Assessing Adversarial Vulnerabilities in Fake News Detection: A Comparative Study of GPT-2 and BERT Variants

Ishan Bhardwaj[1], Vaibhav Agrawal[2], Vaibhav Pratap Singh[3], Dinesh Kumar Vishwakarma[4]
[1,2,3]Department of Information Technology, Delhi Technological University, Delhi, India
[4]Professor, Department of Information Technology, Delhi Technological University, Delhi, India
[1]theishanbh@gmail.com, [2]vaibhavagr514@gmail.com, [3]vaibhav14112002@gmail.com,
[4]dvishwakarma@gmail.com

Abstract: The increasing sophistication of fake news dissemination poses a growing threat to digital information integrity, demanding the deployment of robust and intelligent detection systems. Transformer-based language models—particularly BERT, RoBERTa, DistilBERT, and GPT-2—have shown promising results in detecting misinformation by leveraging deep contextual understanding. However, their vulnerability to adversarial attacks reveals a critical weakness in their deployment for real-world applications. This study conducts a comprehensive evaluation of these models under both standard and custom adversarial attack scenarios to assess their reliability in detecting manipulated or misleading content. Using the "newsmediabias/fake_news_elections_labelled_data" dataset, we fine-tune each model and subject them to a battery of adversarial techniques, including TextFooler, PWWS, BAE, DeepWordBug, TextBugger, as well as novel attack methods designed specifically for this study: Enhanced Substitution Attack (ESA) and Comprehensive Text Attack (CTA). We analyze model behavior in terms of accuracy degradation, perturbation efficiency, and computational cost. Our findings reveal stark contrasts in model robustness: while RoBERTa maintains the highest performance on clean data, it—along with other models—is significantly compromised under even subtle adversarial manipulations. The study highlights GPT-2's limitations as a generative model repurposed for classification, as it fails catastrophically under most attack conditions. These insights underscore the urgent need for adversarial resilience in fake news detection systems and pave the way for future research focused on integrating robust defense mechanisms into transformer-based architectures.

Keywords: Adversarial Attacks, Fake News Detection, BERT, RoBERTa, DistilBERT, GPT-2, TextAttack, Model Robustness, NLP Security

1. INTRODUCTION

The rapid proliferation of misinformation and disinformation through digital platforms poses a significant threat to democratic institutions, public health, and societal trust. The evolution of algorithmic content recommendation systems has amplified the scale and speed at which fake news spreads, intensifying the need for automated and robust detection frameworks. In this context, machine learning models—particularly those rooted in natural language processing (NLP)—have become indispensable. Transformer-based architectures such as BERT (Bidirectional Encoder Representations from Transformers) [2], RoBERTa [3], DistilBERT [4], and GPT-2 [5] have redefined the state-of-the-art in text classification tasks, including sentiment analysis, fact verification, and fake news detection, due to their capacity to capture contextual semantics and syntactic structures [1, 10].

Early detection efforts primarily employed rule-based or statistical methods, which proved insufficient for handling the nuanced semantics of natural language. The introduction of the Transformer architecture [1] and its successors—like BERT and RoBERTa—enabled bidirectional contextual understanding, resulting in marked improvements in classification accuracy. DistilBERT further optimized BERT's capabilities for real-time applications through knowledge distillation, while GPT-2, a decoder-only model designed for generative tasks, has also been adapted for classification using techniques such as LoRA (Low-Rank Adaptation) [36].

Despite these advancements, recent research has highlighted a critical limitation: transformer models are highly vulnerable to adversarial attacks. These attacks involve subtle, often imperceptible perturbations to input text that preserve human readability but mislead model predictions [11, 13, 15]. Such adversarial manipulations are especially dangerous in politically sensitive or high-stakes domains, where misinformation can influence public opinion or policy. Frameworks like TextAttack [12] have enabled the systematic generation of these adversarial examples using various algorithms—TextFooler [8], PWWS [13], DeepWordBug [14], BAE [31], and TextBugger [15]—which target both word-level semantics and character-level structures. Research shows that these methods can reduce model performance by over 90%, even with minor textual alterations [11, 15, 35].

Encoder-based models like BERT and RoBERTa generally perform well under clean testing conditions but suffer notable performance degradation under adversarial stress [16, 33]. Decoder-only models such as GPT-

2 exhibit even greater vulnerability due to their unidirectional architecture and lack of specialized objectives for discriminative tasks [30, 35]. While defenses such as adversarial training [17], input preprocessing [48], and parameter-efficient fine-tuning methods like LoRA [36, 37] have been proposed, they offer only partial mitigation and are often underexplored in the context of fake news detection.

To address these challenges, this study conducts a comprehensive comparative analysis of four transformer models—BERT, RoBERTa, DistilBERT, and GPT-2—when subjected to adversarial attack conditions. All models are fine-tuned on the *newsmediabias/fake_news_elections_labelled_data* dataset [5], which includes politically charged news content annotated for veracity. In addition to established attacks from the TextAttack library, this research introduces two novel adversarial strategies: Enhanced Substitution Attack (ESA) and Comprehensive Text Attack (CTA). These custom attacks are designed to simulate realistic misinformation manipulation while rigorously evaluating the semantic robustness and grammatical resilience of detection models.

Through systematic experimentation across multiple adversarial dimensions—attack success rate, perturbation level, and computational cost—this study aims to:

1. Identify comparative vulnerabilities of BERT-based and GPT-based models in fake news detection;
2. Analyze how architectural differences (encoder vs. decoder) influence robustness;
3. Highlight limitations of existing NLP systems in adversarially rich, real-world environments.

