation of fake news instances across various contexts.

http://.doi.org/10.46792/fuoyejet.v9i4.6 engineering.fuoye.edu.ng/journal

The Fake News Network dataset is made up of 23,197 news stories that have been classified as fake or true. These articles are sourced from a number of websites that are well-known for spreading false information. The 44,058 textual records in the Dockership.io dataset was taken from online forums and social media sites, with a particular emphasis on fake news on current societal topics. The Syrian War Fake News dataset focuses on misinformation relating to the conflict in Syria and includes 805 news stories tagged as such, whereas the American Fake News dataset comprises 166,355 news articles primarily targeting an American audience. Typically, every dataset comprises roughly 50% genuine news and 50% fraudulent news.

Extensive preprocessing procedures were used to guarantee data quality and consistency across all datasets prior to model training and evaluation. Among the preprocessing tasks were text cleaning, metadata removal, and class distribution balance to reduce bias. Furthermore, steps were taken to address duplicate or missing entries, standardise text formats, and eliminate extraneous or noisy content that can negatively impact the performance of the model. The carefully selected datasets permit a thorough review of the effectiveness of the model in various settings, consequently advancing the field of study on countering false information and advancing information integrity on online platforms.

3.3 Architecture of The Proposed Model

The selection of the pre-trained base model forms the basis for our fake news detection. With the help of large text corpora, BERT, a transformer-based model, can provide bidirectional contextual embeddings that accurately represent complex linguistic subtleties. DistillBERT is a streamlined variant of BERT that enhances compression without compromising the model's original functionality. MobileBERT emphasises speed and efficiency over accuracy when optimising BERT for deployment on mobile and edge devices. By using knowledge distillation techniques to condense the information of a larger teacher model—like BERT—into a smaller student model, TinyBERT expands on this idea and creates a more portable yet potent false news detection system. This work leverages the TinyBERT Architecture of Teacher-Student Distillation Process proposed by Jiao et. al. (2020).

To illustrate knowledge distillation applied by TinyBERT in the context of Transformer-layer distillation, where the student model consists of M Transformer layers and the teacher model comprises N Transformer layers, the process begins by selecting M out of N layers from the teacher model. Subsequently, a mapping function $n = g(m)$ is established to define the correspondence between the indices of the student and teacher layers. This function ensures that the m-th layer of the student model learns information from the $g(m)$-th layer of the teacher model. To clarify, we designate 0 as the index of the embedding layer and $M+1$ as the index of the prediction layer. Accordingly, the mappings for these layers are defined as $0 = g(0)$ and $N+1 = g(M+1)$, respectively. The influence of various mapping functions on model performance is rigorously examined in the subsequent experimental analysis. Formally, the student model acquires knowledge from the teacher by minimizing the following objective function.

$$\mathcal{L}_{model} = \sum_{x \epsilon X} \quad \sum_{m=0}^{M+1} \quad \lambda_m \mathcal{L}_{layer}(f_m^S(x), f_{g(m)}^T(x))$$

Where $\mathcal{L}_{layer}$ refers to the loss function of a given model layer (e.g., Transformer layer or embedding layer), $f_m\ (x)$ denotes the behaviour function induced from the m-th layers and λ_m is the hyper-parameter that represents the importance of the m-th layer's distillation.

3.4 Experimental Design

The experimentation process encompasses data partitioning, model training using the Hugging Face TensorFlow-based library, and the selection and evaluation of training metrics. To facilitate robust evaluation, we partition the dataset into training and testing subsets, allocating 78% of the data for training purposes and reserving the remaining 22% for model evaluation.

For model development and training, we leverage the Hugging Face TensorFlow-based library, which offers a comprehensive suite of tools and utilities for building and fine-tuning transformer-based models. Our experiments involve the use of pre-trained transformer models, including BERT, DistillBERT, MobileBERT, and TinyBERT, as the base models for fake news detection. To adjust the parameters of these models to the particulars of the task, they are refined on the training subset of the dataset.

Several training metrics are calculated during the model-training process to track the models' convergence and performance. These metrics consist of F1-score, recall, accuracy, and precision, among others. Each training metric is computed numerous times to ensure robustness and reliability. The impacts of random initialization and training variability are mitigated by taking the average value.

4 Result and Discussion

This section presents the results obtained from the four BERT models that were trained on four different datasets. Additionally discussed are evaluation metrics and their implications.

4.1 Evaluation Metric

4.1.1 Accuracy

Accuracy measures the overall correctness of the model predictions and is a key performance indicator for models. A high accuracy score shows that the model can accurately categorise news stories as real or fake. But accuracy by itself might not give the whole story, particularly if the datasets are unbalanced, meaning that

one class predominates. However, it provides a helpful starting point for assessing the effectiveness of the model.

4.1.2 F1-Score

The F1-score is a harmonic mean of precision and recall, providing a balanced measure of a model's performance, particularly in scenarios with imbalanced datasets. A model's capacity to produce high recall (minimising false negatives) and high precision (minimising false positives) is indicated by a high F1-score. As a result, it provides an extensive evaluation of the predictive power of the model and is particularly pertinent in applications where false positives and false negatives have serious ramifications.

