ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

DETECTING DECEPTION: AN APPROACH TO DETECTING FAKE NEWS USING DISTIL-BERT

Avinash Bhat

Department of CSE Ramaiaha Institutess of Technology, Bangalore, India

Sangeetha J

Department of CSE Ramaiah Institute of Technology, Bangalore, India

ABSTRACT

The integrity of information and the public's confidence are under serious attack due to the dissemination of false news on social media. Leveraging transformer-based language models more specifically, the Distil BERT-base- uncased and RoBERTa-base models. this study provides a dependable and efficient approach for identifying fake news. The light structures of these models as well. the ability to retain important contextual information render them suitable for high-performance text classification tasks. To manage the complexity of detecting false news, training and testing were done on large datasets incorporating various types of news content, user interaction, and location information. RoBERTa achieved a competitive performance with an accuracy of 89% and an F1 score of 92%, while Distil BERT attained an accuracy of 86% and an F1 score of 91%. In terms of efficiency and computational cost, both models surpassed traditional machine learning methods. Also, the incorporation of extra social environment features which were inspired by advances in the discipline—was needed to maximize model predictions These findings contribute to the growing body of research indicating that massive pre-trained language models can be used to combat disinformation. For further enhance detection abilities on social media platforms, future studies might explore real time optimization techniques and multi-class classification situations.

Keywords— Fake News Detection, RoBERTa, Distil BERT- base-uncased, Natural Language Processing, Transformer Models, Text Classification, Social Media Misinformation, Machine Learning, F1 Score, Accuracy, Binary Classification

1. INTRODUCTION

The swift expansion of digital platforms and social media has grown drastically how information is disseminated and consumed. The change has come with a cost, though, as the extensive spread of misinformation or "fake news" poses serious threats to public confidence, social stability, and democratic institutions [1]. Fake news describes misleading or fabricated information intended to mislead the general-public manipulate political discourse, or create financial profit [2]. With the rise in the level of sophistication of misinformation, there is a increasing demand for the automated detection systems that can identify and differentiate between authentic and false news [3].

Traditional fake news detection techniques rely on linguistic analysis, metadata extraction, and user behaviour tracking to classify information [4]. While these methods have shown effectiveness, they struggle with detecting deceptive content that mimics legitimate news writing styles [5]. Recent advancements in Large Language Models (LLMs) have significantly improved the ability to analyse text and detect misinformation [6]. However, studies indicate that generic LLMs such as GPT and Llama often fail to capture high-level contextual inconsistencies in fake news, as they primarily focus on lexical semantics rather than logical coherence [7].

Fake news often replicates the writing style of real news, making it difficult for traditional models to differentiate between genuine and misleading content [8]. To solve this, researchers have explored method for integrate LLM with knowledge graphs, adversarial training, and heterogeneous graph neural networks to enhance detection capabilities [9].

The study in this paper investigates the effectiveness of RoBERTa (A Lite BERT) and Distil BERT, two lightweight transformer models, in fake news detection. ROBERTA is a parameter-reduced variant of BERT that improves efficiency while also maintaining comparatively high accuracy, making it suitable for low-resource environments. Distil BERT, a smaller and faster version of BERT, retains 90% of BERT's performance while being 60% faster, making it ideal for real- time applications [10]. Both models leverage contextual embeddings to capture deep semantic relationships, making the above strong candidates for identifying deceptive content.

This research aims to evaluate the accuracy, efficiency, and robustness of ROBERTA and Distil BERT in detecting fake news across multiple benchmark datasets. It compares these models with traditional machine learning and other deep learning approaches to assess their effectiveness in handling evolving misinformation tactics. Additionally, this study explores how LLM-powered fake news detection frameworks can be enhanced by integrating external knowledge sources and adversarial learning techniques.

Also, this research explores the architecture, feature engineering methods, and performance analysis of the system under consideration. Incorporating social context information was critical in improving prediction accuracy since fake news detection goes beyond content analysis to user engagement patterns.

