Fake News Detection Using PEFT-LoRA in RoBERTa with Retrieval-Augmented Generation

Tushar Raj
Under the Supervision of **Prof. Dr. S. Sumitra**
M.Tech Research Project Report

Abstract

This project addresses the challenge of detecting fake news in the digital age using advanced deep learning techniques. A RoBERTa-based transformer model is fine-tuned using Low-Rank Adaptation (LoRA), a Parameter-Efficient Fine-Tuning (PEFT) method, to classify news articles. Furthermore, Retrieval-Augmented Generation (RAG) is integrated for real-time fact-checking using Wikipedia as an external knowledge source. Despite high-quality retrievals, the generative model often fails to provide accurate predictions. This report delves into the theoretical underpinnings, implementation details, and performance evaluation of the approach.

1. Introduction

The spread of fake news threatens public opinion and decision-making. This research leverages transformer-based language models and retrieval systems to build a scalable, real-time fake news detection framework. The focus lies on three innovations:

- Using RoBERTa for robust textual understanding.
- Employing LoRA for efficient model adaptation.
- Integrating RAG for dynamic fact-checking.

2. RoBERTa Architecture

RoBERTa builds on BERT by optimizing the training strategy. It removes Next Sentence Prediction (NSP), increases batch size and training data, and uses dynamic masking.

The transformer encoder processes tokens $x_1, x_2, \ldots, x_n$, mapped to embeddings $\mathbf{E}(x_i)$. Attention is computed as:

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

where $Q = XW^Q$, $K = XW^K$, $V = XW^V$.

RoBERTa uses multi-head attention:

$$\text{MHA}(X) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{2}$$

Each head is computed separately.

3. Parameter-Efficient Fine-Tuning with LoRA

LoRA introduces low-rank matrices to attention layers, avoiding full fine-tuning.

3.1. Formulation

For a weight matrix $W_0 \in \mathbb{R}^{d\times k}$, LoRA replaces it with:

$$W = W_0 + \Delta W = W_0 + BA \tag{3}$$

where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$, and $r \ll \min(d, k)$.

3.2. Training Strategy

- Freeze pre-trained weights.
- Train only A and B.
- During inference, use $W = W_0 + BA$.

4. LoRA Training Algorithm

Algorithm 1 LoRA Fine-Tuning Algorithm

1: **Input:** Pre-trained model weights W_0, rank r
2: **Initialize:** $A \in \mathbb{R}^{r\times k}, B \in \mathbb{R}^{d\times r}$ randomly
3: **for** each batch (X, y) **do**
4: Compute logits with $W = W_0 + BA$
5: Compute loss: $\mathcal{L} = \text{CrossEntropy}(\text{softmax}(WX), y)$
6: Backpropagate and update A, B
7: **end for**

5. Real-Time Fact Verification

We integrated Wikipedia API to retrieve real-time knowledge for input claims. A snippet is appended to the input before classification.

6. Retrieval-Augmented Generation (RAG)

RAG combines a retriever with a generator:

- Retriever retrieves k documents $D_1, \ldots, D_k$ relevant to input x.
- Generator models $P(y|x, D_1, \ldots, D_k)$.

$$P(y|x) = \sum_{i=1}^{k} P(D_i|x)P(y|x, D_i) \tag{4}$$

Despite retrieving relevant Wikipedia data, hallucinations in generation led to wrong predictions.

7. Evaluation and Results

To evaluate the performance of the fake news detection model, we use standard classification metrics: accuracy, F1 score, precision, recall, and confusion matrix. These metrics are essential for understanding the strengths and weaknesses of the model, especially in scenarios involving class imbalance or noisy data.

7.1. Training and Validation Loss

The training process minimizes a loss function, typically the cross-entropy loss for classification tasks. Given predicted class probabilities $\hat{y}_i$ and true labels $y_i \in \{0, 1\}$, the binary cross-entropy loss for a batch of N samples is:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \tag{5}$$

A well-trained model shows a decreasing training loss with epochs and ideally a converging validation loss.

