Assignment 4: Experimenting BERT on AG news

COMP 551
April 22th, 2025

Androw Abd El Malek (261050237)
Kevin Luo (261038261)
Radu Petrescu (261051351)

Abstract

In this project, we experimented with the pre-trained BERT model on the AG News dataset—a collection of news headlines categorized into four classes: World, Sports, Business, and Sci/Tech. We probed the untrained BERT model using various embedding strategies, including the [CLS] token, the first token, the last token, and the mean of all token embeddings. These embeddings were evaluated using K-Nearest Neighbours (KNN) and multi-class logistic regression classifiers. Among the different strategies, the mean token embedding consistently yielded the highest validation accuracy for KNN and logistic regression, with KNN slightly outperforming logistic regression.

After the probing experiments, we fine-tuned the BertForSequenceClassification model, initialized with "google-bert/bert-base-uncased," on a subset of 10,000 training samples. The fine-tuned model achieved the highest test accuracy at 92.14%, followed by the probed KNN classifier at 91.66% and the probed logistic regression classifier at 91.06%.

Finally, we analyzed the attention weights from the last layer of the fine-tuned BERT model to better understand its interpretability. The model tended to assign higher attention to end-of-sequence punctuation tokens in the samples we examined.

Introduction

Recent advancements in natural language processing (NLP) have been driven by transformer-based models, with Bidirectional Encoder Representations from Transformers (BERT) becoming a widely used pre-trained language model for fine-tuning tasks. In this project, we investigate the performance of BERT on the AG News dataset, consisting of over 120,000 news headlines categorized into four classes: World, Sports, Business, and Sci/Tech.

We probed the frozen pre-trained BERT model using several sentence embedding strategies, including the [CLS] token, first and last token embeddings, and the mean of all token embeddings. These embeddings were evaluated using K-Nearest Neighbours (KNN) and multi-class logistic regression to examine the accuracy of the probed BERT models. Our results show that the mean token embedding strategy consistently outperformed the others on validation accuracy, with KNN slightly outperforming logistic regression.

To further improve performance, we fine-tuned the entire BERT model for the classification task using a subset of the AG News training data. Fine-tuning yielded the highest test accuracy of 92.14%, outperforming the best probing-based KNN (91.66%) and logistic regression (91.05%) models.

Finally, we examined the model's attention weights to gain interpretability insights, revealing that the fine-tuned BERT model often emphasizes end-of-sequence punctuation and other structurally significant tokens. These findings suggest that both probing and fine-tuning can achieve high performance on news topic classification, but fine-tuning offers a marginal yet consistent edge.

Dataset

The AG News dataset is a large-scale text classification dataset that contains news headlines and short descriptions, each labeled with one of four categories: World, Sports, Business, and Sci/Tech. The training and testing splits were loaded

directly, providing 120,000 training samples and 7,600 test samples. Each entry in the dataset includes a single text field and an associated numerical label, where 0 corresponds to World, 1 to Sports, 2 to Business, and 3 to Sci/Tech. The original training dataset is perfectly balanced, with 30,000 examples per class.

To reduce training time during development and experimentation, a smaller, balanced subset of the training set was created. This subset was constructed by sampling 2,500 examples from each of the four classes, resulting in a total of 10,000 samples. The subset was shuffled to ensure a random distribution of classes throughout the dataset and saved as a CSV file for later use. The test dataset was also converted to a Pandas DataFrame but was left unchanged to preserve its original structure and distribution.

Since BERT operates directly on raw text, no manual feature extraction or text preprocessing was required. All tokenization and input formatting will be handled by a pre-trained BERT tokenizer during the model training phase. This streamlined approach makes it easier to prepare the data and take full advantage of BERT's language modeling capabilities.

Benchmark

The BERT model used in this project is the bert-base-uncased variant, which is one of the most commonly used pre-trained models provided by Hugging Face. It was originally developed by Google and introduced in the 2018 paper *"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"* by Devlin et al [1]. The architecture consists of 12 transformer encoder layers, each with 12 self-attention heads and a hidden size of 768, amounting to approximately 110 million parameters. This version is "uncased," meaning all input text is lowercase before tokenization, as the model was pre-trained on a lowercased corpus combining Wikipedia and BookCorpus. The tokenizer uses a WordPiece vocabulary with 30,522 tokens, and the model supports sequences up to 512 tokens in length. BERT was pre-trained using two objectives: masked language modeling (MLM) and next sentence prediction (NSP), enabling it to learn deep bidirectional representations of language. More technical specifications are available on Hugging Face's model page [2] and in the original paper.

