Experimenting BERT on AG news

Gordon Ng Colin Xiong Kynda Nashif

April 22, 2025

Abstract

In this project, we investigated the performance of the BERT base4 model on classifying the AG News dataset using existing PyTorch libraries. We compare fine-tuned BERTS with KNN and multi-class logistic regression probing. We found that BERT achieved better accuracy than traditional ML methods and was significantly slower to train. We achieved test accuracy of 87.50% using probing and 92.39% via end-to-end fine-tuning. We also examined the textual content of correctly and incorrectly classified examples to understand potential model behaviors, hypothesizing how attention mechanisms might focus on relevant or misleading terms.

1. Introduction

In this project, we investigate the effectiveness of using a pre-trained (Bidirectional Encoder Representations from Transformers) BERT model for document classification on the AG News dataset, comparing K-Nearest Neighbors (KNN), multi-class logistic regression for probing with end-to-end fine-tuned BERT. BERT, introduced by Devlin et al. (2019), is a language model pre-trained using masked language modeling and next-sentence predictions. Its contextual understanding of text allows it to be fine-tuned for specific downstream tasks like document classification.

AG News is a widely used dataset for text classification tasks, consisting of more than 1 million news articles categorized into four classes: World, Sports, Business, and Sci/Tech. The goal is to classify each document into one category based on keywords in the description.

This task is a valuable test of the language models' contextualization and classification abilities and has practical applications in information retrieval, news filtering, and summarization. Furthermore, this dataset fuels incredibly crucial research in data mining, data compression, data streaming, and more [1].

Our findings show that BERT's pre-trained representations allow simpler probing methods, like KNN and logistic regression, to perform reasonably well. However, fine-tuning and training BERTS through stochastic gradient descent significantly improves performance. Previous studies on BERT's capabilities align with our findings [2].

We also examine attention maps to gain insights into how BERT processes text, identify patterns in both correct and incorrect prediction patterns, and employ different strategies for sentence-level representation extraction in our experiments to better understand the effects of different parameters.

2. Datasets

We performed exploratory analysis to better understand the data before training, such as inspecting class distributions and viewing representative samples from each category. Each sample had a category label and a short text description. The dataset is distributed evenly among the four categories: World, Sports, Business, and Sci/Tech. The dataset was also split into training (108000 articles), validation (12000 articles), and test (7600 articles) subsets to allow evaluation during probing.

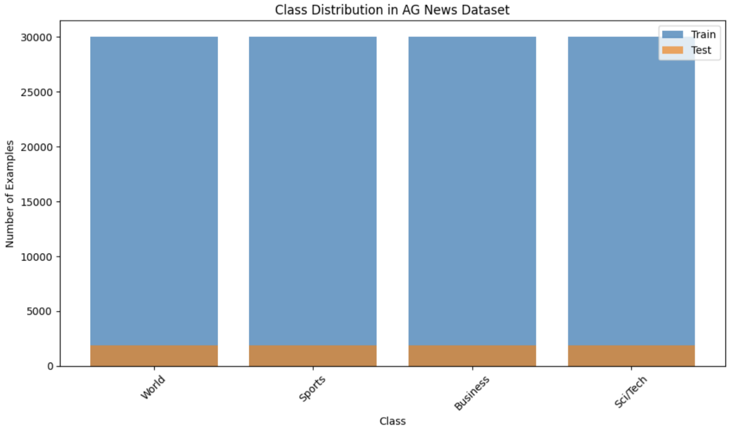


Figure 1: Class Distribution in AG News Dataset

Embeddings were then extracted from training and validation sets using different probing strategies. For efficiency, BERT embeddings were extracted using a subset of the training data.

For fine-tuning, the dataset was prepared by extracting a 15000-sample subset. Custom Pytorch datasets were then created for training, validation, and testing. Texts were tokenized, then padded or truncated to a fixed length (128 tokens), and converted into input tensors, including token IDs and attention masks. The data was then batch-loaded via a DataLoader, ready to be used as input for fine-tuning.

3. Benchmark

The architecture used in this project is the bert-base-uncased model, which is a widely used pre-trained BERT variant containing 12 transformer layers (encoder blocks), 12 self-attention heads per layer, and a hidden size of 768 [2]. This BERT model is loaded using Hugging Face's Transformers library and is initially kept frozen, meaning that the weights are not updated, during the probing experiments. Probing involves extracting sentence-level embeddings from the frozen BERT using four strategies: the [CLS] token embedding, the first token, the last non-padding token, and the mean of all token embeddings (mean pooling). These embeddings are then used as feature vectors for two classifiers: K-Nearest Neighbors (KNN) and logistic regression.

