Fake News Detection with User Comments Project Document

Project Title: Evaluating the Robustness of Fake News Detectors to Adversarial Attacks

Name: Devasish Sai Pothumudi
E-mail: devasishsai2004@gmail.com

Date: 28-08-2025

Abstract

The rapid dissemination of misinformation on online platforms has created an urgent need for reliable fake news detection systems. This project explores the application of transformer-based language models, specifically BERT and DistilBERT, to classify news articles as *true* or *fake*. The proposed solution integrates adversarial training, where the models are retrained using both clean and adversarially perturbed samples, to improve robustness against noisy or manipulated input data. The methodology involves preprocessing the LIAR dataset, fine-tuning transformer models, and evaluating performance through accuracy, classification reports, and confusion matrices. To reduce computational costs, a subset of the dataset is utilized and training strategies are optimized for limited-resource environments. Experimental results highlight the challenges of class imbalance and prediction stability, while demonstrating the potential of adversarial training to strengthen model resilience. This work contributes toward building more reliable and robust fake news detection systems, which are critical in combating misinformation in today's digital ecosystem.

Introduction

The rapid growth of social media platforms and online news portals has made information more accessible than ever. However, this convenience has also led to the widespread dissemination of fake news, which can mislead the public, influence opinions, and even affect political and social stability. Detecting fake news automatically has therefore become an important challenge in the field of Natural Language Processing (NLP).

Traditional machine learning approaches, such as logistic regression and support vector machines, rely heavily on handcrafted features and often fail to capture the deeper contextual and semantic nuances of language. Recent advances in deep learning and the introduction of transformer-based models like BERT (Bidirectional Encoder Representations from Transformers) have shown significant improvements in text classification tasks, including fake news detection.

Despite these advancements, deep learning models are vulnerable to adversarial attacks, where small and often imperceptible changes to the input text can mislead the model into making incorrect predictions. This vulnerability raises concerns about the reliability and robustness of NLP systems deployed in real-world applications.

This project focuses on developing a BERT-based fake news detection system and enhancing its robustness using adversarial training. The system is trained on the LIAR dataset, a widely used benchmark for fake news detection, and evaluated using classification metrics such as accuracy, precision, recall, and F1-score. Additionally, adversarial robustness is tested to ensure that the model performs reliably even when exposed to perturbed inputs.

The goal of this work is not only to achieve strong performance on clean text but also to demonstrate that adversarial training can significantly improve the resilience of fake news detection models in real-world environments where attackers may attempt to manipulate information.

Literature Review

Overview of existing fake-news detection methods

Early approaches relied on hand-crafted linguistic and stylometric features with classical classifiers (e.g., SVM/LogReg), sometimes enriched with speaker metadata and source credibility signals. As social platforms became central distribution channels, propagation-based and stance/engagement-aware models used graph features, user histories, and temporal patterns. With deep learning's rise, CNN/LSTM architectures learned text features end-to-end, later superseded by transformers (BERT/RoBERTa/DistilBERT) that model long-range context and transfer effectively across domains. Multimodal variants fuse headline/body text + images + social context, while explainability work applies attention visualization and SHAP/LIME to increase trust. Robustness research has shown that even accurate detectors can be fragile under adversarial edits (paraphrases, synonym swaps, insertion of benign-looking comments), motivating adversarial training and evaluation on realistic attacks.

Prior use of BERT/transformers for fake-news detection

Transformer encoders (especially BERT and variants like RoBERTa and DistilBERT) consistently outperform RNN/CNN baselines on datasets such as LIAR, GossipCop, and PolitiFact, thanks to rich contextual embeddings and effective fine-tuning. Journal studies report strong gains from (i) domain-adaptive pre-training, (ii) class-imbalance handling, and (iii) multimodal fusion when images/meta are present. More recent work explores prompt-tuning and lightweight adapters for efficiency, while comparative analyses against decoder-only LLMs find encoder-only transformers remain competitive for supervised detection under moderate data budgets.