In our experiments, we implemented two approaches for text classification using BERT: probing and fine-tuning. For the probing approach, we used a frozen BERT model to extract sentence-level embeddings for each document. We experimented with four embedding strategies: taking the [CLS] token from the final hidden layer, selecting the first token embedding, selecting the last non-padding token, and computing the mean of all token embeddings (excluding padding). These embeddings were then used as input features for two classifiers: K-Nearest Neighbors (KNN) and logistic regression.

The KNN classifier was tested using three different values of K: 1, 3 5, 7, 9, 11 ,13, 15, 17 and 19. For logistic regression, we used the LogisticRegression class from the Scikit-learn library with a maximum iteration limit of 1000. Each classifier was trained and validated on a split of the original training set to identify the best-performing configuration. The embedding strategy and classifier combination with the highest validation accuracy—mean embeddings paired with logistic regression—was then retrained on the full training data and used to make predictions on the test set.

For the fine-tuning approach, we fine-tuned all parameters of the bert-base-uncased model using the BertForSequenceClassification implementation from the HuggingFace Transformers library. The fine-tuning was performed on a randomly sampled subset of 10,000 training examples, of which 80% were used for training and 20% for validation. Input texts were tokenized using the standard BERT tokenizer, with sequences padded or truncated to a fixed length of 512 tokens. Data was loaded in mini-batches of size 32.

The training was conducted over three epochs using the AdamW optimizer with a learning rate of 2e-5. We applied a linear learning rate scheduler with 10% warm-up steps to gradually increase the learning rate at the beginning of training to the initial learning rate, which helps stabilize updates and can enhance model convergence, particularly in the early stages of fine-tuning large language models [3]. The model was trained using cross-entropy loss, and evaluation was performed after each epoch to monitor validation performance.

Result

We experimented with two primary approaches for classifying the AG News dataset using BERT: probing with frozen embeddings and end-to-end fine-tuning. For the probing setup, we extracted sentence embeddings from a frozen BERT model using four strategies: the [CLS] token, the first token, the last token, and the mean of all token embeddings (excluding padding). These embeddings were then used to train K-Nearest Neighbors (KNN) and Logistic Regression (LR) classifiers. Validation accuracy was recorded across different values of *K* for KNN and compared with logistic regression for each embedding strategy.

Table 1 summarizes the validation accuracies for different combinations of method, embedding strategy, and *K*. We observed that the mean embedding strategy consistently outperformed others across both classifiers. KNN with a k of 13 has the highest validation accuracy of 91.77%, followed by logistic regression with a validation accuracy of 91.42%.

Table 1: Validation accuracy for probing strategies

Method	Strategy	Best K	Validation Accuracy (%)
KNN	CLS	17	88.83
KNN	First	17	88.83
KNN	Last	7	89.00
KNN	Mean	13	91.77
Logistic Regression	CLS	N/A	90.36
Logistic Regression	First	N/A	86.71
Logistic Regression	Last	N/A	90.04
Logistic Regression	Mean	N/A	91.42

Among the probing methods, the best configuration, KNN with K=13 using mean token embeddings, achieved a test accuracy of 91.66%. The best-performing logistic regression model, also using mean embeddings, achieved a slightly lower test accuracy of 91.05%. Fine-tuning the entire BERT model led to further performance gains, as shown in Figure 1. After training for three epochs on a balanced subset of 10,000 training examples, the fine-tuned model reached a validation accuracy of 92.60% and a test accuracy of 92.14%. While end-to-end fine-tuning does improve the test accuracy on the AG News dataset, the gain is marginal. This suggests that the pre-trained BERT model, even without fine-tuning, already captures much of the task-relevant information needed for this dataset. One possible explanation is that AG News is a relatively simple classification task for which an untrained BERT is already suited. In contrast, fine-tuning may offer more substantial benefits on more complex or domain-specific classification problems.

