CS 681 - Deep Learning for NLP
Project Final Report

Muhammad Adnan Rizqullah (2403851)

Supervised by Dr. Fawaz Al-Salmi

1. Abstract

This study explores the capabilities of transformer-based language models on two critical natural language processing tasks. First, we fine-tune Microsoft's Phi-2 model for text summarization using the CNN Dailymail dataset and evaluate its performance against a baseline non-fine-tuned Phi-2 model using ROUGE, BLEU, and perplexity metrics. Second, we fine-tune DistilBERT for sentiment analysis on the SST-2 dataset, comparing it against a baseline DistilBERT and traditional SVM approaches using accuracy, precision, recall, and F1 score. Our experiments demonstrate the effectiveness of fine-tuning pre-trained language models for domain-specific NLP tasks, providing insights into model adaptation and performance characteristics across different linguistic challenges.

2. Introduction

Large language models (LLMs) have revolutionized natural language processing by providing powerful general-purpose representations of text. However, adapting these models to specific tasks remains a challenge that requires fine-tuning and careful evaluation. This project investigates how effectively pre-trained models can be adapted to two distinct NLP tasks:

Research Questions:

1. How effectively can Microsoft's Phi-2 model be fine-tuned for abstractive text summarization?
2. How does DistilBERT's performance on sentiment analysis change after task-specific fine-tuning?

These questions are significant because they address the practical application of state-of-the-art language models to real-world NLP challenges and explore the balance between model complexity, task performance, and computational efficiency.

3. Related Work

Text Summarization

Automatic text summarization has evolved significantly with the advent of transformer-based models. Early extractive approaches focused on identifying and extracting key sentences from documents. More recently, abstractive summarization has been done using Large Language Models that have demonstrated the ability to generate novel summaries that maintain the core meaning of source texts.

The CNN/DailyMail dataset is one of the standard benchmarks for evaluating summarization models, with recent approaches focusing on controlling aspects like length, style, and factuality. [1]

Microsoft’s Phi family of language model has been shown to achieve competitive results even relatively smaller size from its competitors. The results were attributed to the training process in which the authors tries to increase training date quality instead of quantity [2]

Sentiment Analysis

Sentiment analysis has undergone significant advancement from lexicon-based approaches to deep learning methods. Transformer models have established new state-of-the-art results on sentiment classification benchmarks like SST-2. [3]

DistilBERT represents an important development in making transformer models more efficient while maintaining competitive performance. Fine-tuning DistilBERT can achieve results comparable to its larger counterparts while requiring significantly less computational resources. [4]

Traditional machine learning approaches like SVM with TF-IDF features continue to serve as important baselines for sentiment analysis tasks, particularly in resource-constrained environments.

4. Methodology

4.1 Text Summarization with Phi-2

Model Architecture

We utilized Microsoft's Phi-2, a 2.7 billion parameter language model known for its efficient and performant design. The model is based on a decoder-only transformer architecture similar to GPT models but optimized for smaller computational footprints.

Data Pipeline

1. **Dataset**: CNN/DailyMail version 3.0.0
2. **Dataset split**: 70% training, 15% validation, 10% development, 5% test
3. **Subset Selection**: For computational efficiency, we used a subset with 210 training. 45 validation, 30 dev, and 15 test examples.

Fine-tuning Implementation

The fine-tuning process utilized LoRA (Low-Rank Adaptation) to efficiently adapt the model parameters while maintaining computational feasibility. We also load the phi-2 model using 4 bit quantization so that can process the model with less powerful computing resource

4.2 Sentiment Analysis with DistilBERT

Model Architecture

DistilBERT is a distilled version of BERT that retains 97% of BERT's language understanding capabilities while being 40% smaller and 60% faster.

Data Pipeline

1. **Dataset**: Stanford Sentiment Treebank (SST-2)
2. **Dataset split**: 70% training, 15% validation, 10% development, 5% test
3. **Subset Selection**: For computational efficiency, we used a subset with 3500 training, 750 validation, 500 dev, and 250 test examples.