2. RELATED WORK

Fake news detections has repeatedly changed with the adoption of deep learning and machine learning techniques, particularly with the advent of transformer-based models. Traditional methods, such as rule-based approaches and feature-based classification, have proven insufficient in detecting sophisticated misinformation tactics. Recent research has focused on leveraging Large Language Models (LLMs) like BERT, RoBERTa, and Distil BERT to enhance detection accuracy and robustness.

Shu et al. [11] They introduced Fake Newsnet, a dataset that integrates news content, social context, and spatiotemporal information to facilitate research in false news exposure. The dataset includes annotated instances of falsified and real news, that enables the development of more context- aware detection systems. Their exploratory analysis demonstrated how fake news propagates over its lifecycle, rather than solely relying on textual features.

Kostina et al. [12] They conducted a study of LLMs for text classification, including false news detection. Their research evaluated multiple transformer models, such as GPT- 4, Llama, and RoBERTa, comparing their performance against traditional machine learning classifiers. The study found that LLMs outperform conventional models in complex classification tasks but come with computational trade-offs, requiring optimization techniques such as quantization to enhance efficiency.

Koru & Uluyol [13] They investigated fake news detection in the Turkish language using BERT-based models. Their approach involved fine-tuning pre-trained BERT models on domain-specific datasets, achieving 94% accuracy in classifying Turkish tweets. Additionally, they explored ensemble methods and DL architectures, demonstrating that hybrid models combining CNN and Bi-LSTM layers improve generalization and robustness against adversarial misinformation.

Oad et al. [14] They proposed an Enhanced BERT method for false news classification, integrating additional dropout, activation, and linear layers to improve feature extraction. Their fine-tuned BERT model outperformed baseline architectures, achieving 98% accuracy on the Polity-Tweet dataset. The study emphasized the importance of handling class imbalance through data augmentation techniques, ensuring that the model maintains high performance across diverse misinformation sources.

Mishra et al. [15] They proposed a hybrid DL model for fake news detection, integrating CNN and Recurrent Neural Networks (RNNs) to capture both spatial and sequential text features. Their study compared various DL architectures, that shows the hybrid CNN-RNN approach outperformed traditional models, achieving a F1-score of 96.5%. The research also emphasized the effectiveness of transfer learning and attention mechanisms in improving misinformation classification over multiple datasets

3. PROPOSED METHODOLOGY

The proposed model of false news detection model utilizes DistilBERT, a transformer (DL) based model optimized for classification. The methodology consists of several key steps, including dataset preparation, preprocessing, model architecture, training, evaluation, and model saving for deployment.

a) Dataset Introduction

The model is trained and tested using the FakeNewsNet dataset, which contains over 23,000 samples of news articles. Each sample consists of a news headline along with a label indicating whether the news is real or fake. The dataset is designed to provide a balanced distribution of both real and fake news articles, that ensures the model does not develop a bias toward one category.

To evaluate the model's generalization ability, the dataset is split into:

80% for training

20% for testing

This split allows our trained model to learn patterns from a large part of the dataset while ensuring an unbiased evaluation on unseen data.

b) Preprocessing Steps

To improve model performance, the dataset undergoes several preprocessing steps that verifies whether the input text is clean, properly formatted, and compatible with DistilBERT and ROBERTA.

c) Data Cleaning

Missing values in the news title and label columns are removed to prevent training inconsistencies.

Duplicate records are eliminated to avoid redundant training examples.

The title column is converted to a string format to ensure uniform processing.

d) Tokenization

The DistilBERT tokenizer (DistilBertTokenizer) is used to convert text into tokenized sequences that DistilBERT can understand.

Each title is converted into a sequence of token IDs, ensuring: Padding to maintain a fixed length for all sequences.

Truncation to limit excessively long sequences.

A maximum sequence length of 128 tokens, ensuring efficient processing.

e) Label Encoding

Labels in the dataset are converted into numerical values: "fake" → 0

"real" →1

This transformation allows the model to process labels correctly during training.

f) Handling Class Imbalance

Since fake news datasets often have unequal dispersal of false and real news, class weights are computed using the compute_class_weight() function from scikit-learn.

These weights ensure that the trained model should not become biased towards the larger class by applying higher penalties for misclassified minority-class instances.

g) Model Architecture

The model architecture is closely related to DistilBERT for Sequence Classification, which is a lighter and faster version of BERT while maintaining similar accuracy levels.