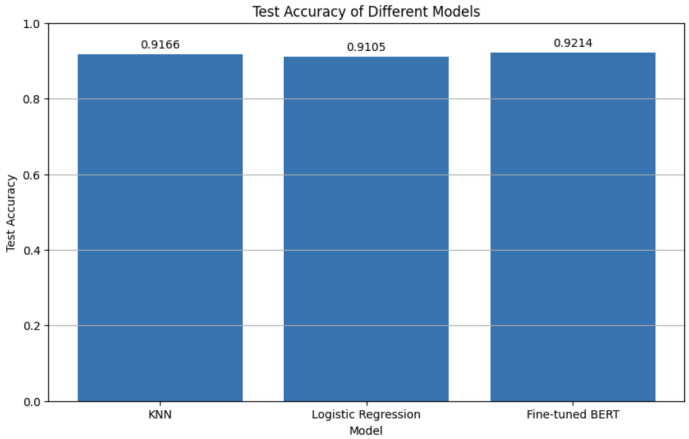

Figure 1: Bar plot comparing test accuracies for best probing and fine-tuned BERT

In this experiment, we examined the attention patterns within BERT's transformer architecture to gain insight into how the model attends to different tokens when performing classification on the AG News dataset. Specifically, we focused on Layer 11, Head 0 of the transformer and analyzed both correctly and incorrectly classified examples from each of the four categories: World, Sports, Business, and Sci/Tech. Our analysis centered on the [CLS] token, which serves as a summary representation of the entire input sequence and plays a critical role in the model's final prediction. By examining the attention weights between [CLS] and other tokens, we identified which words the model considered most important for its decisions.

As illustrated in Figure 2, for a correctly classified World example, the most attended token was ")", followed by "press," "sister," and "slams." The heatmap highlights the salience of ")" with a strong red intensity, indicating a high attention score. While seemingly non-semantic tokens like punctuation can sometimes receive attention due to formatting patterns in the training data, the presence of content-rich words like "press" and "slams" reinforces the model's alignment with class-relevant cues. To better understand how attention might contribute to misclassification, we conducted the same analysis on two incorrectly predicted examples from the Business and Sci/Tech classes. As shown in Figure 3, these heatmaps reveal more diffuse or misplaced attention, with the [CLS] token focusing on less informative or ambiguous words. Notably, terms such as "Britain" or "Olympics" appeared prominently, terms that often correlated with Sports but, in these cases, belonged to articles labeled as World. This confusion suggests that BERT occasionally over-relies on surface-level associations rather than deeper semantic context, contributing to misclassification.

For the Sports category, the correctly classified example (Figure 4) showed strong alignment between BERT's attention and semantically relevant terms such as "medley," "minutes," and "individual," all of which are contextually tied to athletic performance or event timing. The attention weights for these top tokens ranged from 0 to 0.20, indicating a moderately concentrated distribution focused on key sport-related vocabulary. In contrast, the misclassified sample (Figure 5) exhibited diffuse and low-impact attention weights, less than 0.035, spread across less meaningful or ambiguous words such as "appear," "like," "doesn't," and "adjustments." This suggests a failure of the model to identify or emphasize salient features necessary for accurate classification in this instance.

In the Business class, the correct prediction (Figure 6) was anchored by high-attention tokens including "pension," "workers," "unions," "firm," and "federal," all of which resonate strongly with labor and economic discourse. The attention weights in this case fell within the range of 0 to 0.15, reflecting a focused and coherent distribution that contributed meaningfully to the final classification. Conversely, the incorrectly predicted example (Figure 7), misclassified as Sci/Tech, included prominent attention on words such as "apparel," "fashion," "retailers," and "style." These terms, though potentially overlapping in consumer-focused content, reflect a domain drift. Notably, the attention weights in the incorrect example reached up to 0.16, comparable in magnitude to the correct case, but likely misallocated semantically.

For the Sci/Tech class, correct predictions (Figure 8) were driven by tokens such as "X," "space," and "launch," all of which are high-salience indicators of scientific or technological content. However, the attention distribution in this instance was relatively flat, with weights ranging only between 0 and 0.05, suggesting a broader and more uniform spread of focus across tokens. This may reflect BERT's confidence in prediction stemming from a collective signal across many low-weighted but relevant terms. In contrast, the misclassified Sci/Tech example (Figure 9), incorrectly labeled as Business, demonstrated a significantly broader range of attention weights, spanning from 0 to 0.16. Key words receiving higher attention included "cards," "card," "nets," "fraud," and "stolen." These words, while potentially indicating cybersecurity themes, may have triggered business-related associations in the model, resulting in categorical confusion. The wider weight range here suggests less discriminative attention targeting compared to the more narrowly distributed focus observed in correct classifications.