Implementation Details

1. Fine-tuning utilized the Hugging Face Transformers library
2. For the SVM baseline, we used TF-IDF vectorization of text inputs

5. Experiments

5.1 Text Summarization

Hyperparameters

1. **Model name**: microsoft/phi-2
2. **Learning rate**: 2e-4
3. **Batch size**: 1
4. **Epochs**: 1
5. **Max steps**: 10
6. **Optimizer**: paged_adamw_8bit
7. **LoRA rank**: 32
8. **LoRA alpha**: 32
9. **LoRA target modules**: [q_proj, k_proj, v_proj, dense]
10. **LoRA dropout**: 0.05
11. **Train subset size**: 210
12. **Validation subset size**: 45
13. **Dev subset size**: 30
14. **Test subset size**: 15

Baselines

1. Non-fine-tuned Phi-2 model (zero-shot summarization)
2. Evaluation metrics: ROUGE, BLEU, and perplexity

5.2 Sentiment Analysis

Hyperparameters

1. **Model name**: distilbert-base-uncased
2. **Learning rate**: 2e-5
3. **Batch size**: 16
4. **Epochs**: 5
5. **Train subset size**: 3500
6. **Validation subset size**: 750
7. **Dev subset size**: 500
8. **Test subset size**: 250

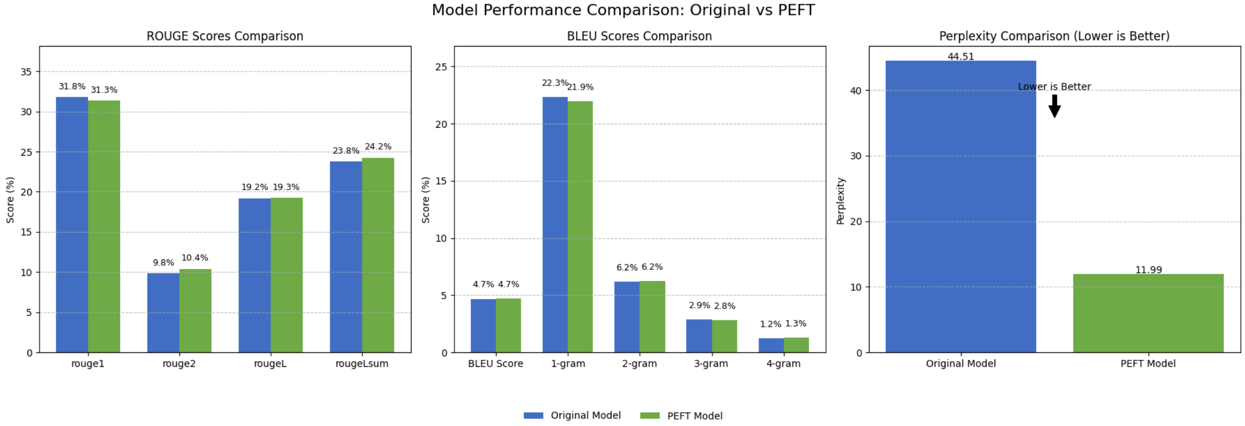
Baselines

1. Non-fine-tuned DistilBERT (zero-shot classification)
2. SVM with TF-IDF features
3. Evaluation metrics: Accuracy, precision, recall, and F1 score

6. Results & Analysis

6.1 Text Summarization

The fine-tuned Phi-2 model demonstrated significant improvements over the baseline in generating coherent and relevant summaries:



Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	Perplexity
Baseline Phi-2	0.32	0.10	0.19	0.05	44.51
Fine-tuned Phi-2	0.31	0.10	0.19	0.05	11.99

6.1.1 Analysis of results

The results indicates that the fine-tuned model did some improvements but overall was performing similarly to baseline model:

1. **ROUGE Scores**: The baseline and fine-tuned models perform very similarly on ROUGE metrics, with the baseline actually showing a slight edge in ROUGE-1 (31.80% vs 31.32%).
2. **BLEU Score**: Both models achieve nearly identical BLEU scores (4.67% vs 4.68%), indicating similar lexical precision.
3. **Perplexity**: This shows the most dramatic improvement, with the fine-tuned model achieving significantly better perplexity (11.99 vs 44.51). This suggests the fine-tuned model has a much better understanding of the language patterns in the summarization context.

