LLMEMBED REPRODUCTION AND IMPROVEMENT FOR TEXT CLASSIFICATION USING LIGHTWEIGHT LANGUAGE MODELS

Hassaan Ullah, Ali Tanveer, Ahsan Saqib

FAST National University of Computer and Emerging Sciences, Islamabad

ABSTRACT

This project reproduces and extends the ACL 2024 paper "LLMEmbed: Rethinking Lightweight LLM's Genuine Function in Text Classification." The original paper demonstrates that lightweight Large Language Models can be used as efficient embedding extractors without parameter-intensive fine-tuning. We re-implemented the methodology using the Phi-2 model, extracted multi-layer embeddings, and evaluated them on the AG News dataset using logistic regression. The reproduced baseline achieved 87.80% accuracy and 87.25% Macro-F1 on a reduced dataset due to hardware limitations. We introduced original improvements including contrastive embedding post-training, lightweight data augmentation, and advanced fusion ensembles. Although improvements did not surpass the baseline primarily because of CPU-only constraints, they demonstrate extensibility and robustness of LLM embedding pipelines. This project confirms that lightweight LLMs can perform competitively in text classification tasks.

Index Terms— LLMs, Embeddings, NLP, Text Classification, Contrastive Learning

1. INTRODUCTION

Large Language Models have become foundational in natural language processing. Most classification pipelines rely on fine-tuning models like BERT, but this requires significant computational resources. The LLMEmbed paper proposes that lightweight LLMs can serve as strong fixed embedding extractors without any fine-tuning. The key insight is that multi-layer hidden states contain rich semantic information that can be fused into a single embedding.

This project aims to reproduce the core LLMEmbed methodology, evaluate its performance on a benchmark dataset, and improve the pipeline with additional lightweight techniques. Hardware limitations influenced our experimental scale but still allowed meaningful results.

2. RELATED WORK

Early non-contextual methods such as TF-IDF and Bag of Words provided baseline approaches for text classification. BERT and Transformers introduced contextualized embeddings with fine-tuning capabilities [3]. Sentence-BERT and SimCSE used contrastive learning to generate sentence-level embeddings [2]. The LLMEmbed paper presented at ACL 2024 showed that lightweight LLMs can outperform supervised and fine-tuned models using only embeddings with fusion techniques [1]. Our improvements draw inspiration from SimCSE and ensemble learning methodologies while representing original contributions to the embedding extraction pipeline.

3. METHODOLOGY OVERVIEW

The project consists of two main components. Part 2 focuses on reproduction using the Phi-2 model with 2.7 billion parameters on the AG News dataset. We extract hidden states from multiple layers, apply mean fusion of embeddings, and train logistic regression on the fused representations. Part 3 introduces improvements including contrastive embedding refinement inspired by SimCSE, data augmentation techniques such as synonym swapping and random deletion, and advanced fusion with ensemble classifiers.

4. DATASET DESCRIPTION

We utilize the AG News dataset containing four classes: World, Sports, Business, and Science/Technology. The original dataset includes 120,000 training samples and 7,600 test samples. Due to CPU-only computational constraints, we reduced the dataset to 2,000 training samples and 500 test samples while maintaining class balance across all categories.

5. EMBEDDING EXTRACTION PIPELINE

The embedding extraction follows the LLMEmbed methodology precisely. We load the Phi-2 model and tokenize input texts. During the forward pass, we extract hidden states from multiple layers of the transformer architecture. Each layer's token embeddings are mean-pooled to create sentence-level representations. Finally, we fuse multi-layer embeddings through averaging to produce a single comprehensive vector per sample that captures information from different levels of the network hierarchy.

6. FUSION MECHANISMS

6.1. Baseline Fusion

The baseline reproduction employs mean fusion across selected layers where the final embedding is computed as the average of hidden states from multiple layers. This approach balances information from different network depths.

6.2. Advanced Fusion

Our improvements explore concatenation fusion that combines layer outputs directly, weighted scalar fusion that learns optimal layer contributions, and ensemble fusion using both probability averaging and majority voting strategies to combine predictions from multiple fusion configurations.