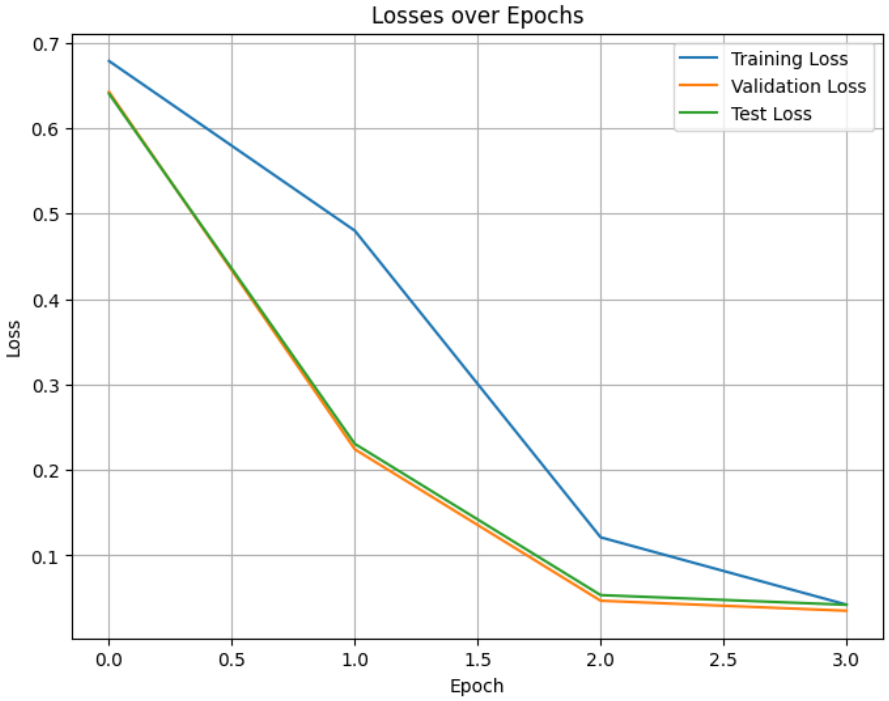

Figure 1: Training and Validation Loss over Epochs

The plot demonstrates smooth convergence, indicating stable optimization. However, divergence between training and validation losses may indicate overfitting.

7.2. F1 Score Over Epochs

The F1 Score is the harmonic mean of precision and recall:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

It is particularly useful for evaluating models on imbalanced datasets, as it balances false positives and false negatives.

Let:

$$\text{TP} = \text{True Positives}$$
$$\text{FP} = \text{False Positives}$$
$$\text{FN} = \text{False Negatives}$$

Then:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{8}$$

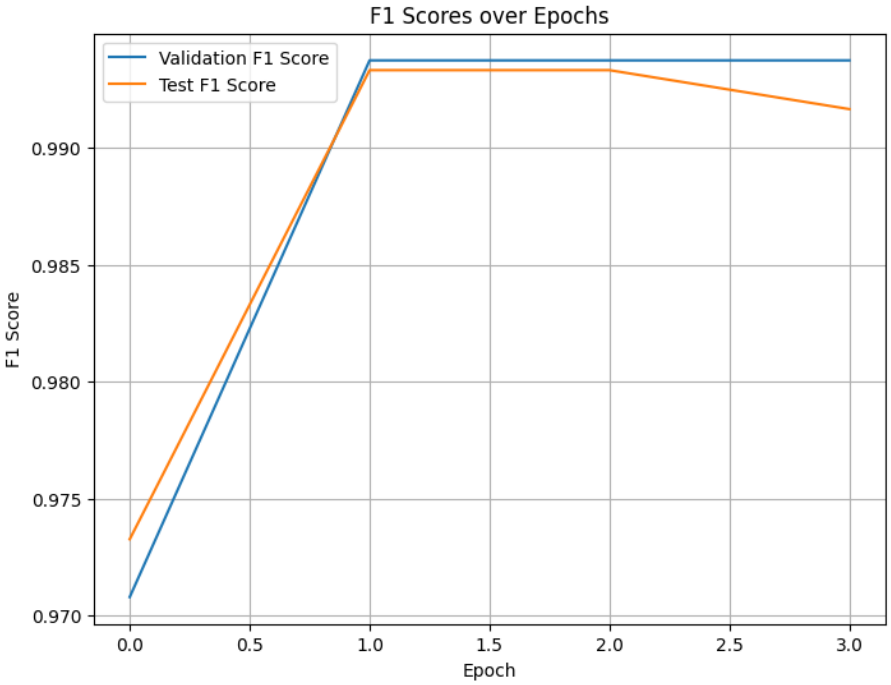

Figure 2: F1 Score over Epochs

The plot shows an improvement in F1 score, signifying better balance between precision and recall over epochs.

7.3. Accuracy Over Epochs

Accuracy is the ratio of correctly predicted samples to total samples:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

It is easy to interpret but may not be reliable for imbalanced classes. In this task, fake and real news distributions may not be uniform, so accuracy alone is insufficient.