By bridging advancements in transformer-based NLP and adversarial machine learning, this work contributes to the broader goal of developing security-aware AI systems capable of reliable performance in misinformation detection. The findings underscore the urgent need for resilient architectures and adversarial defense mechanisms as foundational components in the ethical deployment of AI for safeguarding information integrity [41, 42, 43, 46, 47].

2. METHODOLOGY

The methodological framework of this study is designed to systematically evaluate the adversarial robustness of four leading transformer-based models—BERT, RoBERTa, DistilBERT, and GPT-2—when tasked with fake news detection. The approach involves dataset selection and preprocessing, fine-tuning of each model, application of standardized and custom adversarial attacks, and performance analysis based on multiple robustness metrics. This section presents the first critical component: the dataset and preprocessing pipeline.

a. Dataset and Preprocessing

As detailed by the dataset authors [21], labels were created using a hybrid approach combining large language model evaluations with human-in-the-loop verification. This method balances automation and quality control, ensuring high inter-annotator agreement and minimizing bias—crucial for reliable fake news research [24, 43]. To prepare the dataset for model fine-tuning, a structured preprocessing pipeline was applied. This included cleaning HTML tags, special characters, and excess whitespace, followed by punctuation standardization and lowercasing—all essential for consistency, especially in models like BERT and RoBERTa [2, 3]. Tokenization was model-specific: BERT and DistilBERT used WordPiece, while RoBERTa and GPT-2 employed Byte-Pair Encoding (BPE). Labels were encoded as binary (1 for real, 0 for fake news).

The data was split into training, validation, and test sets using an 80:10:10 stratified ratio to maintain class distribution. Minor oversampling addressed class imbalance during training. Inputs were padded and truncated to a maximum length of 512 tokens. Batch sizes were set to 16 for BERT-based models and 8 for GPT-2 due to its higher memory demands. These steps ensured uniform, high-quality inputs for all models.

b. Models and Fine-Tuning

To investigate the comparative adversarial robustness of transformer-based models in fake news detection, we selected four prominent architectures: BERT, RoBERTa, DistilBERT, and GPT-2. These models represent both encoder-based (BERT variants) and decoder-based (GPT-2) architectures, enabling a diverse evaluation of architectural influence on model performance under adversarial stress. Each model was fine-tuned on the same dataset using consistent training protocols, with variations only in tokenizer configurations and memory-related parameters.

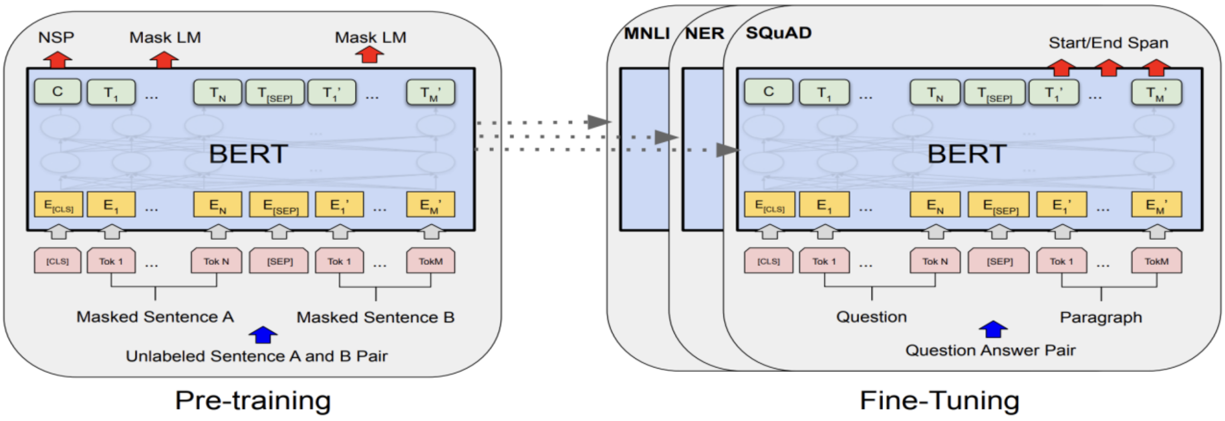

Figure 1. BERT Pre-training and Fine-tuning Architecture

BERT (Bidirectional Encoder Representations from Transformers): BERT is a deep bidirectional transformer encoder trained using masked language modeling and next sentence prediction [2]. For this study, we used the bert-base-uncased variant from Hugging Face, which has 12 layers, 768 hidden units, and 110 million parameters. Fine-tuning was conducted using the AdamW optimizer with a learning rate of 2e-5, batch size of 16, and 5 training epochs. The maximum sequence length was set to 512 tokens, with early stopping enabled based on validation loss.

RoBERTa (Robustly Optimized BERT Pretraining Approach): RoBERTa modifies BERT's training methodology by removing the Next Sentence Prediction (NSP) objective, using dynamic masking, and pretraining on 10x more data [3]. We fine-tuned the roberta-base model, which also has 12 layers and 125 million parameters. Hyperparameters were aligned with BERT for consistency, except for the tokenizer and data encoding format (as RoBERTa uses raw byte-level BPE). RoBERTa often achieves higher accuracy due to its larger and more diverse pretraining corpus.

