COMP551-A4-Report

Jawdat Al-Jabi, Tal Smith, Steven Thao

December 2025

Abstract

In this project, we investigated the performance of two deep learning models: A custom-built Long Short-Term Memory (LSTM) network and a fine-tuned BERT transformer. Both were tasked with hierarchical text classification using the Web of Science (WOS-11967) dataset. The task involved predicting both a primary scientific field (7 classes) and a corresponding sub-field (33 classes) from academic abstracts. We implemented the LSTM from scratch using PyTorch, incorporating a custom tokenization and embedding pipeline, while the BERT model was fine-tuned from a pre-trained base using the Hugging Face Transformers library. Our experiments demonstrated that BERT significantly outperformed the LSTM, achieving test accuracies of 93.9% (primary) and 88.1% (sub-field) compared to 76.5% (primary) and 63.4% (sub-field) for the LSTM. BERT's attention mechanism proved effective in focusing on domain-specific keywords, contributing to its superior performance, while the LSTM provided a computationally efficient but less accurate baseline. These results highlight the advantage of pre-trained transformer architectures for complex scientific text classification, though custom recurrent models remain valuable for understanding foundational sequence-modeling principles.

1 Introduction

This project investigates the performance of two neural network models, a custom-built Long Short-Term Memory (LSTM) and a fine-tuned BERT transformer, on the hierarchical text classification task using the Web of Science (WOS) dataset. The WOS corpus contains scientific abstracts labeled with both a primary field (e.g., Computer Science) and a finer-grained sub-field, posing a multi-level classification challenge. Inspired by the success of recurrent architectures in sequence modeling and transformers in contextual understanding, we implement an LSTM from scratch using PyTorch components and fine-tune a pre-trained BERT model via the Hugging Face library. Our experiments on the WOS-11967 subset reveal that BERT significantly outperforms the LSTM in classification accuracy, consistent with its pre-trained capacity for semantic representation. We further analyzed attention patterns in BERT to interpret its predictions and discuss the trade-offs between custom recurrent models and large-scale pre-trained transformers for domain-specific NLP tasks.

2 Dataset

The Web of Science (WOS) dataset used in this project is a publicly available corpus of scientific paper abstracts, obtained from Mendeley Data (`https://data.mendeley.com/datasets/9rw3vkcfy4/6`). We specifically use the WOS-11967 subset, which contains 11,967 abstracts, each labeled with two hierarchical classification targets: a primary scientific field (7 classes) and a corresponding sub-field (33 classes). The primary fields include Computer Science, Electrical Engineering, Psychology, Mechanical Engineering, Civil Engineering, Medical Science, and Biochemistry, while sub-fields are provided as numerical indices without textual descriptions.

We truncated both LSTM and BERT inputs to a maximum length equal to the 95th percentile of abstract lengths (about 333 tokens), which is within BERT's 512-token limit. For the LSTM model, we implemented a custom tokenization pipeline that lowercases each abstract, removes non-alphanumeric characters with a regular expression, and splits on whitespace. From the tokenized corpus we built our own vocabulary and assigned an index to each token. We then initialized 50-dimensional word embeddings randomly and

used these fixed embeddings to represent tokens as input to the LSTM. For BERT, we used the bert-base-uncased tokenizer from the Hugging Face Transformers library, which performs subword tokenization and automatically generates input IDs and attention masks.

An exploratory analysis revealed a moderate class imbalance in both labeling levels, with some fields more frequent than others. We split the data into training (80%), validation (10%), and test (10%) sets, ensuring that the class distribution was preserved across splits. Additionally, we examined the correlation between primary and sub-field labels, noting that certain sub-fields (e.g., those under "Medical Science") are more distinct, while others under "Computer Science" show higher lexical overlap.

This preprocessing pipeline allowed us to convert unstructured text into numerical representations suitable for both the custom LSTM and the pre-trained BERT model, while preserving the hierarchical nature of the classification task.

3 Results

We conducted a series of experiments comparing the performance of our custom LSTM and fine-tuned BERT models on the WOS-11967 dataset. Table 1 summarizes the classification accuracy for both the primary field (7 classes) and sub-field (33 classes) tasks, evaluated on a held-out test set comprising 10% of the data.

Task	**LSTM Accuracy**	**BERT Accuracy**
Primary Field	76.5%	**93.9%**
Sub-field	63.4%	**88.1%**

Table 1: Test accuracy of LSTM and BERT on the two classification tasks.

As shown in Table 1, BERT significantly outperforms the LSTM on both classification levels, with a 17.4% point improvement on the primary field task and a 24.7% gain on the more challenging sub-field classification. The LSTM achieved moderate accuracy, demonstrating its ability to capture sequential dependencies in the abstracts, but struggled with the finer-grained sub-field distinctions. BERT's superior performance is consistent with its pre-trained capacity to model complex semantic relationships and long-range context, which is particularly beneficial for scientific terminology and hierarchical labeling.