For KNN, we test k values of 1, 3, 5, 7, 9, 11, 13, 15, using Scikit-learn's KNeighborsClassifier. Each embedding strategy is evaluated using all k values, and the best-performing pair is chosen based on validation accuracy. Logistic regression is implemented as a multi-class classification task, using Scikit-learn's LogisticRegression. Like KNN, it is trained on the BERT embeddings and evaluated on validation data to determine the best strategy.

In the fine-tuning stage, the full BERT model is trained end-to-end using BertForSequenceClassification, which adds a classification head on top of the final [CLS] embedding. The model is trained on a subset of the AG News training data for 2–3 epochs using the AdamW optimizer, and the learning rate is set to 2e-5. This fine-tuning process allows all of BERT's internal weights to be updated using stochastic gradient descent.

4. Results

This section presents the experimental results, starting with the performance of the probing strategies on the validation set, followed by the fine-tuning process, the final comparison on the test set, and an analysis of selected classification examples.

4.1. Probing Performance on Validation Set

We evaluated the four sentence-level embedding strategies ([CLS] token, first token, last non-padding token, mean pooling) using KNN and multi-class logistic regression. Embeddings were extracted from a 5,000-sample subset of the training data, and models were evaluated on embeddings from a 1,000-sample validation set.

For KNN, we tested K values from 1 to 15 (odd numbers). The best validation accuracies were:

- **[CLS] Token:** 84.60% (at K=7)
- **First Token:** 84.60% (at K=7)
- **Last Token:** 83.90% (at K=5 and K=13)
- **Mean Pooling:** 89.50% (at K=9)

Mean pooling significantly outperformed other strategies for KNN.

For multi-class logistic regression, the validation accuracies were:

- **[CLS] Token:** 86.80%
- **First Token:** 86.80%
- **Last Token:** 87.50%
- **Mean Pooling:** 89.60%

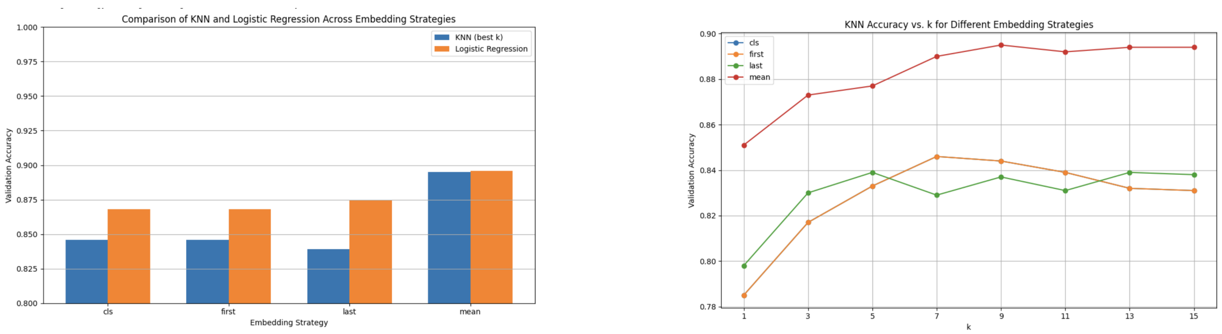
Again, mean pooling yielded the highest accuracy for logistic regression.

Table 1 summarizes the best validation accuracies achieved for each strategy. Based on these results, the 'mean' pooling strategy combined with logistic regression (89.60% accuracy) was identified as the best overall probing approach on the validation set.

Embedding Strategy	Best KNN Accuracy (K)	Logistic Regression Accuracy
[CLS] Token	84.60% (K=7)	86.80%
First Token	84.60% (K=7)	86.80%
Last Token	83.90% (K=5, K=13)	87.50%
Mean Pooling	89.50% (K=9)	89.60%

Table 1: Best Validation Accuracies for Probing Strategies (on 1000 validation samples)

Figure 2 shows the validation findings graphically. Figure 2a presents a bar chart comparing the peak KNN accuracy and the logistic regression accuracy for each embedding method, visually emphasizing the strong performance of the mean pooling strategy. Figure 2b displays line plots illustrating KNN accuracy as a function of K for each embedding strategy, highlighting how accuracy changes with K and showing the peak for mean pooling occurring at K=9.