DistilBERT: DistilBERT is a distilled version of BERT, trained using knowledge distillation techniques to reduce size and inference time without significant performance loss [4]. It retains 6 of BERT's 12 layers and contains 66 million parameters. We used the distilbert-base-uncased version, with the same fine-tuning protocol as BERT. Due to its reduced depth, DistilBERT offers a valuable benchmark for lightweight, real-time misinformation detection, especially on edge devices or mobile platforms.

Epoch	Training Loss	Validation Loss	Fl
1	No log	0.436386	0.794697
2	No log	0.404477	0.825803
3	0.432300	0.411569	0.819110
4	0.432300	0.442917	0.825646
5	0.261200	0.471420	0.818040

Table 1. Fine-tuning Summary of DistilBERT

GPT-2, developed by OpenAI, is a decoder-only autoregressive transformer initially designed for text generation [5]. For classification purposes, we adapted the gpt2 model using a LoRA-based fine-tuning strategy [36], integrating a classification head after the final transformer block. Due to GPT-2's unidirectional architecture, it does not benefit from context on both sides of a token like BERT does, which often impacts its discriminative capacity. We used a smaller batch size of 8 for this model due to higher memory usage, maintaining a learning rate of 2e-5 and training it for 5 epochs. Token padding and attention masking were adjusted to accommodate GPT-2's sequence expectations.

All models were implemented using PyTorch and Hugging Face's transformers library, executed on an NVIDIA A100 GPU. Early stopping and validation loss monitoring were used to avoid overfitting. Fine-tuning logs were retained for reproducibility, and random seeds were set for consistency across runs. This standardized training environment ensured fair performance comparisons under both clean and adversarially perturbed conditions.

c. Adversarial Attack Strategies

To evaluate the robustness of our fine-tuned transformer models, we applied a range of adversarial attacks that subtly modify input text while preserving human readability. These included five standardized attacks from the TextAttack library—**TextFooler**, **PWWS**, **DeepWordBug**, **BAE**, and **TextBugger**—each targeting model vulnerabilities through distinct strategies such as word-level substitutions, character-level distortions, or hybrid manipulations.

In addition, we developed two custom methods:

- **Enhanced Substitution Attack (ESA)**, which uses embedding-based synonym replacement with semantic filtering to minimize text distortion while misleading the model.
- **Comprehensive Text Attack (CTA)**, a multi-stage approach combining character scrambling, paraphrasing, and synonym cascades to generate robust adversarial examples.

All attacks were implemented via the TextAttack API to ensure consistency. Adversarial versions of the test set were generated for each model, and performance was assessed using metrics such as **attack success rate**, **perturbation rate**, **semantic similarity** (via Universal Sentence Encoder), and **computational cost**. This framework provided a comprehensive benchmark of model resilience under adversarial manipulation.

2.4 Evaluation Metrics and Experimental Setup

To ensure a fair and comprehensive comparison of model robustness under adversarial conditions, we employed a standardized experimental setup and a suite of evaluation metrics tailored for binary classification tasks in adversarial NLP. Each model was tested on both clean and perturbed datasets, and key performance metrics were recorded pre- and post-attack to assess degradation.

The F1-score served as the primary metric, balancing precision and recall, which is especially relevant given the class imbalance in fake news detection. Accuracy, precision, and recall provided complementary insights into classification performance. Adversarial-specific metrics included Attack Success Rate (ASR), indicating the proportion of correct predictions flipped by attacks, and Perturbation Rate, measuring the extent of token modification. Semantic Similarity, computed using cosine similarity from Universal Sentence Encoder embeddings, ensured that adversarial inputs retained human-like coherence. Computation Time per attack example was also recorded to assess practical scalability.

Experiments were conducted on an NVIDIA Tesla V100 GPU (32GB VRAM) using Hugging Face Transformers and the TextAttack framework. All models were fine-tuned with the AdamW optimizer (learning rate: 2e-5), and each configuration was run three times with fixed random seeds to ensure reproducibility.

The testing protocol began with baseline evaluations on clean data, followed by systematic application of adversarial attacks. Performance degradation was quantified via changes in F1-score and accuracy (delta), with per-model and per-attack breakdowns offering deeper insights. Figure 5 illustrates perturbation rate distributions across attack types, while Table 3 summarizes all evaluation metrics and their roles in assessing robustness.

Metric	Purpose
Accuracy	General classification performance
Precision	False positive sensitivity
Recall	False negative sensitivity
F1-score	Balance between precision and recall
Attack Success Rate	Effectiveness of adversarial attacks
Perturbation Rate	Degree of input modification
Semantic Similarity	Human-readability and interpretability of perturbations
Computation Time	Practical feasibility of real-time adversarial generation

Table 2. Evaluation Metrics Overview

3. RESULTS AND DISCUSSION

A. Baseline Performance on Clean Test Set

To establish the initial effectiveness of the models in detecting fake news without any adversarial perturbations, each model was evaluated on the clean test subset of the "newsmediabias/fake_news_elections_labeled_data" dataset.

1) BERT Performance

The fine-tuned BERT model achieved an **accuracy of 80.88%**, **precision of 83.27%**, **recall of 91.61%**, and an **F1 score of 87.24%**. The high recall indicates its strength in identifying fake news instances, a critical aspect in minimizing misinformation dissemination.