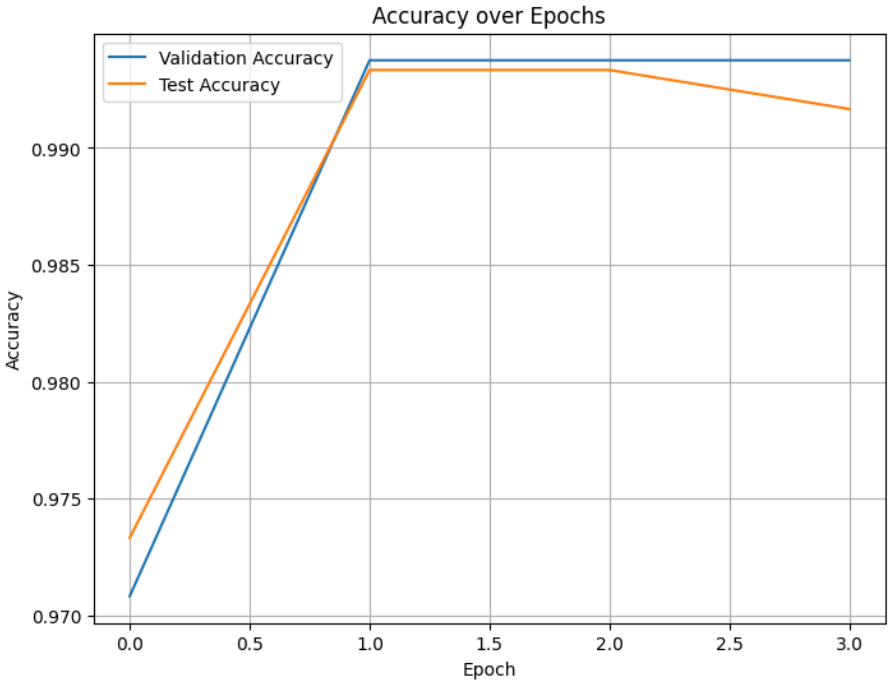

Figure 3: Accuracy over Epochs

The accuracy curve indicates that the model steadily learns to distinguish between classes, with improvements tapering off as it converges.

7.4. Confusion Matrix (Final Epoch)

The confusion matrix presents the raw classification outcomes:

	Predicted Fake	Predicted Real
Actual Fake	TP	FN
Actual Real	FP	TN

Confusion Matrix Layout

Where:

- TP = Correctly predicted fake news
- TN = Correctly predicted real news
- FP = Real news wrongly predicted as fake
- FN = Fake news wrongly predicted as real

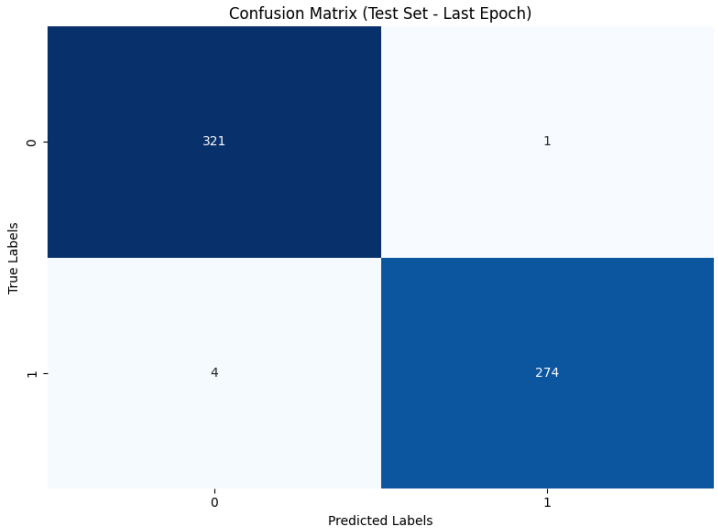

Figure 4: Confusion Matrix (Final Epoch)

From the confusion matrix, we can derive all classification metrics and analyze the types of errors the model makes. High FN (false negatives) are particularly problematic in fake news detection, as it means fake articles are being misclassified as real.

7.5. Interpretation of Results

- **Model Learning:** The consistent decrease in loss and increase in F1 score demonstrate successful model learning.
- **LoRA Effectiveness:** The model trains efficiently using fewer parameters, as expected from the PEFT LoRA approach.
- **Impact of RAG:** While the retriever retrieves relevant Wikipedia data, the generative response often deviates from factual content, indicating hallucination—this results in wrong predictions despite correct evidence retrieval.

Conclusion of Evaluation

This evaluation shows that LoRA-based RoBERTa achieves strong baseline performance in fake news detection. The visualized metrics confirm effective learning and generalization. However, the current RAG pipeline introduces instability in prediction due to hallucinated text generation, warranting further refinement.

Acknowledgments

I thank Prof. Dr. S. Sumitra for her mentorship and guidance.

References

[1] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[2] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chaudhary, "LoRA: Low-Rank Adaptation of Large Language Models," *arXiv preprint arXiv:2106.09685*, 2021.

[3] H. Liu, J. Mu, T. Scao, A. Fan, A. Tamkin, E. Perez, M. K. Muzio, P. Xu, M. Shoeybi, and B. Zoph, "Parameter-efficient fine-tuning of large language models: A survey," *arXiv preprint arXiv:2303.15647*, 2023.

[4] P. Lewis, E. Pérez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kandpal, B. Yogatama, S. Riedel, and A. Kandpal, "Retrieval-augmented generation for knowledge-intensive NLP tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[5] S. Sammut and G. Webb, "Confusion matrix," in *Encyclopedia of Machine Learning*, pp. 160–161, Springer, 2010.

[6] D. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.