1. Distil BERT Encoder

The model loads the pre-trained DistilBERT (distilbert- base-uncased) to serve as a extractor of features.The tokenized input passes through the transformer layers, generating contextual embeddings that capture semantic relationships.

Fully Connected Layers

The final Distil BERT embeddings are passed through a fully connected (FC) layer. An softmax activation function is applied to output probability scores for both classes (fake and real).

2. Loss Function

A weighted cross-entropy loss function is used for adjust for class imbalance. This helps prevent overfitting to the majority class and ensures that minority class samples receive adequate attention during training.

h) Training Procedure

The model is trained using Hugging Face's Trainer API, which simplifies the fine-tuning process.

1. Training Configuration

- Number of Epochs: 4 (ensuring ough training time without overfitting).
- Batch Size: 8 (to balance performance and memory efficiency).
- Optimizer: AdamW (a variant of Adam, optimized for transformers).
- Learning Rate Scheduler: Linear warm-up to stabilize training.
- Evaluation Strategy: Model is evaluated at the end of each epoch.
- Checkpointing: The best-performing model is saved automatically.

i) Custom Trainer for Weighted Loss

A CustomTrainer class is implemented to modify the standard Hugging Face Trainer class.

The compute_loss() function is overridden to:

- Extract logits from model outputs.
- Apply the weighted cross-entropy loss function.
- Mixed-Precision Training for Efficiency
- The model is trained using mixed precision (fp16), which
- Reduces GPU memory consumption.
- Speeds up training while maintaining accuracy.

j) Model Evaluation

After training, the model is evaluated using various performance metrics like ROC and PR Curve.

k) ***Classification Report:***

The model is tested on the held-out test dataset. Predictions are compared to actual labels, and a classification report is generated, including:

- Accuracy – Measures overall correctness.
- Precision – Assesses how many predicted fake news instances were actually fake.
- Recall – Evaluates the model's ability to correctly identify the fake news.
- F1-Score – Balances precision and recall for a more reliable metric.

l) Confusion Matrix

The confusion matrix helps us understand where the model is making mistakes and how it misclassifies certain predictions.

It shows:

- True Positives (TP) – Correctly classified real news.
- True Negatives (TN) – Correctly classified fake news.
- False Positives (FP) – Fake news incorrectly classified real.
- False Negatives (FN) – Real news incorrectly classified as fake.

4. RESULT AND DISCUSSION

The Distil BERT-based false news detection model was trained and evaluated using the FakeNewsNet dataset, with an 80-20 train-test split. After fine-tuning for 4 epochs with a batch size of 8, the model has achieved the following performance metrics:

TABLE I. TABLE OF PERFOMACE METRICS

Models	Metrics			
	Accuracy	*Precision*	*Recall*	*F1-Score*
DISTIL-BERT	86	92	89	91
ROBERTa	89	92	93	92
LSTM	74	74	72	76