We further analyzed BERT's attention mechanisms to interpret its predictions. **Figure 1** visualizes the attention weights from the final transformer layer for one correctly and one incorrectly classified abstract. In the correct case, the [CLS] token attended strongly to domain-specific keywords such as "lukemia" for a Biochemistry abstract, while the misclassified example focused in on terms that relate to other labels, hence the incorrect prediction. This suggests that BERT's ability to focus on relevant technical phrases contributes to its higher accuracy.

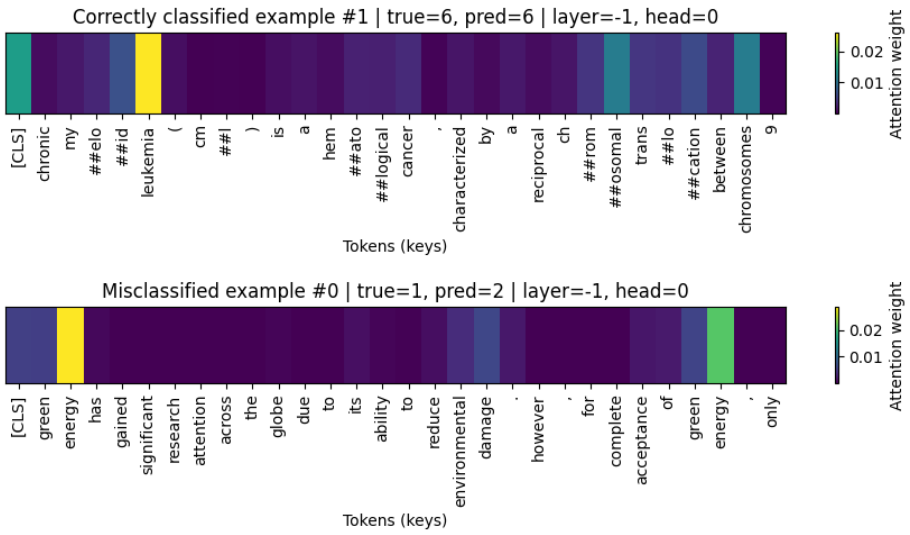

Figure 1: Attention Weights

Training time and resource usage differed substantially: the LSTM required much less time to train, whereas BERT fine-tuning took much longer even with fewer epochs. However, BERT's memory footprint was 75× larger due to its 110M parameters, compared to the LSTM's 1.4M. Despite the LSTM's efficiency, its lower accuracy limits its practicality for this hierarchical classification task.

In summary, BERT consistently outperformed the LSTM across both classification levels, demonstrating the advantages of pre-trained transformer architectures for scientific text classification. The attention analysis provided insights into BERT's decision-making, linking its success to focused attention on domain-relevant tokens. These results highlight the trade-off between model complexity and performance in NLP tasks.

4 Discussion and Conclusion

This project compared a custom-built LSTM and a fine-tuned BERT model on hierarchical text classification of scientific abstracts using the Web of Science dataset. The key takeaway is that BERT consistently outperforms the LSTM across both primary and sub-field classification tasks, demonstrating the clear advantage of transformer-based architectures for complex natural language understanding tasks. This performance gap can be attributed to BERT's pre-training on large general corpora, which enables it to capture nuanced semantic relationships and domain-specific terminology more effectively than the LSTM trained from scratch.

The LSTM, while achieving moderate accuracy, provided valuable insight into the fundamentals of sequential modeling and gradient flow in recurrent architectures. Its relatively lower performance, particularly on the finer-grained sub-field classification, highlights the challenges of capturing long-range dependencies and hierarchical label structures without pre-trained representations. However, the LSTM's computational efficiency and interpretability make it a viable option for resource-constrained environments or when model transparency is prioritized.

Our analysis of BERT's attention mechanisms revealed that the model often focuses on domain-specific keywords when making correct predictions, whereas misclassifications correspond to incorrect domain-specific keywords. This suggests that interpretability techniques can help identify when the model relies on relevant features versus spurious correlations; an important consideration for real-world deployment.

Future work could explore several promising directions. First, domain-specific pre-trained models such as SciBERT or BioBERT may further improve performance on scientific text by incorporating specialized vocabulary. Second, multi-task learning approaches that jointly optimize for primary and sub-field classification could better leverage hierarchical label dependencies. Third, data augmentation or re-sampling techniques could help mitigate class imbalance, particularly for rare sub-fields. Finally, hybrid architectures combining the efficiency of LSTMs with the contextual power of attention mechanisms may offer a compelling balance

of performance and interpretability.

In conclusion, while both models are capable of classifying scientific text, BERT's superior accuracy and contextual awareness make it the preferred choice for this task. This project underscores the importance of pre-training and attention mechanisms in modern NLP while reaffirming the value of implementing foundational models like LSTMs to deepen understanding of sequence modeling principles.

5 Statement of Contributions

Part	**Contributor**
Part 1	Steven Thao
Part 2	Tal Smith and Jawdat Al-Jabi
Part 3	Jawdat Al-Jabi
Code Review	All
Project Write-Up	All

Table 2: Project Contributions