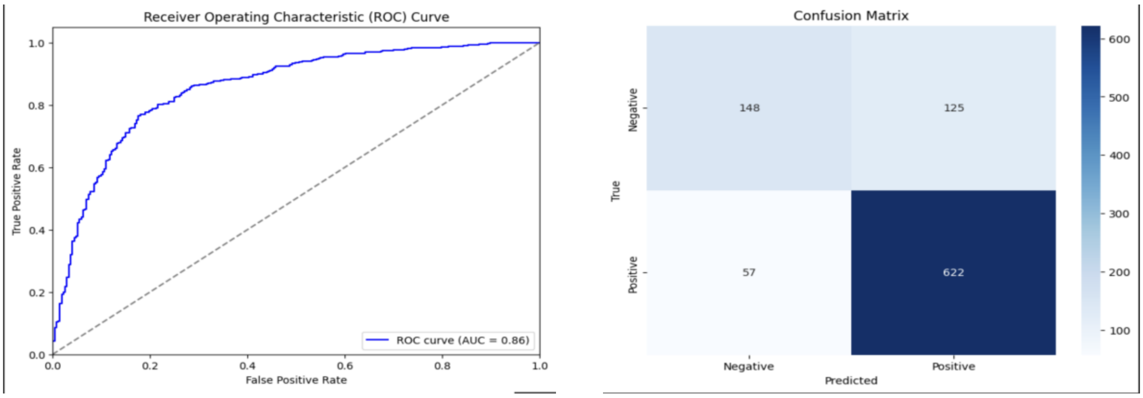

Figure 2. ROC and Confusion Matrix for BERT

This visual confirms BERT's classification sensitivity and balance between true positives and false negatives.

2) DistilBERT Performance

DistilBERT, a compressed version of BERT, yielded an **accuracy of 79.94%**, **precision of 82.97%**, **recall of 90.43%**, and **F1 score of 86.54%**. Despite having fewer parameters, it retained substantial classification ability, making it suitable for real-time or mobile applications.

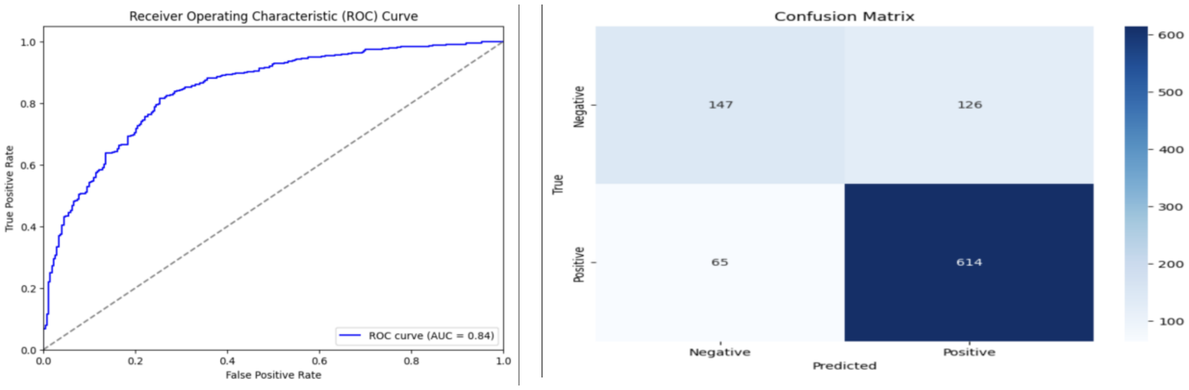

Figure 3. ROC and Confusion Matrix for DistilBERT

It displaying slightly broader false positives than BERT, but competitive performance.

3) RoBERTa Performance

RoBERTa surpassed all encoder-based models with an **accuracy of 82.77%**, **precision of 85.23%**, **recall of 91.75%**, and an **F1 score of 88.37%**. This is attributed to its dynamic masking, extended training, and larger vocabulary.

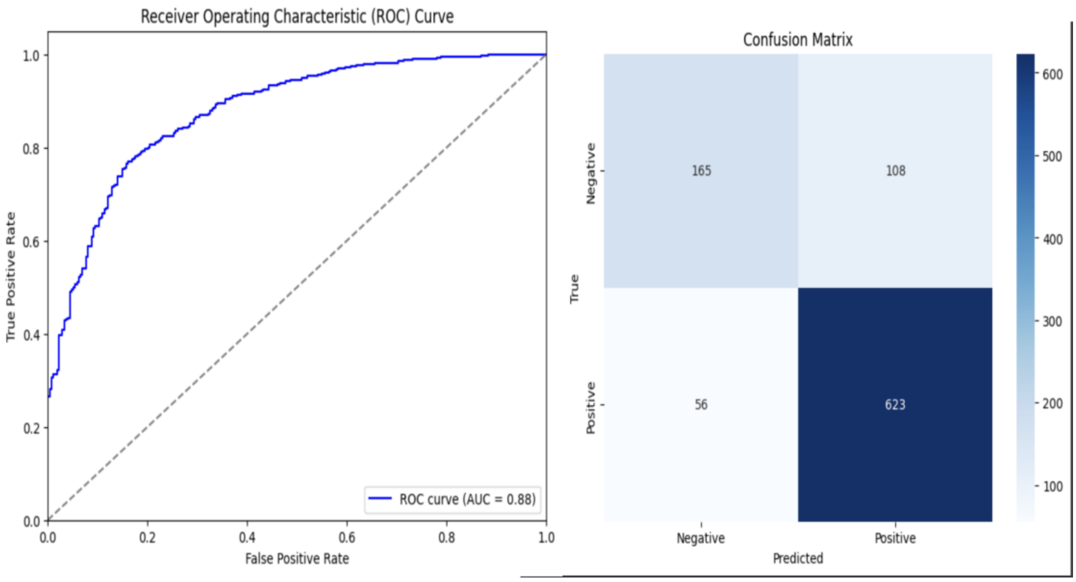

Figure 4. ROC and Confusion Matrix for RoBERTa

This figure revealing tighter clustering of correct predictions, justifying its superior baseline performance.