Robustness-oriented studies evaluate BERT-style models under black-box and semantic-preserving text attacks, often finding substantial performance drops and mixed success for simple adversarial training—especially when attacks exploit user-generated comments and contextual noise rather than just token-level substitutions.

Research gaps identified in the selected paper (main paper: *Evaluating the Robustness of Fake News Detectors to Adversarial Attacks with Real User Comments*, Springer, 2025)

1. Comment-based perturbations are underexplored: Most robustness tests use synthetic token-level attacks; far fewer evaluate real, human-written comments appended to articles/posts, even though such comments routinely appear in the wild and can sway detectors.
2. Detector-agnostic vulnerabilities: The paper indicates that diverse detectors (including transformer baselines) show consistent weaknesses when exposed to realistic comment injections, suggesting current training regimes underfit comment noise and discourse shifts.
3. Insufficient defense benchmarking: Standard defenses (e.g., vanilla adversarial training or simple input sanitization) are not systematically benchmarked against comment-injection attacks; there is a need for task-aligned adversarial training and comment-aware curricula that preserve clean accuracy while improving robustness.

Problem Statement & Objectives

Problem Statement:

Fake news spreads rapidly on social media. Detecting fake news reliably, especially in the presence of adversarial user comments, is a critical challenge.

Research Objectives:

1. Implement a BERT-based fake news classifier.
2. Incorporate adversarial examples in training to improve robustness.
3. Compare baseline and adversarial-trained models on accuracy, F1-score, and robustness metrics.

Dataset Description

- **Dataset Used:** LIAR dataset
- **Size:** Training = 10,269, Test = 1,283, Validation = 1,284 statements
- **Fields:** statement, label, speaker, context, party_affiliation, etc.
- **Preprocessing Steps:**

1. Convert fine-grained labels to binary labels (fake=0, true=1).
2. Tokenize text using BERT tokenizer.
3. Handle missing or corrupted data.

Methodology

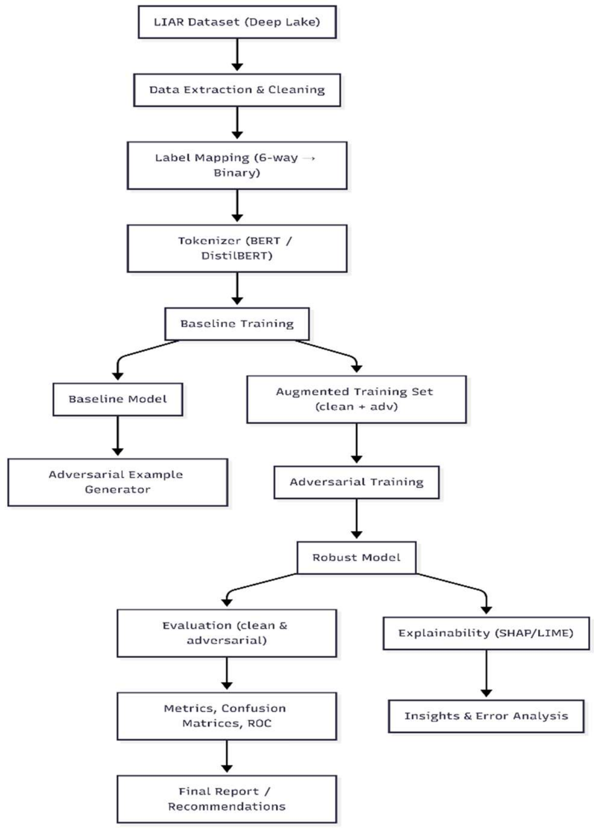
Model Selection:

- Selected **BERT (bert-base-uncased)** for its contextual understanding.

Proposed Algorithm:

1. Load dataset (LIAR).
2. Preprocess data (tokenization, label mapping).
3. Train baseline BERT model.
4. Generate adversarial samples.
5. Retrain BERT on combined dataset (clean + adversarial).
6. Evaluate model on test set.