(a) Comparison of Best KNN and LogReg Accuracy Across Embedding Strategies

(b) KNN Accuracy vs. k for Different Embedding Strategies

Figure 2: Performance Comparison of Different Embedding Strategies for Text Classification

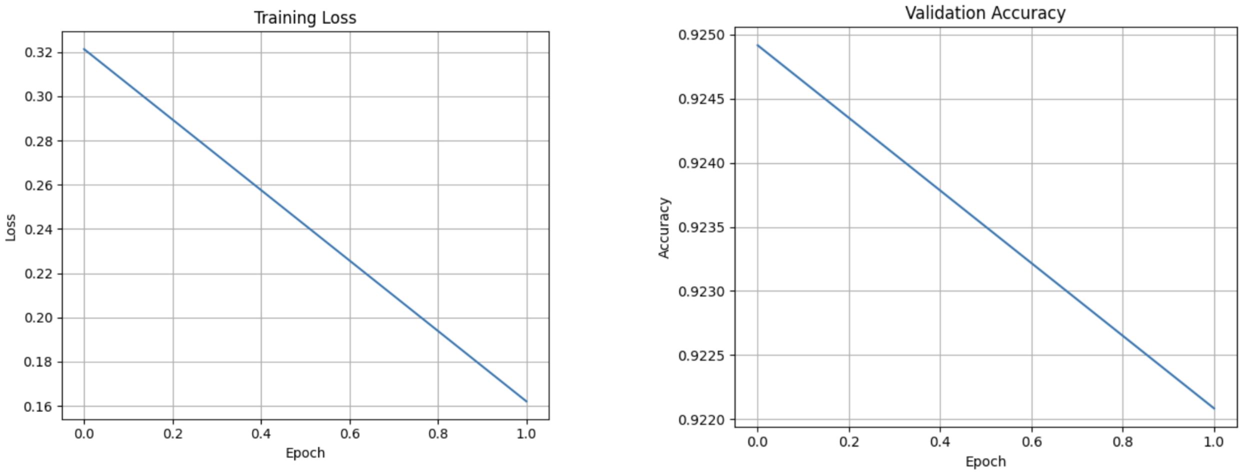
4.2. Fine-tuning Performance

The ‘BertForSequenceClassification‘ model was fine-tuned on a 15,000-sample subset of the training data for 2 epochs, using the 12,000-sample validation set for monitoring. The performance over epochs was:

1. **Epoch 1:** Average training loss = 0.3213, Validation accuracy = 92.49%
2. **Epoch 2:** Average training loss = 0.1621, Validation accuracy = 92.21%

The training loss decreased substantially, indicating effective learning. Validation accuracy peaked after the first epoch, suggesting that further training might lead to slight overfitting on this specific validation set or that optimal performance was reached quickly.

Figure 3 illustrates these trends. Figure 3a shows the decrease in average training loss across the two epochs. Figure 3b shows the validation accuracy achieved after each epoch, with the peak at epoch 1.



(a) Training Loss per Epoch

(b) Validation Accuracy per Epoch

Figure 3: BERT Fine-Tuning Performance Metrics Across Training Epochs

4.3. Reporting Classification Performances on Test Set

The final performance was evaluated on the 7,600-sample AG News test set. We compared the best probing models (determined using validation data: KNN with K=9 and Logistic Regression, both using mean-pooled embeddings) against the fine-tuned BERT model (using the state after epoch 1, which had the highest validation accuracy).

Model	Test Accuracy
KNN (mean, K=9)	86.40%
Logistic Regression (mean)	87.50%
Fine-tuned BERT	92.39%

Table 2: Test Accuracy Comparison Between Probing Methods and Fine-Tuned BERT

The fine-tuned BERT model significantly outperformed both probing methods on the unseen test data. The improvement over the best probing method (Logistic Regression) was 4.89 percentage points, representing a relative increase of 5.59%.