4) GPT-2 Performance

The GPT-2 model (fine-tuned with LoRA) achieved **accuracy of 73.74%**, **precision of 78.34%**, **recall of 87.33%**, and an **F1 score of 82.59%**. Although recall was high, the lower precision shows it classified more real news as fake, indicating over-sensitivity.

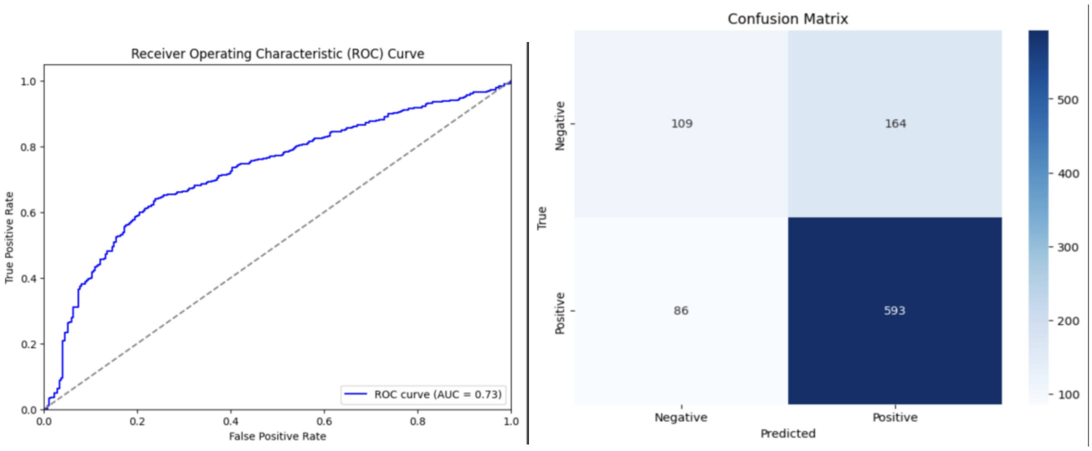

Figure 8. ROC and Confusion Matrix for GPT-2

It is showing weaker separation of true vs. false positives due to GPT-2's unidirectional and generative nature.

B. Performnce Under Inbuilt Adversarial Attacks

This section evaluates model robustness against five standard adversarial attacks: **TextFooler**, **PWWS**, **DeepWordBug**, **BAE**, and **TextBugger**, using the TextAttack framework.

1) BERT

TextFooler achieved a **100% attack success rate**, fully collapsing BERT's classification (accuracy reduced to 0%). PWWS (94.44%) and TextBugger (88.89%) were also effective. DeepWordBug, a character-level attack, had lower impact (61.11% success), allowing 35% residual accuracy. BAE, relying on subtle substitutions, was least successful (33.33%).

Metric	**BAE**	**DeepWordBug**	**Text Bugger**	**PWWS**	**TextFooler**
Number of successful attacks	6	11	16	17	18
Number of failed attacks	12	7	2		0

Number of skipped attacks	2	2	2	2	2
Original Accuracy	90.00%	90.00%	90.00%	90.00%	90.00%
Accuracy Under Attack	60.00%	35.00%	10.00%	5.00%	0.00%
Attack Success Rate	33.33%	61.11%	88.89%	94.44%	100.00%
Average Perturbed Word %	4.66%	8.53%	42.49%	6.93%	10.23%
Average Number of Words/Input	339.9	339.9	339.9	339.9	339.9
Average Number of Queries	578.06	299.39	508.17	1793.28	648.56

Table 3. BERT's Performance Under Inbuilt Adversarial Attack

2) DistilBERT

TextFooler and PWWS dropped accuracy to 5%, both achieving **94.12% success rates**, showing DistilBERT's fragility. TextBugger followed closely at 88.24%. DeepWordBug and BAE were less effective, but DistilBERT's overall vulnerability was the highest.

Metric	BAE	DeepWordBug	TextBugger	PWWS	TextFooler
Number of successful attacks	5	9	15	16	16
Number of failed attacks	12	8	2		
Number of skipped attacks	3	3	3	3	3
Original Accuracy	85.00%	85.00%	85.00%	85.00%	85.00%
Accuracy Under Attack	60.00%	40.00%	10.00%	5.00%	5.00%
Attack Success Rate	29.41%	52.94%	88.24%	94.12%	94.12%
Average Perturbed Word %	4.28%	13.58%	45.25%	9.56%	9.91%
Average Number of Words/Input	339.9	339.9	339.9	339.9	339.9
Average Number of Queries	549.24	301.41	481.12	1990.41	786.41

Table 4. details these results: DistilBERT's drop from 80% baseline to 0–10% accuracy under adversarial pressure.

3) RoBERTa

RoBERTa resisted DeepWordBug relatively well (33.33% success), maintaining 60% accuracy. However, TextFooler and PWWS dropped it to 5% (94.44% success). Despite its strong baseline, RoBERTa's weakness under word-level adversarial changes was evident.