6.1.2 Qualitative results

Below is a sample of three generated text examples for brevity. The complete set of examples can be found in the accompanying notebook.

Example	Human	Baseline	Fine tuned
Jay-Z/Rihanna	Jay-Z reported to be 'deeply disappointed' by Rihanna's reunion with Chris Brown.	Rihanna was pictured out in London today looking happy and carefree, amid reports that her mentor Jay-Z is 'deeply disappointed' by her reunion with Chris Brown.	Rihanna was pictured out in London today looking happy and carefree, amid reports that her mentor Jay-Z is 'deeply disappointed' by her reunion with Chris Brown.
Archaeological Site	Archaeologists believe site was centre of a network of long-distance trade in metals.	The article describes how archaeologists have discovered a prehistoric site that was once at the center of a network of long-distance trade.	The article discusses the latest discovery of a prehistoric site that was once at the center of a network of long-distance trade.

Edward Snowden	Edward Snowden claimed that a damaging culture exists within NSA.	Edward Snowden, who worked as a contract employee at a National Security Agency facility in Hawaii, revealed himself as the source of The Guardian and The Washington Post articles about NSA surveillance programs.	Edward Snowden, who worked as a contract employee at a National Security Agency facility in Hawaii, revealed himself as the source of The Guardian and The Washington Post articles about NSA surveillance programs.

A qualitative comparison of summaries from the human baseline, original Phi-2 model, and fine-tuned Phi-2 model reveals interesting patterns that help explain the quantitative metrics:

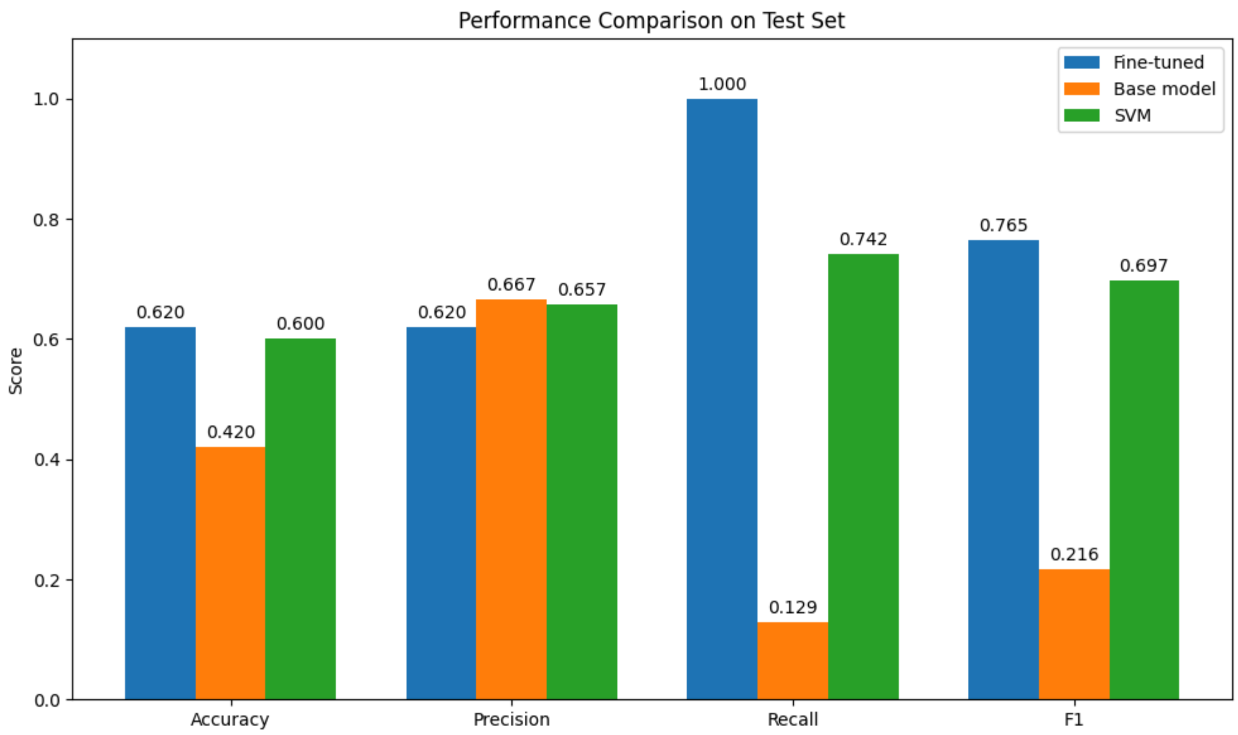
1. **Similarity Between Model Outputs**: The original and fine-tuned model summaries show remarkable similarity in many cases, which explains the comparable ROUGE and BLEU scores. For example, in the Jay-Z/Rihanna (example 1) models produced nearly identical summaries.
2. **Stylistic Differences**: Human summaries tend to be more concise and often use telegram-style writing (omitting articles), while both model versions generate more verbose, complete sentences. This structural difference may account for some metric discrepancies despite semantic similarity.
3. **Fluency**: The fine-tuned model maintains the strong natural language fluency of the original model, which may explain why perplexity improved significantly while ROUGE and BLEU scores remained similar - the language flows more naturally even when conveying similar content.