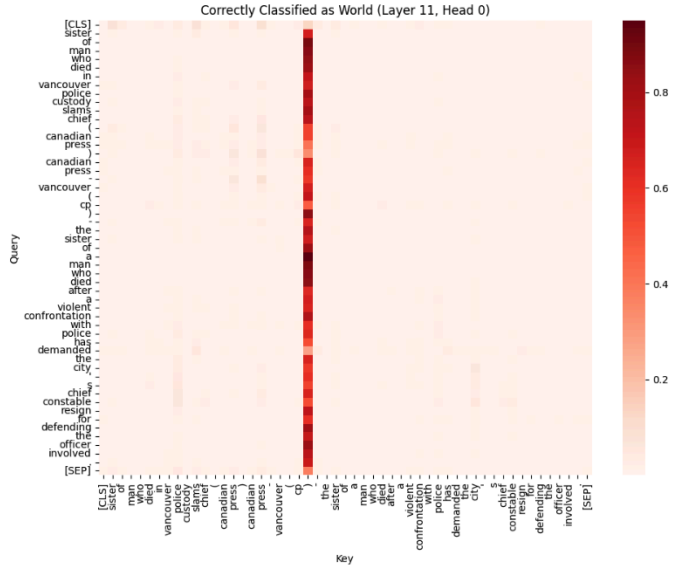

Figure 2. Correctly Classified World Attention Matrix and Attention Between The Top Words and the [CLS] Token

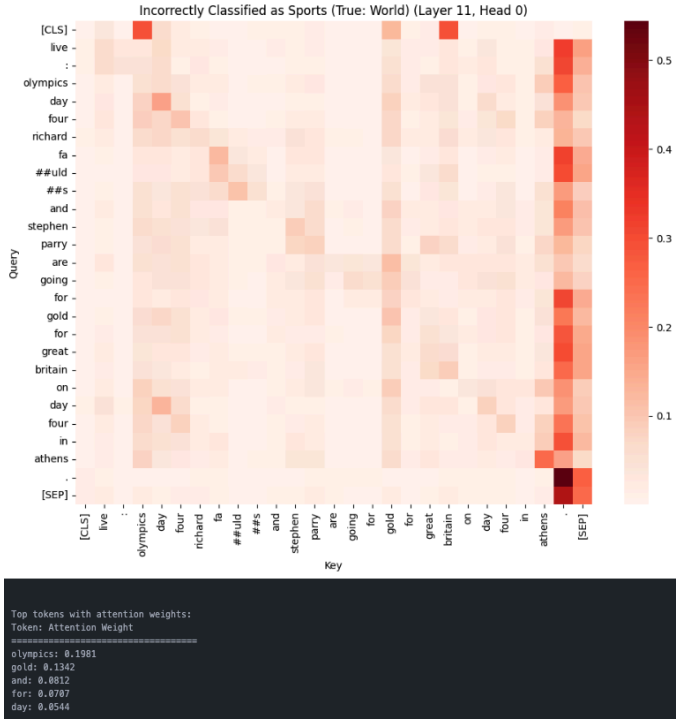

Figure 3. Incorrectly Classified World Attention Matrix and Attention Between The Top Words and the [CLS] Token