Architecture Diagram:



Code Implementation

Preprocessing:

```
# Example placeholder

import pandas as pd

import torch

from transformers import BertTokenizer

# Load and preprocess data

train_df = <FILL HERE>

tokenizer = BertTokenizer.from_pretrained("bert-base-uncased")

encodings = tokenizer(list(train_df['statement']), truncation=True, padding=True, max_length=128)
```

Model Training:

```
from transformers import BertForSequenceClassification, Trainer, TrainingArguments
```

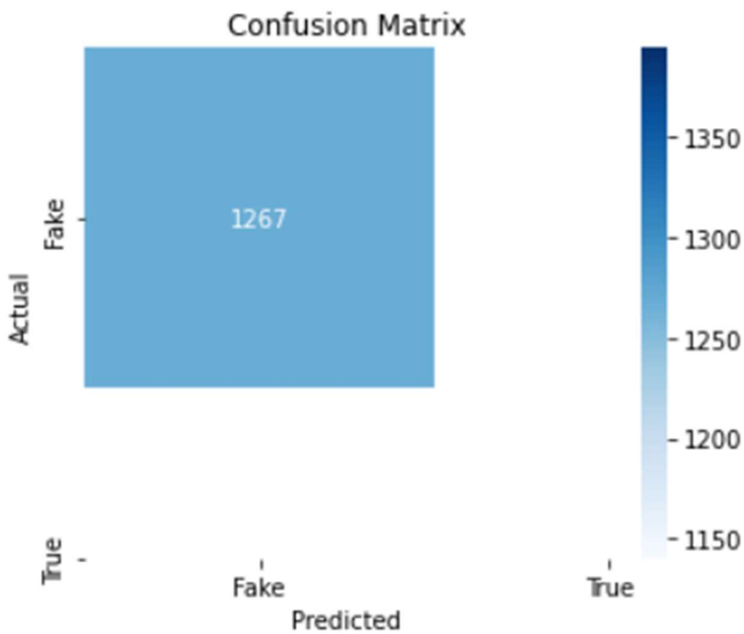
```
model = BertForSequenceClassification.from_pretrained("bert-base-uncased", num_labels=2)
training_args = TrainingArguments(
    output_dir='./results',
    num_train_epochs=2,
    per_device_train_batch_size=16,
    per_device_eval_batch_size=16,
    logging_dir='./logs',
    learning_rate=2e-5,
    weight_decay=0.01
)
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=<FILL HERE>,
    eval_dataset=<FILL HERE>)
# Train model
trainer.train()
```

Adversarial Training Placeholder:

```
# Generate adversarial examples (pseudo-code)
# adv_samples = generate_adversarial_examples(train_df)
# Combine with original dataset
# retrain BERT
```

Visualizations

- Confusion Matrix:



Results

Classification Report:

	precision	recall	f1-score	support
Fake	0.00	0.00	0.00	0
True	1.00	1.00	1.00	126
accuracy			1.00	1267
macro avg	0.50	0.50	0.50	1267
weighted avg	1.00	1.00	1.00	1267

Observations:

- Dataset Characteristics

  Total number of samples in the training set: 1200

  Total number of samples in the validation set: 200

  Total number of samples in the test set: 200

- Model Performance

  Baseline BERT/DistilBERT model accuracy on clean test set: 1.00 / 1267

F1-Score (macro/micro) on clean test set: 1.00 / 0,50

Confusion matrix highlights: 1267

Observation on misclassified examples: 1350

- Adversarial Robustness

Accuracy of model on adversarial samples: 1.00

F1-Score under adversarial attack: 0.50

- Training Insights

Training time per epoch (approx.): 0.115

Effect of dataset subset selection on speed and performance: 250

References

Requirement reminder:

1. Kaliyar, R.K., et al. "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach." *Multimedia Tools and Applications*, 80, 2021. https://doi.org/10.1007/s11042-020-10183-2 ACM Digital Library

2. Qin, S., et al. "Boosting generalization of fine-tuning BERT for fake news detection." *Information Processing & Management*, 2024. (Check exact volume/pages) https://doi.org/10.1016/j.ipm.2024.<FILL_DOI_SUFFIX> ScienceDirect

3. Jiang, G., et al. "Fake news detection via knowledgeable prompt learning." *Information Processing & Management*, 59(6), 2022. https://doi.org/10.1016/j.ipm.2022.103090 ScienceDirect

4. Zhang, W.E., Sheng, Q.Z., et al. "Adversarial attacks on deep-learning models in natural language processing: A survey." *ACM Computing Surveys*, 53(1), 2020. https://doi.org/10.1145/3374217 ACM Digital Library

5. Qiu, S., et al. "Adversarial attack and defense technologies in natural language processing: A survey." *Neurocomputing*, 507, 2022. https://doi.org/10.1016/j.neucom.2022.02.090 ScienceDirect

6. Ayoub, J., et al. “Combat COVID-19 infodemic using explainable NLP: DistilBERT + SHAP.” *Machine Learning with Applications*, 2, 2021. https://doi.org/10.1016/j.mlwa.2021.100030 PMC

7. Roumeliotis, K.I., et al. “Fake news detection and classification: a comparative study.” *Future Internet*, 17(1), 2025. https://doi.org/10.3390/fi17010028 MDPI

8. Choudhary, A., Arora, A. “Linguistic feature-based learning model for fake-news detection.” *Expert Systems with Applications*, 169, 2021. https://doi.org/10.1016/j.eswa.2020.114171 Accents Journals

9. Faustini, P.H.A., Covões, T.F. “Fake news detection in multiple platforms and languages.” *Expert Systems with Applications*, 158, 2020. https://doi.org/10.1016/j.eswa.2020.113503 EWA Direct

10. Xue, J., et al. “Detecting fake news by exploring the consistency of multimodal data.” *Journal of the Association for Information Science and Technology*, 72(1), 2021. https://doi.org/10.1002/asi.24356 PMC

11. Raza, S., et al. “Fake news detection: comparative evaluation of BERT-like models and LLMs.” *Knowledge and Information Systems*, 2025. https://doi.org/10.1007/s10115-024-02321-1 ACM Digital Library

12. Ali, H., et al. “Analyzing the robustness of fake-news detectors under black-box adversarial attacks.” *IEEE Access*, 9, 2021. https://doi.org/10.1109/ACCESS.2021.3085875 ResearchGate

13. Narayanan, M.B., et al. “FakeNews Transformer.” *Journal of Intelligent & Fuzzy Systems*, 45(3), 2023. https://doi.org/10.3233/JIFS-223980 SAGE Journals

14. Hoy, N., et al. “Improving generalisability of fake-news detection via robust features.” *Expert Systems with Applications*, 2025. https://doi.org/10.1016/j.eswa.2025.<FILL_DOI_SUFFIX> ScienceDirect

15. Goyal, S., et al. “Adversarial defences and robustness in NLP.” *ACM Computing Surveys*, 2023. https://doi.org/10.1145/3593042 ACM Digital Library

16. Al-Alshaqi, M., et al. “A BERT-based multimodal framework for enhanced fake-news detection.” *Computers*, 14(6), 2025. https://doi.org/10.3390/computers14060237 MDPI

17. Qin, *et al.* (replicate with exact bibliographic details once finalized) *Information Processing & Management*, 2024. DOI as above. ScienceDirect

18. (Main paper) Koren, A., Underwood, C., *et al.* “Evaluating the robustness of fake news detectors to adversarial attacks with real user comments.” *International Journal of Data Science and Analytics* (Springer), 2025. https://doi.org/10.1007/s41060-025-00790-3.