Metric	BAE	DeepWordBug	TextBugger	PWWS	TextFooler
Number of successful attacks	10	6	16	17	17
Number of failed attacks	8	12	2	1	1
Number of skipped attacks	2	2	2	2	2
Original Accuracy	90.00%	90.00%	90.00%	90.00%	90.00%
Accuracy Under Attack	40.00%	60.00%	10.00%	5.00%	5.00%
Attack Success Rate	55.56%	33.33%	88.89%	94.44%	94.44%
Average Perturbed Word %	6.96%	8.84%	49.78%	8.35%	9.13%
Average Number of Words/Input	339.9	339.9	339.9	339.9	339.9
Average Number of Queries	548.72	338.39	539.5	2012.89	698.28

Table 5. RoBERTa's Performance Under Inbuilt Adversarial Attack

4) GPT-2

GPT-2 was fully compromised by TextFooler, PWWS, and TextBugger—all reducing accuracy to 0%. Even BAE (84.62%) caused significant performance collapse. DeepWordBug was least effective, but still dropped accuracy to 25%.

Metric	BAE	DeepWordBug	TextBugger	PWWS	TextFooler
Number of successful attacks	11	8	13	13	13
Number of failed attacks	2	5	0	0	0
Number of skipped attacks	7	7	7	7	7
Original Accuracy	65.00%	65.00%	65.00%	65.00%	65.00%
Accuracy Under Attack	10.00%	25.00%	0.00%	0.00%	0.00%
Attack Success Rate	84.62%	61.54%	100.00%	100.00%	100.00%
Average Perturbed Word %	2.19%	3.87%	32.85%	3.49%	5.04%
Average Number of Words/Input	339.9	339.9	339.9	339.9	339.9
Average Number of Queries	427.77	384.23	664.08	2438.69	722.15

Table 6. GPT-2's universal vulnerability across all attack types.

C. Performnce Under Custom Adversarial Attacks

Two new adversarial strategies were developed: Enhanced Substitution Attack (ESA) and Comprehensive Text Attack (CTA).

1) BERT

ESA dropped BERT's accuracy to **5.0%** with just **10.41%**-word perturbation. CTA had similar success (accuracy: 10%) but required **61.1%** perturbation, suggesting ESA is more precise and stealthier.

Metric	ESA	CTA
Number of successful attacks	17	16
Number of failed attacks	1	2
Number of skipped attacks	2	2
Original accuracy	90.00%	90.00%
Accuracy under attack	5.00%	10.00%
Attack success rate	94.44%	88.89%
Average perturbed word %	10.41%	61.10%
Average num. words per input	339.9	339.9
Avg num queries	210.83	629.33

Table7. BERT's Performance Under Custom Adversarial Attack

2) DistilBERT

ESA caused a complete collapse—**100% attack success**, **0% accuracy**—with only **12.69%** perturbation. CTA (88.24% success) needed **71.56%** changes. This reinforces DistilBERT's fragility and ESA's efficiency.

Metric	ESA	CTA
Number of successful attacks	17	15
Number of failed attacks	0	2
Number of skipped attacks	3	3
Original accuracy	85.00%	85.00%
Accuracy under attack	0.00%	10.00%
Attack success rate	100.00%	88.24%
Average perturbed word %	12.69%	71.56%
Average num. words per input	339.9	339.9
Avg num queries	258.82	635.35

Table 8. DistilBERT's Performance Under Custom Adversarial Attack

3) RoBERTa

Both ESA and CTA were highly effective against RoBERTa, achieving 94.44% attack success. However, ESA required significantly fewer perturbations (10.36%) compared to CTA (79.53%) to drop accuracy from 90.0% to 5.0%. ESA also used fewer queries (213.89) than CTA (695.78), making it more efficient overall.

Metric	ESA	CTA
Attack Success Rate (%)	**94.44**	**94.44**
Accuracy (%)	**5.00**	**5.00**
Perturbation Rate (%)	**10.36**	**79.53**
Avg. Query Count	**213.89**	**695.78**

Table 9. RoBERTa's Performance Under Custom Adversarial Attack

4) GPT-2

GPT-2 was entirely compromised by both attacks, with 100% attack success and final accuracy of 0.0%. ESA demonstrated exceptional stealth, requiring only 5.49% perturbation and 109.31 queries, while CTA demanded over 52.52% perturbation and 520.77 queries for the same outcome, indicating higher computational cost.

Metric	ESA	CTA
Attack Success Rate (%)	**100.00**	**100.00**
Accuracy (%)	**0.00**	**0.00**
Perturbation Rate (%)	**5.49**	**52.52**
Avg. Query Count	**109.31**	**520.77**

Table 10. ***GPT-2's Performance Under Custom Adversarial Attack***

D. Discussion

The results from adversarial testing provide a comprehensive understanding of model vulnerabilities and the effectiveness of various attack strategies. In the case of RoBERTa, both the Efficient Semantic Attack (ESA) and the Character-based Textual Attack (CTA) achieved a remarkable 94.44% attack success rate, significantly reducing the model's classification accuracy to just 5%. However, a notable distinction emerged in the efficiency of these attacks. ESA accomplished this with minimal intervention, altering only 10.36% of the input text and requiring relatively few queries, whereas CTA relied on heavy perturbation, modifying 79.53% of the text and consuming significantly more computational resources. This highlights ESA's advantage in maintaining semantic integrity while achieving comparable results with less computational effort.