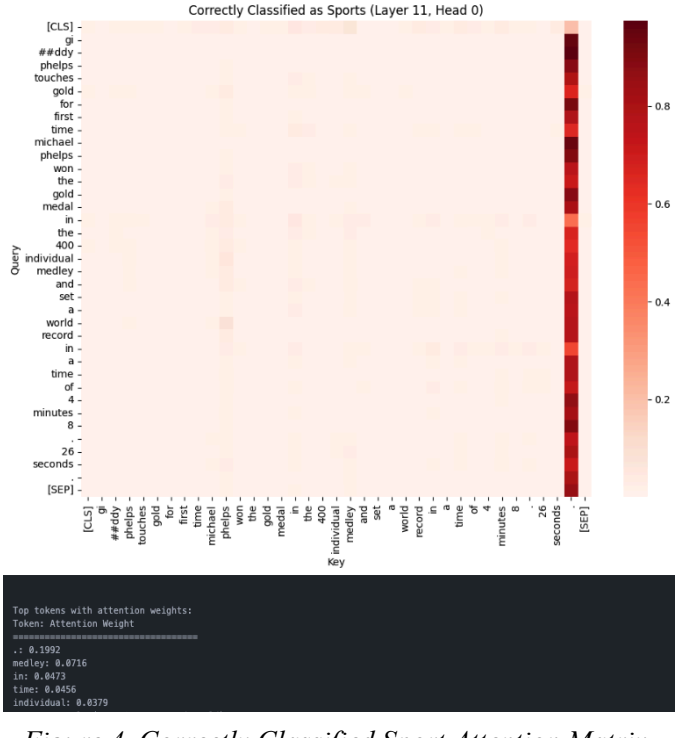

Figure 4. Correctly Classified Sport Attention Matrix and Attention Between The Top Words and the [CLS] Token

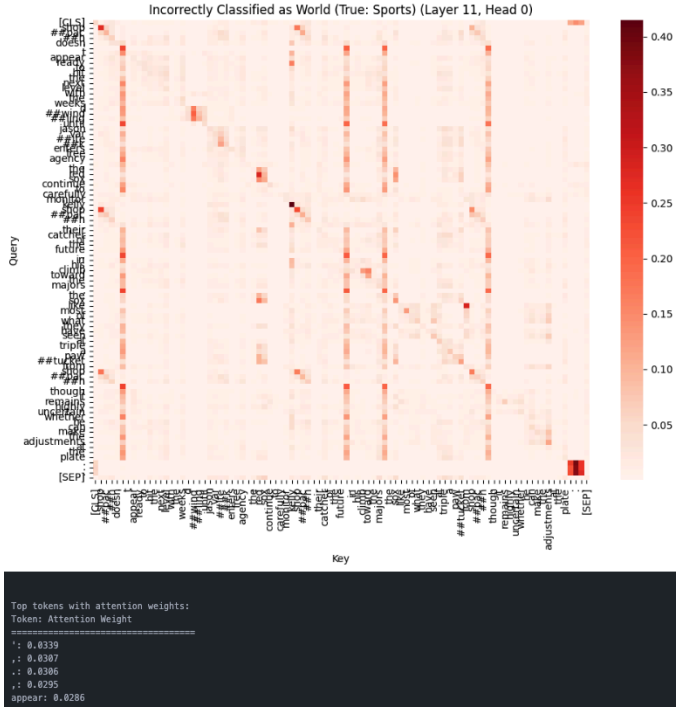

Figure 5. Incorrectly Classified Sport Attention Matrix and Attention Between The Top Words and the [CLS] Token

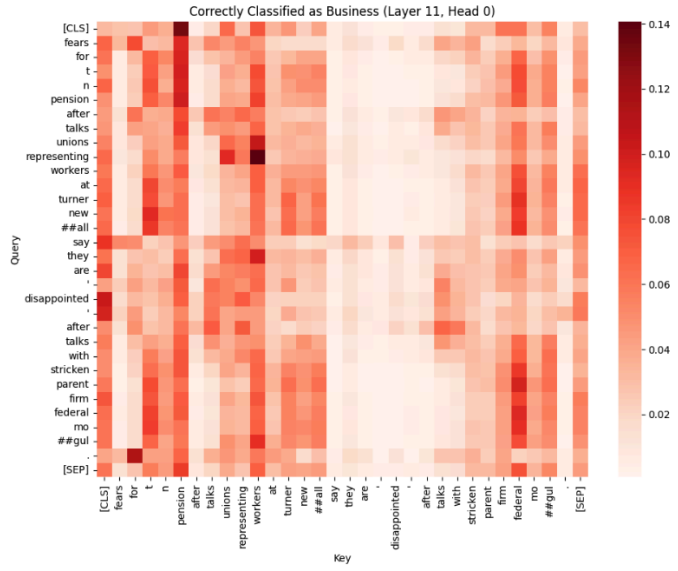

Figure 6. Correctly Classified Business Attention Matrix and Attention Between The Top Words and the [CLS] Token