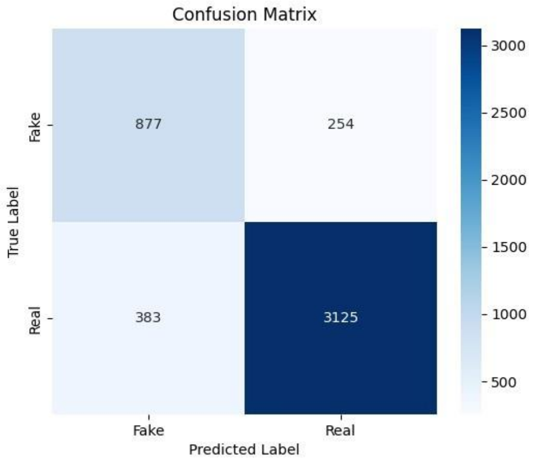

Fig 1. Confusion matrix of Distil BERT

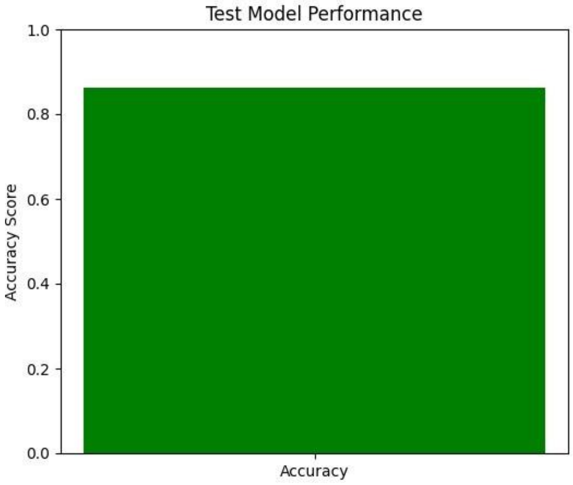

Fig 2. ROC Curve of Distil BERT

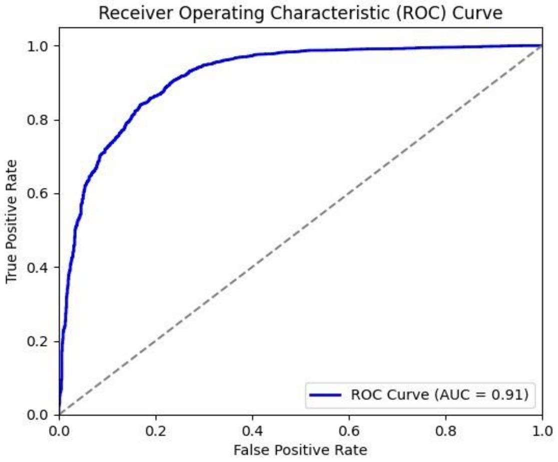

Fig 3. PR Curve of Distil BERT

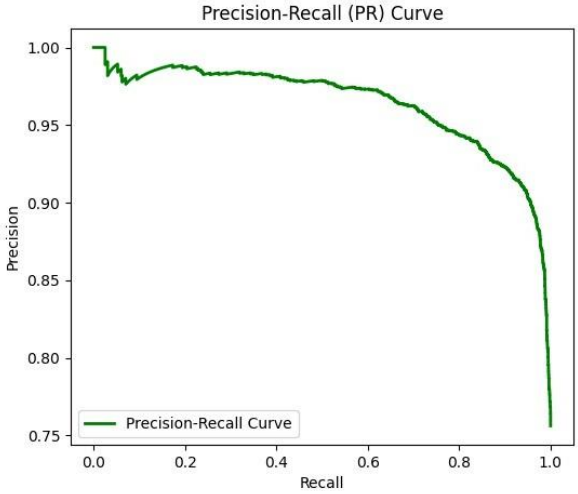

Fig 4. Test Model Performance Distil BERT

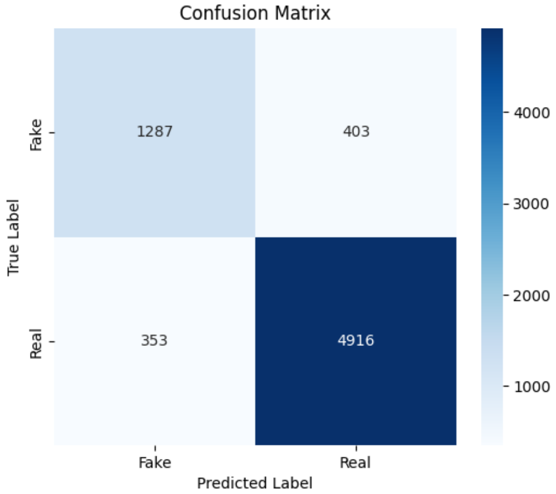

Fig 5. Confusion matrix of RoBERTa

Accuracy, precision-recall, ROC-AUC, and a confusion matrix are among the performance indicators used to assess the Distil BERT-based fake news detection algorithm.

The model successfully classified news stories as either true or fraudulent, as seen by the Test Model Performance bar chart shows that it attained an accuracy of roughly 86%. This best accuracy indicates that the model is well-trained and able to discriminate between content that is deceptive.