4.1.3 Precision and Recall

Recall represents the percentage of correctly predicted positive instances among all actual positive instances in the dataset, whereas precision measures the percentage of correctly predicted positive instances (true positives) among all instances projected as positive. When minimising false positives is critical, precision is important; when minimising false negatives is critical, recall is important. Depending on the particular requirements of the application, recall and precision must be balanced.

4.1.4 Training Time and Model Size

Factors such as training time and model size affect the scalability and viability of using false news detection models in practical applications. Longer training durations could impede the speed at which models are iterated and implemented, particularly in settings with limited resources. Similar to this, higher model sizes could make deployment more difficult, especially for apps with little memory or storage. To guarantee scalability and practical application, model performance must be balanced with training time and model size.

4.2 Results

Followings are the results for each dataset trained on each BERT variant as a base model:

Table 1: Fake News Network Dataset

Model	Acc	Prec	Rec	F1	Model Size (MB)	Training Duration (seconds)
BERT	0.96	0.99	0.91	0.95	320.43	8989.30
DistillBERT	0.94	0.98	0.90	0.94	267.30	4650.00
MobileBERT	0.54	0.61	0.64	0.62	292.00	6802.54
TinyBERT	0.94	0.94	0.95	0.95	22.40	286.80

Table 2: Dockership.io Dataset

Model	Acc	Prec	Rec	F1	Model Size (MB)	Training Duration (seconds)
BERT	0.99	0.99	0.91	0.96	346.13	8690.45
DistillBERT	0.96	0.97	0.95	0.96	292.12	5033.00
MobileBERT	0.58	0.59	0.51	0.55	266.40	4940.70
TinyBERT	0.89	0.90	0.88	0.89	26.30	246.60

Table 3: America Fake News Dataset

Model	Acc	Prec	Rec	F1	Model Size (MB)	Training Duration (seconds)
BERT	0.96	0.99	0.91	0.95	335.09	7773.33
DistillBERT	0.97	0.98	0.96	0.97	267.14	5400.20
MobileBERT	0.58	0.58	0.53	0.56	247.65	5054.60
TinyBERT	0.90	0.91	0.89	0.90	24.1	246.60

Table 4: Syria War Fake News Dataset

Model	Acc	Prec	Rec	F1	Model Size (MB)	Training Duration (seconds)
BERT	0.53	0.54	0.61	0.57	335.09	793.33
DistillBERT	0.53	0.48	0.48	0.48	267.14	592.35
MobileBERT	0.49	0.47	0.70	0.57	233.92	554.60
TinyBERT	0.56	0.51	0.30	0.38	24.1	28.60

Table 5: Two-Ways Anova Analysis Result

Source of Variation	Sum of Squares	df	Mean Square	F	P-value
Between Datasets	10919694.99	3	3639898.329	2.194628991	0.0990028
Between Metrics	144340901.9	4	36085225.47	21.75711375	0
Between Models	24807709.62	12	2067309.135	1.246456948	0.2767566
Within	91220160.17	55	1658548.367		
Total	**271288466.7**	**74**			

4.2.1 Two-Way Anova with Replication for Comparative Analysis

This section delves deeply into the comparison measures that are used to evaluate the performance of various models in the context of detecting false news. Our goal is to obtain a thorough grasp of the interactions and combined effects of important independent factors, such as accuracy, model size, and training time, on model performance. A two-way ANOVA with a replication design that takes into consideration probable sources of variation and sheds light on how these factors interact is used to guarantee the validity of our results. Finding statistically significant variations in the models' or datasets' performance is the main objective, especially with regard to the identification of fake news and its usefulness in resource-constrained contexts.

4.2.2 Two-Way Anova with Replication

To examine how the three independent variables—accuracy, model size, and training time—and the model selection (categorical variable) affected the dependent variable of model performance, a two-way ANOVA with replication was used. Replication is used in this research to improve within-group variance estimation, which increases the sensitivity and precision of our findings.

http://.doi.org/10.46792/fuoyejet.v9i4.6 engineering.fuoye.edu.ng/journal

4.2.3 Analysis and Interpretation

1. Accuracy vs. Model Size: We looked at how the interaction between model selection and model size affect accuracy. This study sheds light on whether larger or smaller models often have higher accuracy rates.
2. Accuracy vs. Training Time: It was evaluated how training time and model selection work together to affect accuracy. It is necessary to understand this relationship in order to choose models wisely.
3. Model Size vs. Training Time: The trade-offs between model size and training time were explored. This analysis reveals the practical implications of choosing models of different sizes.

4.3 Discussion

The last section contains the main substance for discussion, there are four sources of variations in the report:

Between dataset: Variation between the levels of dataset factor.

Between metric: Variation between the levels of accuracy, training period and model size.

Between models: Interaction effect between the models.