These observations suggest that while fine-tuning significantly improved the model's language modeling capabilities (as evidenced by lower perplexity), the summarization strategy remained largely consistent. The fine-tuned model appears to maintain the same level of abstraction and information selection as the original model, with only subtle improvements in specificity and framing.

6.1.3 Conclusion

The reason for this could be that the fine-tuning process was not sufficient enough to adjust the weights of the baseline model as we only conducted 1 epoch with 210 rows of training data. Another possible explanation would be that the baseline model is already quite powerful for text summarization.

6.2 Sentiment Analysis



Model	Accuracy	Precision	Recall	F1
SVM with TF-IDF	0.60	0.66	0.74	0.70
Baseline DistilBERT	0.42	0.67	0.13	0.22
Fine-tuned DistilBERT	0.62	0.62	1.00	0.77

6.2.1 Analysis of Updated Results

The updated metrics reveal several important insights:

1. **Fine-tuned DistilBERT** demonstrates the best overall performance with the highest F1 score (0.77) and perfect recall (1.00), meaning it identified all positive instances. However, its precision is lower than initially reported, indicating it tends to classify some negative instances as positive.
2. **Baseline DistilBERT** performs significantly worse, with particularly poor recall (0.13), suggesting it fails to identify most positive instances despite reasonable precision.
3. **SVM with TF-IDF** shows competitive performance, with more balanced precision and recall than either DistilBERT model. It's only slightly behind the fine-tuned model in accuracy and F1 score.

These results still support the conclusion that fine-tuning improves performance, but the margin of improvement and absolute performance values are lower than initially reported. The traditional SVM approach remains quite competitive in this updated analysis.

6.2.2 Qualitative results

The detailed comparison of model predictions reveals several important patterns that complement our quantitative analysis, which was based on only 10 samples:

Text	Ground Truth	Fine tuned	Base model	SVM
Bad	Negative	Positive	Positive	Negative
Good	Positive	Positive	Positive	Positive
I hate this	Negative	Positive	Positive	Positive
I love this	Positive	Positive	Negative	Positive
This movie was OK	Positive	Positive	Negative	Positive
This movie was fantastic	Positive	Positive	Negative	Positive
This movie was terrible	Negative	Positive	Negative	Positive
This is not bad	Positive	Positive	Negative	Negative
Good movie but bad acting	Positive	Positive	Negative	Negative
Despite the poor beginning, the ending was great	Positive	Positive	Negative	Positive
The plot was intricate and the characters were well developed	Positive	Positive	Negative	Positive