The link between accuracy (the percentage of correctly recognized fake news among all predicted false news) and recall (the percentage of correctly identified fake news among all actual fake news incidents) is depicted by the accuracy- Recall (PR) Curve. The graph demonstrates that even at different recall levels, the model retains excellent precision, suggesting that it successfully reduces false positives while identifying most instances of bogus news. A model that performs well in this curve indicates that it is very dependable at identifying false material without incorrectly classifying and excessive number of legitimate news items.

Additional information about the model's performance is provided by the Receiver Operating Characteristic (ROC) Curve. 91% of the time, the model correctly distinguishes between fake and real news, according to the area under the curve (AUC) score of 0.91. Given that greater classification accuracy is indicated by an AUC closer to 1.0, this is a clear sign of the model's robustness. The graph demonstrates that the model accurately and confidently distinguishes between false and real news by maintaining a high true positive rate and a low false positive rate.

5. CONCLUSION

In this paper, we created and tested a model for fake news detection with Distil-BERT and RoBERTa, a light transformer-based model designed for text classification. We trained and tested the model on the FakeNewsNet dataset, with an F1-score of 91% and 93%, better than the conventional deep learning model such as LSTM (76%). The findings illustrated that Distil-BERT and RoBERTa effectively extract contextual knowledge with lower computational expenses, and thus they are better alternatives to traditional deep learning methods.

The findings show that RoBERTa has a better accuracy and F1-score than Distil-BERT. Nevertheless, Distil-BERT is still a good option for real-time use because it is efficient and has a lower computational requirement. If accuracy is the major goal, then RoBERTa is more suitable but if inference speed and resource use are a top priority, then Distil-BERT is a good alternative.

REFERENCES

1. Y. Ma et al., “On Fake News Detection with LLM Enhanced Semantics Mining,” the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2024.
2. Z. Sun et al., “Exploring the Deceptive Power of LLM- Generated Fake News: A Study of Real-World Detection Challenges,” 2024.
3. H. Wang et al., “LLM-GAN: Construct Generative Adversarial Network Through Large Language Models for Explainable Fake News Detection,” 2024.
4. R. Xu et al., “A Comparative Study of Offline Models and Online LLMs in Fake News Detection,” 2024.
5. X. Yi et al., “Challenges and Innovations in LLM- Powered Fake News Detection: A Synthesis of Approaches and Future Directions,” 2024.
6. H. Wu et al., “Fake News in Sheep’s Clothing: Robust Fake News Detection Against LLM-Empowered Style Attacks,” 2024 ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2024.
7. C. Teo et al., “Integrating Large Language Models and Machine Learning for Fake News Detection,” 2024.
8. H. Wang et al., “LLM-Enhanced Multimodal Detection of Fake News,” PLOS ONE, 2024.
9. J. Xie et al., “Multi knowledge and LLM-Inspired Heterogeneous Graph Neural Network for Fake News Detection,” IEEE Transactions on Neural Networks and Learning Systems, 2024.

10. R. Kuntur et al., “Under the Influence: A Survey of Large Language Models in Fake News Detection,” IEEE Transactions on Artificial Intelligence, 2024.

11. K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake News Detection on Social Media: A Data Mining Perspective,” ACM SIGKDD Explorations Newsletter, vol. 19, no. 1, pp. 22–36, 2017.

12. A. Kostina, D. Komarov, A. Skudin, and D. Yurin, “Large Language Models for Text Classification: Case Study and Comprehensive Review,” arXiv:2501.08457, 2025.

13. G. K. Koru and Ç. Uluyol, “Detection of Turkish Fake News from Tweets with BERT Models,” IEEE Access, vol. 12, pp. 14918–14932, 2024.

14. A. Oad, M. H. Farooq, A. Zafar, B. A. Akram, R. Zhou, and F. Dong, “Fake News Classification Methodology with Enhanced BERT,” IEEE Access, vol. 12, pp. 164491– 164506, 2024.

15. A. Mishra, R. Singh, and P. Kumar, “Fake News Detection Using Hybrid Deep Learning Models: A Comparative Analysis,” IEEE Transactions on Computational Social Systems, vol. 11, no. 2, pp. 217–229,

16. 2024.