Within: Variation within the combinations of the dataset and metric can be used as a baseline for other summations.

After conducting individual two-way ANOVA analyses for each pair of variables, a comprehensive comparison of results was undertaken to discern commonalities and disparities, maintaining a 5% significance level. By synthesizing these findings, a holistic perspective on the collective impact of accuracy, model size, and training time on overall model performance was achieved.

The obtained between dataset P-value is 9.9%, surpassing our 5% benchmark, indicating a lack of statistical significance difference in performance of the models between datasets. Additionally, the between models P-value stands at 27.7%, asserting that there is no significant difference in the interaction between the considered models. Consequently, it can be inferred that TinyBERT exhibits efficiency comparable to other BERT models, with the added advantage of resource efficiency. This positions TinyBERT favourably, especially in resource-constrained environments such as mobile devices. TinyBERT's efficacy and efficiency make it possible to implement it on a variety of devices, guaranteeing accessibility and dependability even in situations with limited resources where conventional deep-learning methods would not be as useful.

5 Conclusion

In conclusion, our research project has demonstrated the potential of BERT-based models in the detection of fake news. The effectiveness of these models, combined with insights into dataset characteristics and practical considerations, contributes to the ongoing efforts to combat misinformation in the digital age. By leveraging the strengths of BERT-based models and considering the nuances of fake news, we move closer to a more informed and resilient society capable of discerning truth from deception.

As the landscape of fake news continues to evolve, our research offers valuable insights and recommendations for researchers, practitioners, and stakeholders dedicated to the pursuit of information integrity and the preservation of the public's trust in reliable news sources.

6 References

Alghamdi, J., Lin, Y., & Luo, S. (2023). Towards COVID-19 fake news detection using transformer-based models. *Knowledge-Based Systems*, 274, 110642.

Alotaibi, F. L., & Alhammad, M. M. (2022). Using a rule-based model to detect arabic fake news propagation during covid-19. *International Journal of Advanced Computer Science and Applications*, 13(1).

Amahan, P. A. (2023). The perspective of data mining: the study of fake news on social media. *Dynamic Journal of Pure and Applied Sciences*, Ozamiz City, Philippines.

Bahmanyar, R., & Miyaneh, E. K. (2024). A Novel Content-based Approach for Fake News Detection using Transformer Model: A Case Study of Covid-19 Dataset. *In 2024 10th International Conference on Web Research (ICWR)* (pp. 364-369). IEEE.

Broda, E., & Strömbäck, J. (2024). Misinformation, disinformation, and fake news: lessons from an interdisciplinary, systematic literature review. *Annals of the International Communication Association*, 48(2), 139-166.

Clark, K. (2019). What Does Bert Look At? An Analysis of Bert's Attention. In *Proceedings of the 2019 ACL Workshop. BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*.

Guo, Z., Zhang, Q., Ding, F., Zhu, X., & Yu, K. (2023). A novel fake news detection model for context of mixed languages through multiscale transformer. *IEEE Transactions on Computational Social Systems*, 11(4), 5079-5089.

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., ... & Liu, Q. (2019). Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*.

Liu, C., Wu, X., Yu, M., Li, G., Jiang, J., Huang, W., & Lu, X. (2019). A two-stage model based on BERT for short fake news detection. In Knowledge Science, Engineering and Management: 12th International Conference, KSEM 2019, Athens, Greece, August 28–30, 2019, Proceedings, Part II 12 (pp. 172-183). Springer International Publishing.

Liu, Chao & Wu, Xinghua & Yu, Min & li, Gang & Jiang, Jianguo & Huang, Weiqing & Lu, Xiang. (2019). A Two-Stage Model Based on BERT for Short Fake News Detection. 10.1007/978-3-030-29563-9_17.

Odumuyiwa, V. T., Shoyemi, O.O., & Fagoroye, A. E. (2024). Sentiment Analysis of Low-Resource Yorùbá Tweets Using Fine-Tuned Bert Models. *University of Ibadan Journal of Science and Logics*

in ICT Research, 12(1), 110-122

Praseed, A., Rodrigues, J., & Thilagam, P. S. (2023). Hindi fake news detection using transformer ensembles. *Engineering Applications of Artificial Intelligence*, 119, 105731.

Ruchansky, N., Seo, S., & Liu, Y. (2017, November). Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 797-806)

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1), 22-36.

Shu, K., Wang, S., & Liu, H. (2017). Exploiting tri-relationship for fake news detection. *arXiv preprint arXiv:1712.07709*, 8.

Tandoc Jr, E. C., Lim, Z. W., & Ling, R. (2018). Defining "fake news" A typology of scholarly definitions. *Digital journalism*, 6(2), 137-153.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *science*, 359(6380), 1146-1151.

Zhang, T., Wang, D., Chen, H., Zeng, Z., Guo, W., Miao, C., & Cui, L. (2020). BDANN: BERT-based domain adaptation neural network for multi-modal fake news detection. *In 2020 international joint conference on neural networks (IJCNN)* (pp. 1-8). IEEE.