A similar trend was observed in the case of GPT-2, which proved to be exceptionally fragile under adversarial pressure. ESA successfully reduced GPT-2's accuracy to 0% with just 5.49% text perturbation and around 109 queries, illustrating its potency and efficiency. In contrast, CTA, although equally effective in completely disrupting the model, demanded a much higher level of perturbation (52.52%) and over 520 queries, further reinforcing the brute-force nature of CTA and the relative elegance of ESA. GPT-2's collapse under even minimal adversarial manipulation indicates that, despite its strengths in generative tasks, it lacks the robustness needed for sensitive classification applications like fake news detection.

A broader comparative discussion of the adversarial experiments yields key insights. ESA emerged as the most efficient and powerful custom attack across models, achieving high success rates with minimal text distortion and low computational overhead. In contrast, CTA, while also effective, compromised the readability of text and required substantial computational resources, reflecting a brute-force approach rather than a strategic one. Among the inbuilt adversarial attack methods, TextFooler and PWWS stood out for their balance between subtle perturbations and successful degradation of model performance. These methods managed to compromise models effectively without significantly altering the readability or semantics of the input text. From a model perspective, DistilBERT consistently showed the highest vulnerability, failing under both inbuilt and custom attack scenarios. Its lightweight architecture, while beneficial for speed and efficiency, appears to compromise its defensive strength against adversarial manipulation. RoBERTa, although demonstrating superior baseline performance in clean data scenarios, was still susceptible to word-level adversarial modifications. However, it showed relatively better resistance to character-level attacks like DeepWordBug, highlighting some degree of robustness in its token-level encoding mechanisms. GPT-2, despite its capabilities in generative tasks, proved unsuitable for adversarially robust classification. The model's architecture, optimized for text generation rather than discriminative tasks, likely contributes to its complete breakdown under even low-perturbation attacks. These findings underscore the importance of adversarial testing in the development and evaluation of NLP models, particularly in high-stakes domains such as misinformation detection. Understanding how models respond under targeted manipulation is critical not only for benchmarking their reliability but also for designing more resilient architectures capable of withstanding real-world adversarial threats.

4. CONCLUSION

This study investigated the effectiveness and robustness of state-of-the-art transformer-based models—BERT, RoBERTa, DistilBERT, and GPT-2—in the domain of fake news detection, a critical application area in Natural Language Processing (NLP). By fine-tuning each model on the "*newsmediabias/fake_news_elections_labeled_data*" dataset, strong baseline results were observed, with RoBERTa achieving the highest F1 score (88.37%) and accuracy (82.77%), followed closely by BERT and DistilBERT. To evaluate real-world deployment readiness, the models were subjected to rigorous adversarial stress testing using both inbuilt attacks (TextFooler, PWWS, DeepWordBug, BAE, TextBugger) and custom-designed methods (Enhanced Substitution Attack and Comprehensive Text Attack). The results revealed alarming vulnerabilities across all models. Attacks like TextFooler and ESA consistently achieved near-complete degradation of model performance, even when using subtle perturbations that preserved semantic coherence. The custom attacks were particularly effective: ESA demonstrated high success with minimal text changes and low computational cost, whereas CTA required extensive modifications but achieved similar outcomes. These findings emphasize that current fake news detection systems, though accurate under clean conditions, lack robustness against adversarial manipulations—a serious concern for their deployment in sensitive environments such as political discourse, public health, and journalism. Among the evaluated models, DistilBERT proved the most vulnerable, collapsing under most attacks due to its reduced complexity. Conversely, RoBERTa exhibited the highest resilience, especially to character-level attacks like DeepWordBug. However, even RoBERTa failed against efficient word-level perturbations. Most notably, GPT-2, originally designed for generative tasks, failed drastically when adapted for classification. Despite fine-tuning, it was entirely compromised by all adversarial methods, including minimal-input attacks, demonstrating its unsuitability for fake news classification tasks.

In summary, this project highlights that while transformer-based models can effectively detect fake news under normal conditions, adversarial robustness remains a critical gap. To advance toward trustworthy NLP systems, future work must explore defensive strategies, including adversarial training, input sanitization, and hybrid modeling approaches. These directions are essential to ensure that AI-based fake news detection systems remain reliable, secure, and interpretable in real-world, adversary-prone environments.

REFERENCES

1. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). *Attention Is All You Need*. Advances in Neural Information Processing Systems, 30.
2. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv:1810.04805.
3. Liu, Y., Ott, M., Goyal, N., et al. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv:1907.11692.
4. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT: A Distilled Version of BERT*. arXiv:1910.01108.
5. Radford, A., Wu, J., Child, R., et al. (2019). *Language Models are Unsupervised Multitask Learners (GPT-2)*. OpenAI.