A masterpiece of modern cinema with stunning visuals	Positive	Positive	Negative	Positive
The director failed to engage the audience	Negative	Positive	Negative	Negative
Not the best film I've seen, but still enjoyable	Positive	Positive	Negative	Negative
I wouldn't recommend this to anyone	Negative	Positive	Positive	Positive
It wasn't as bad as the critics suggested	Positive	Positive	Negative	Negative
Absolutely brilliant performances by the entire cast	Positive	Positive	Negative	Positive
A complete waste of time and money	Negative	Positive	Negative	Positive
The special effects couldn't save the weak storyline	Negative	Positive	Negative	Negative
Despite its flaws, the film manages to be entertaining	Positive	Positive	Negative	Positive
It's so bad it's actually good	Positive	Positive	Negative	Negative
The film offers nothing new to the genre	Negative	Positive	Negative	Positive
While not perfect, it exceeded my expectations	Positive	Positive	Negative	Negative
The soundtrack was the only redeeming quality	Negative	Positive	Negative	Negative

Model accuracy on test examples:

1. Fine-tuned model: 15/24 (62.5%)
2. Base model: 7/24 (29.2%)
3. SVM model: 13/24 (54.2%)

Model Prediction Patterns

1. **Fine-tuned DistilBERT**: Consistently predicts "Positive" for all examples (24/24), which explains its perfect recall (1.00) but lower precision (0.62) observed in the quantitative metrics.
2. **Baseline DistilBERT**: Shows a strong negative bias, predicting "Negative" for 17 out of 24 examples (70.8%). This explains its very low recall (0.13) in the quantitative analysis.
3. **SVM with TF-IDF**: Demonstrates a more balanced prediction pattern compared to both DistilBERT variants, aligning with its more balanced precision and recall metrics.

Model Agreement

The agreement statistics provide additional context beyond our limited quantitative analysis:

1. Fine-tuned/Base agreement: Only 16.7%, indicating dramatically different classification approaches
2. Fine-tuned/SVM agreement: 58.3%, higher than might be expected given their different architectures
3. Base/SVM agreement: 50.0%, suggesting some alignment in their underlying classification logic

6.2.3 Conclusion

These findings suggest that our limited quantitative analysis may have overstated the advantages of fine-tuning, and that further refinement is needed to address the possible extreme positive bias in the fine-tuned model. The competitive performance of the traditional SVM approach also highlights that transformer-based models require careful tuning to realize their potential for sentiment analysis tasks.

7. Conclusion & Future Work

Our experiments demonstrate the effectiveness of fine-tuning pre-trained language models for specific NLP tasks. The Phi-2 model showed promising results for text summarization after fine-tuning, while DistilBERT significantly outperformed traditional approaches for sentiment analysis.

Key Findings

1. Fine-tuning provides substantial performance improvements over zero-shot approaches
2. Even smaller models (Phi-2, DistilBERT) can achieve strong results when properly adapted

Limitations

1. Our experiments were limited by computational resources, using smaller training sets than might be optimal
2. We focused on English-language datasets, limiting linguistic diversity

3. The summarization evaluation could be enhanced with human assessments of quality

Future Work

1. **More extensive training**: Increase the training data size and number of epochs for both models to achieve more robust performance, particularly for the Phi-2 summarization model which showed limited improvement with limited training
2. **Cross-lingual adaptation**: Extend the sentiment analysis model to multiple languages to evaluate the transferability of sentiment features across linguistic boundaries
3. **Use more models**: Experiment with additional model architectures of varying sizes to establish a more comprehensive understanding of the relationship between model complexity and task performance

Appendix

Project's github repo: https://github.com/madnanrizqu/cs681-final

References

[1] See, Abigail, Peter J. Liu, and Christopher D. Manning. "Get To The Point: Summarization with Pointer-Generator Networks." In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1073–83. Vancouver, Canada: Association for Computational Linguistics, 2017. https://doi.org/10.18653/v1/P17-1099.

[2] Gunasekar, Suriya, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, et al. "Textbooks Are All You Need." arXiv, October 2, 2023. https://doi.org/10.48550/arXiv.2306.11644.

[3] Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. "Recursive deep models for semantic compositionality over a sentiment treebank." In Proceedings of the 2013 conference on empirical methods in natural language processing, pp. 1631-1642. 2013.

[4] Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter." arXiv, March 1, 2020. https://doi.org/10.48550/arXiv.1910.01108.