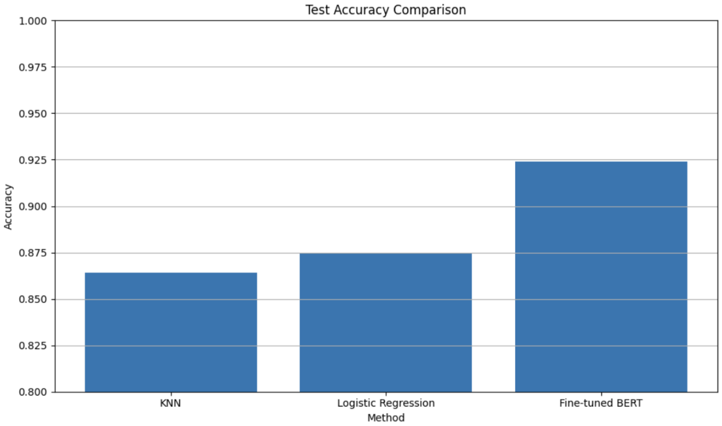


Figure 4: Test Accuracy Comparison: Best Probing vs. Fine-tuned BERT

4.4. Attention Matrix Visualization Analysis

As required, we examined examples of correctly and incorrectly classified documents from the fine-tuned model to hypothesize about the model's reasoning, particularly how attention mechanisms might function. We consider how attention might be distributed across tokens, influencing the final classification via the [CLS] token aggregation.

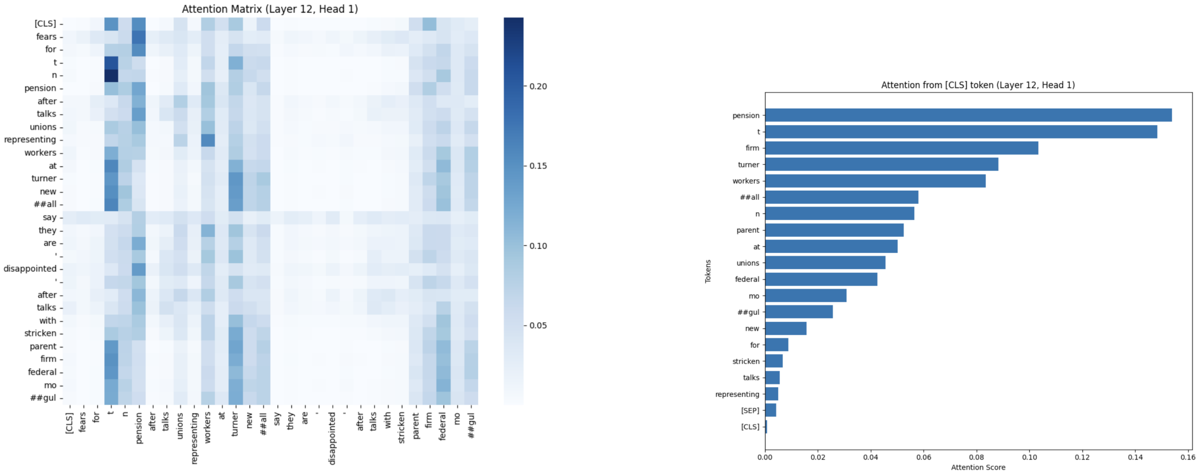
Correctly Predicted Examples:

1. **Text:** ”fears for t n pension after talks unions representing workers at turner newall say they are 'disappointed' after talks with stricken parent firm federal mogul.”

- True: Business, Predicted: Business (Confidence: 90.50%)
- *Analysis:* High attention was likely placed on keywords strongly indicative of business news, such as ”pension,” ”unions,” ”workers,” company names (”turner newall,” ”federal mogul”), and ”parent firm.” The model correctly identified the corporate/labor context.

2. **Text:** "the race is on : second private team sets launch date for human spaceflight (space. com) space. com - toronto, canada - - a second team of rocketeers competing for the #36; 10 million ansari x prize, a contest for privately funded suborbital space flight, has officially announced the first launch date for its manned rocket."

- True: Sci/Tech, Predicted: Sci/Tech (Confidence: 99.26%)
- *Analysis:* The very high confidence suggests strong, focused attention on highly discriminative Sci/Tech terms like "spaceflight," "space. com," "rocketeers," "ansari x prize," "suborbital," "launch date," and "rocket."