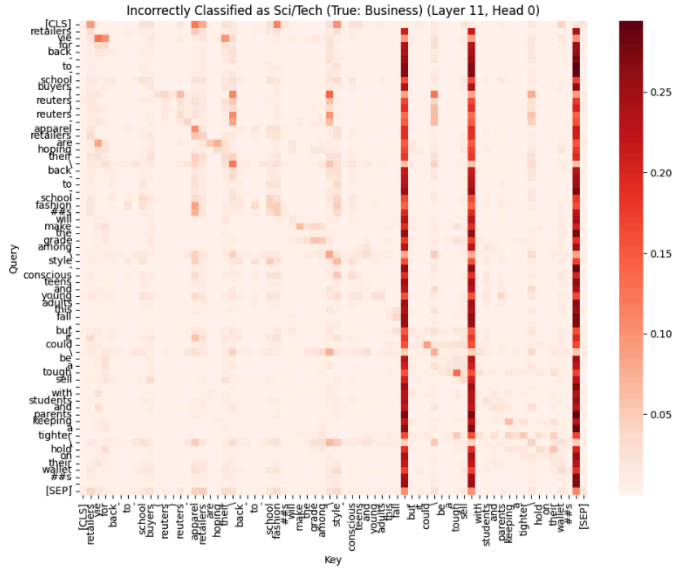

Figure 7. Incorrectly Classified Business Attention Matrix and Attention Between The Top Words and the [CLS] Token

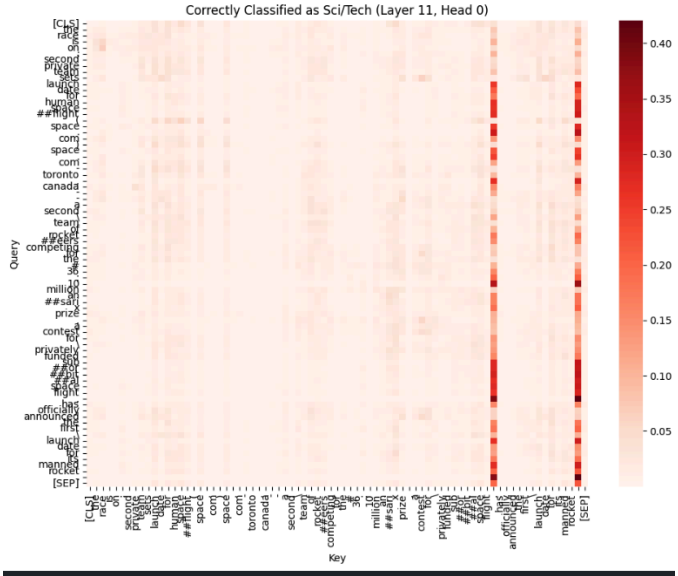

Figure 8. Correctly Classified Sci/Tech Attention Matrix and Attention Between The Top Words and the [CLS] Token

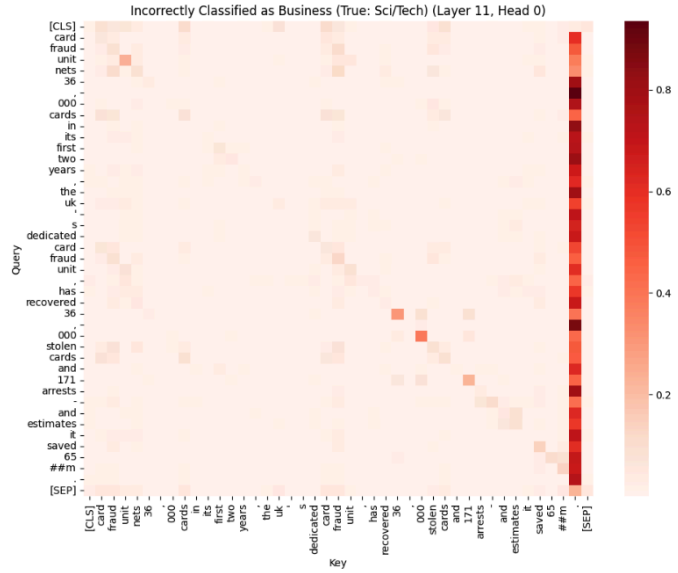

Figure 9. Incorrectly Classified Sci/Tech Attention Matrix and Attention Between The Top Words and the [CLS] Token