7. CLASSIFIER

A logistic regression classifier is employed for both reproduction and improvement experiments. This choice offers several advantages including computational efficiency, lightweight memory footprint, consistency with the original paper's methodology, and strong performance with fixed embedding representations. The simplicity of logistic regression also allows us to focus evaluation on embedding quality rather than classifier complexity.

8. EXPERIMENTAL SETUP

8.1. Hardware

All experiments were conducted on an Intel i5 CPU with 12 GB RAM and no GPU acceleration. Phi-2 inference on CPU required approximately three hours total processing time. These constraints significantly influenced our experimental design and dataset size decisions.

8.2. Software

The implementation utilizes Python with HuggingFace Transformers library for model loading and inference, PyTorch for tensor operations, and scikit-learn for classification and evaluation metrics.

9. REPRODUCTION RESULTS

The baseline reproduction achieved an accuracy of 0.8780 and Macro-F1 score of 0.8725 on the reduced test set. The confusion matrix demonstrates strong diagonal values indicating successful classification across all four categories. Class-wise performance shows particularly strong results for the Sports category with 135 correct predictions out of 145 samples. These results successfully reproduce the LLMEmbed methodology and demonstrate strong performance despite the reduced dataset size imposed by hardware limitations.

10. IMPROVEMENT RESULTS

10.1. Contrastive Embedding Post-Training

We implemented in-batch contrastive loss inspired by SimCSE to refine embeddings through self-supervised learning. This approach achieved an accuracy of 0.8380 and Macro-F1 of 0.8329. The performance decrease relative to baseline is attributed to the combination of small dataset size and CPU limitations causing underfitting during the contrastive training phase.

10.2. Data Augmentation

We applied data augmentation techniques randomly to 20% of training samples using synonym replacement, random word swapping, and random deletion. These techniques increase training data diversity and provide robustness against lexical variations without requiring additional labeled data.

10.3. Advanced Fusion and Ensemble

Three fusion configurations were combined through ensemble methods. Both soft probability averaging and majority voting ensembles maintained the baseline performance with accuracy of 0.8780 and Macro-F1 of 0.8725. This stability confirms the robustness of the embedding representations across different fusion strategies.

11. COMPARATIVE ANALYSIS

The following table summarizes results across all experimental conditions:

Method	**Accuracy**	**Macro-F1**
Baseline (Reproduction)	0.8780	0.8725

Method	Accuracy	Macro-F1
Contrastive Learning	0.8380	0.8329
Ensemble (Probability)	0.8780	0.8725
Ensemble (Voting)	0.8780	0.8725

The baseline reproduction establishes a strong foundation. While contrastive learning showed reduced performance under current constraints, ensemble methods maintained stability and confirmed the reliability of the embedding pipeline.

12. DISCUSSION

Several key observations emerge from our experiments. The reproduction successfully matched expectations from the LLMEmbed methodology, validating the approach on reduced data. Baseline performance remained strong despite significant hardware constraints. The improvement techniques were conceptually meaningful but limited by small dataset size, CPU-only processing, and few contrastive training epochs. However, our experiments confirmed the stability and robustness of lightweight LLM embeddings across different processing configurations. Future work with GPU resources and larger datasets would likely demonstrate the full potential of the proposed improvements.

13. CONCLUSION

This project successfully reproduced the LLMEmbed methodology and implemented several novel improvements for text classification. While enhancements did not exceed baseline performance under current hardware constraints, they demonstrate how LLM-based embedding pipelines can be extended and studied in resource-limited environments. The methodology proves reproducible, flexible, and suitable for future expansion. With GPU acceleration and larger datasets, the proposed improvements show promise for advancing lightweight LLM embedding techniques in text classification tasks.

14. REFERENCES

[1] C. Liu et al., "LLMEmbed: Rethinking Lightweight LLM's Genuine Function in Text Classification," in Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 2024. https://aclanthology.org/2024.findings-acl.87/.

[2] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple Contrastive Learning of Sentence Embeddings," in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021.https://aclanthology.org/2021.emnlp-main.823/

[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, 2018.https://aclanthology.org/N19-1423/.

[4] AG News Dataset. Available: https://huggingface.co/datasets/ag_news

[5] Phi-2 Model (HuggingFace)
Microsoft Phi-2: lightweight LLM used for embedding extraction.https://huggingface.co/microsoft/phi-2