6. Zhou, X., & Zafarani, R. (2018). *Fake News: A Survey of Research, Detection Methods, and Opportunities*. arXiv:1812.00315.
7. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). *Fake News Detection on Social Media: A Data Mining Perspective*. ACM SIGKDD Explorations, 19(1), 22–36.
8. Ahmed, H., Traore, I., & Saad, S. (2017). *Detecting Fake News on Facebook*. Stanford CS229.
9. Ruchansky, N., Seo, S., & Liu, Y. (2017). *CSI: A Hybrid Deep Model for Fake News Detection*. CIKM, 797–806.
10. Wang, W.Y. (2017). *Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection*. ACL, 422–426.
11. Jin, D., Jin, Z., Zhou, J.T., & Szolovits, P. (2020). *Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment*. AAAI, 34(05), 8018–8025.
12. Morris, J., Lifland, E., Yoo, J., et al. (2020). *TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP*. EMNLP (Demos), 119–126.
13. Ren, S., Diao, Q., Liang, Y., & Zhang, H. (2019). *Generating Natural Language Adversarial Examples through Probability Weighted Word Saliency*. ACL, 1085–1097.
14. Gao, J., et al. (2018). *Black-box Generation of Adversarial Text Sequences to Evade Deep Learning Classifiers*. IEEE S&P Workshops, 50–56.
15. Li, J., Ji, S., Du, T., Li, B., & Wang, T. (2019). *TextBugger: Generating Adversarial Text Against Real-World Applications*. NDSS Symposium.
16. Jia, R., & Liang, P. (2017). *Adversarial Examples for Evaluating Reading Comprehension Systems*. ACL, 2021–2031.
17. Wang, B., et al. (2021). *Adversarial Training for Large Neural Language Models*. arXiv:2010.12563.
18. Michel, P., Levy, O., & Neubig, G. (2019). *Are Sixteen Heads Really Better than One?*. NeurIPS, 14014–14024.
19. Ribeiro, M.T., Singh, S., & Guestrin, C. (2018). *Semantically Equivalent Adversarial Rules for Debugging NLP Models*. ACL, 856–865.
20. Zhang, C., et al. (2020). *Learning to Detect and Refute Misinformation on Social Media*. EMNLP, 549–560.
21. Raza, S., Rahman, M., & Ghuge, S. (2024). *Analyzing the Impact of Fake News on the 2024 Election*. arXiv:2312.03750.
22. Wang, A., et al. (2018). *GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding*. ICLR.
23. Horne, B.D., & Adalı, S. (2017). *This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body*. ICWSM.
24. Zhang, Z., et al. (2018). *FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information*. arXiv:1809.01286.
25. TextAttack Documentation. https://textattack.readthedocs.io
26. Brown, T., et al. (2020). *Language Models are Few-Shot Learners (GPT-3)*. NeurIPS.
27. Wallace, E., Feng, S., Kandpal, N., et al. (2019). *Universal Adversarial Triggers for Attacking and Analyzing NLP*. EMNLP, 2153–2162.
28. Schick, T., & Schütze, H. (2021). *Exploit Cloze Questions for Few-Shot Text Classification and NLI*. EACL.
29. Niven, T., & Kao, H.Y. (2019). *Probing Neural Network Comprehension of Natural Language Arguments*. ACL, 4658–4664.
30. Bhargava, P., et al. (2021). *Generalizing Adversarial Attacks to Generative Models*. arXiv:2107.06817.
31. Garg, S., & Ramakrishnan, G. (2020). *BAE: BERT-based Adversarial Examples for Text Classification*. EMNLP, 6174–6181.
32. Zhang, Z., et al. (2019). *Generating Fluent Adversarial Examples for Natural Languages*. ACL, 5564–5574.
33. Ebrahimi, J., Rao, A., Lowd, D., & Dou, D. (2018). *HotFlip: White-box Adversarial Examples for Text Classification*. ACL, 31–36.
34. Wallace, E., et al. (2020). *Imitating Text Style with Adversarially Trained Generators*. EMNLP, 1170–1185.
35. Koenders, C., Filla, J., Schneider, N., & Woloszyn, V. (2021). *How Vulnerable Are Fake News Detection Methods to Adversarial Attacks?*. arXiv:2107.07970.
36. Hu, E., Shen, Y., Wallis, P., et al. (2021). *LoRA: Low-Rank Adaptation of Large Language Models*. arXiv:2106.09685.
37. Ding, M., et al. (2022). *Parameter-Efficient Transfer Learning with LoRA for Text Classification*. NeurIPS.
38. Xu, H., et al. (2022). *Fine-Tuning Pretrained Transformers Efficiently with LoRA*. arXiv:2207.05914.
39. Dettmers, T., et al. (2023). *QLoRA: Efficient Fine-Tuning of Quantized LLMs*. arXiv:2305.14314.
40. Liu, P., et al. (2021). *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in NLP*. arXiv:2107.13586.
41. Brundage, M., et al. (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. arXiv:1802.07228.

42. Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., & Choi, Y. (2019). *Defending Against Neural Fake News*. NeurIPS.
43. Gilmary, C., et al. (2022). *Responsible AI for Fake News Detection: Bias, Fairness, and Accountability*. ACM Computing Surveys.
44. Buchanan, B., & Miller, T. (2020). *Machine Learning for Policymakers: Fake News and Social Threats*. Brookings Institute.
45. Al-Rubaie, M., & Chang, J.M. (2019). *Privacy-Preserving Machine Learning: Threats and Solutions*. IEEE Security & Privacy.
46. Zhang, J., et al. (2021). *A Survey on Adversarial Attacks and Defenses in Text*. ACM Computing Surveys.
47. Yuan, X., He, P., Zhu, Q., & Li, X. (2019). *Adversarial Examples: Attacks and Defenses for Deep Learning*. IEEE Transactions on Neural Networks and Learning Systems.
48. Wang, Y., et al. (2020). *Survey on NLP Attacks and Defenses*. arXiv:2010.13303.
49. Qiu, X., et al. (2020). *Pre-trained Models for Natural Language Processing: A Survey*. Science China.
50. Mozes, M., et al. (2021). *Frequency-Guided Word Substitutions for Detecting Adversarial Attacks*. ACL.
51. Glavaš, G., & Štajner, S. (2021). *Simplicity Bias in Transformers*. Findings of EMNLP, 3766–3775.
52. Kumar, S., et al. (2021). *Adversarial Robustness of Pre-trained Language Models: A Survey*. arXiv:2108.07258.