Extra Experiments

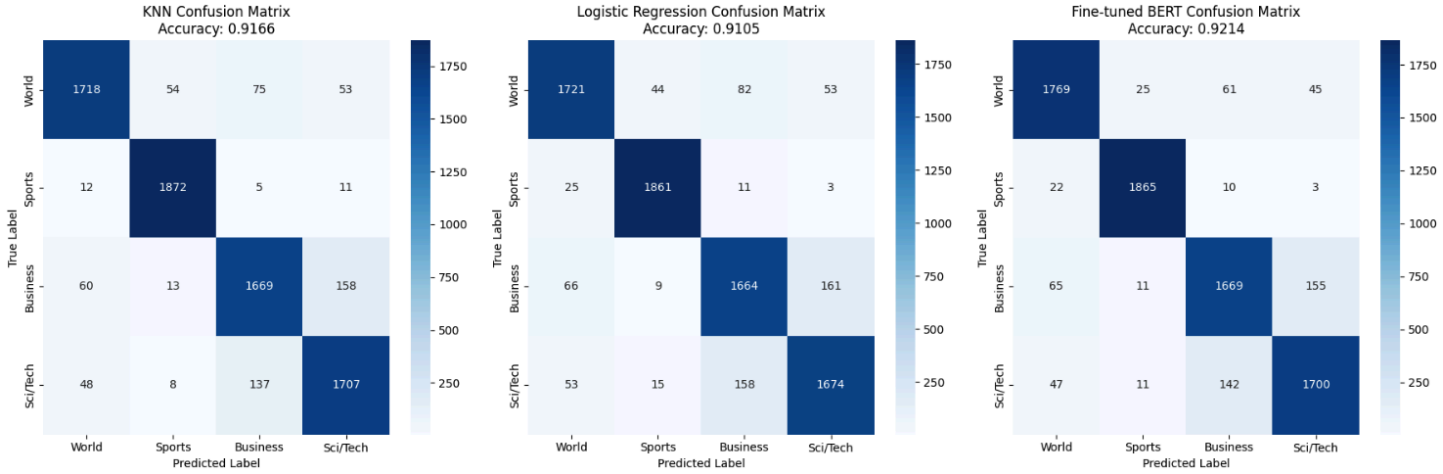

Figure 10. Confusion Matrix of the probed models and the fine-tuned BERT on the test set

For the extra experiment, we opted to analyze the confusion matrix (Figure 10) of the probed KNN, logistic regression, and the fine-tuned BERT to give us further insight into the accuracy of our models. While overall accuracy provides a general performance measure, the confusion matrices summarize how each model performs across individual classes. This helps identify specific misclassification patterns. For instance, all three models showed notable confusion between the Business and Sci/Tech categories, suggesting feature overlap between these classes. However, overall, the misclassification does not seem to have a bias towards one class. By examining these matrices, we better understand each model's strengths and weaknesses, which can guide future improvements in data preprocessing, class balancing, or model selection.

Discussion and Conclusion

In this assignment, we delved into the performance of a pre-trained BERT model on the AG News dataset. Our evaluation began with the untrained BERT embeddings, which we tested using KNN and logistic regression. We then fine-tuned the BERT model for comparison. The results of our experiments were reassuring, with both KNN and logistic regression achieving their highest accuracy using the mean embedding strategy. The fine-tuned BERT model, however, outperformed them all, reaching an impressive test accuracy of 92.14%.

The comparable performance of the untrained BERT embeddings and the fine-tuned BERT model hints at the potential of BERT's untrained embeddings. It suggests that these embeddings already encapsulate crucial patterns for classifying the AG News dataset. This reveals BERT's capability in classifying relatively simplistic classification problems such as on the AG News dataset.

Our analysis of the confusion matrices provided us with a deeper understanding of each model's performance. As for the future, we can explore more complex datasets for text-based classification. This research might reveal a more significant improvement in fine-tuning compared to the probed models.

Statement of Contribution

Each team member contributed to the coding and writing part rather equally. Pair programming sessions were held where all members contributed to ensure full understanding of the models and implementation before going on to produce results (graphs) and writing the report.

References

[1] Devlin, Jacob, et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv, 2018, https://arxiv.org/abs/1810.04805.

[2] Hugging Face. *google-bert/bert-base-uncased*. Hugging Face, https://huggingface.co/google-bert/bert-base-uncased.

[3] Y. Lu, "Glossary: LLM fine-tuning hyperparameters." Modal.com. https://modal.com/blog/fine-tuning-llms-hyperparameters-glossary-article (accessed Apr 18, 2025).