(a) Attention Matrix for Correctly Classified Example

(b) Attention Bar Graph for Correctly Classified Example

Figure 5: Attention Visualizations for Correct Examples

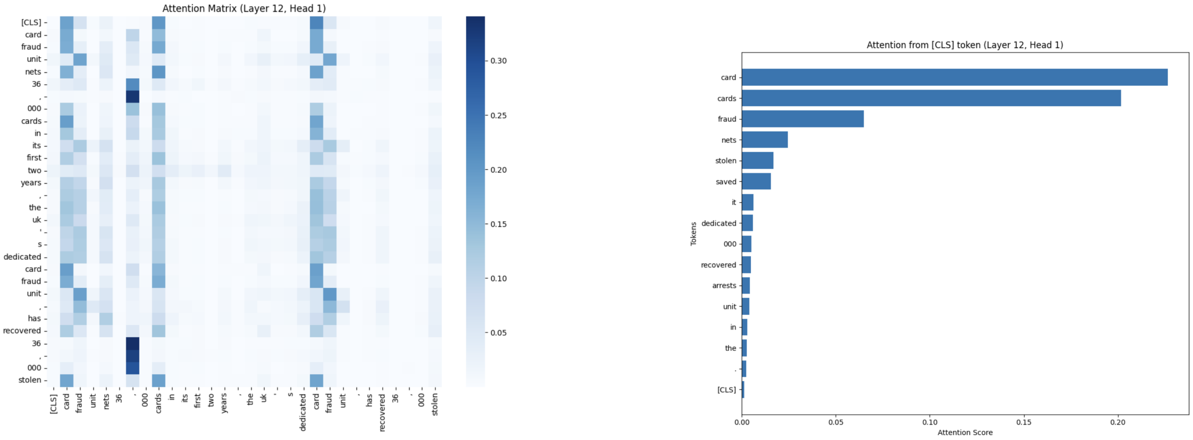
Incorrectly Predicted Examples:

1. **Text:** "card fraud unit nets 36, 000 cards in its first two years, the uk's dedicated card fraud unit, has recovered 36, 000 stolen cards and 171 arrests - and estimates it saved 65m."

- True: Sci/Tech, Predicted: World (Confidence: 38.80%)
- *Analysis:* The model likely focused on "uk's," "fraud unit," "stolen cards," and especially "arrests." While "card fraud unit" could hint at technology, the emphasis on location (UK) and crime/enforcement ("stolen," "arrests") might have outweighed the technological aspect, leading to the "World" prediction. The low confidence reflects this ambiguity.

2. **Text:** "rivals try to turn tables on charles schwab by michael liedtke san francisco (ap) - - with its low prices and iconoclastic attitude, discount stock broker charles schwab corp. (sch) represented an annoying stone in wall street's wing - tipped shoes for decades..."

- True: Sci/Tech, Predicted: Business (Confidence: 69.69%)
- *Analysis:* The text is dominated by terms strongly associated with finance and business: "Charles Schwab," "stock broker," "Wall Street," "prices," "corp." High attention on these terms likely led to the confident misclassification as "Business." The actual "Sci/Tech" label might be due to the source (e.g., a tech news outlet covering business) or a nuance missed by the model, which focused on the overt topic.



(a) Attention Matrix for Incorrectly Classified Example

(b) Attention Bar Graph for Incorrectly Classified Example

Figure 6: Attention Visualizations for Incorrect Examples

5. Discussion and Conclusion

Our experiments demonstrate BERT's effectiveness for AG News classification, highlighting a significant performance advantage for end-to-end fine-tuning over probing techniques. Fine-tuning achieved a test accuracy of 92.39%, substantially higher than the 87.50% obtained by the best probing method (logistic regression with mean-pooled embeddings). This superiority stems from fine-tuning's ability to adapt BERT's pre-trained representations specifically to the nuances of the news classification task, whereas probing relies on fixed, general-purpose embeddings. While probing offers a computationally cheaper alternative, giving respectable results, fine-tuning unlocks BERT's full potential for this task.

Within the probing strategies, mean pooling consistently provided the most informative sentence representation compared to using single tokens like [CLS], suggesting that aggregating information across the sequence is beneficial. Logistic regression slightly outperformed KNN on these embeddings. This confirms that even without task-specific training, BERT embeddings capture significant semantic information usable by traditional classifiers. This is where there is a trade-off: probing offers speed and resource efficiency, while fine-tuning prioritizes maximal accuracy at a higher computational cost.

Qualitative analysis of attention patterns suggests the fine-tuned model focuses on category-relevant keywords for correct classifications and can be misled by ambiguous terms or dominant but incorrect topic signals in misclassifications. Key limitations include the use of data subsets and limited training epochs/tuning. Future work could involve training on the full dataset, exploring more extensive hyperparameter optimization, employing different BERT variants, and conducting a more systematic analysis of attention mechanisms to further understand the model's decision-making process. Overall, BERT, particularly when fine-tuned, proves to be a highly capable tool for text classification on the AG News dataset.

Statement of Contributions

All team members contributed equally to the implementation, experimentation, and report writing for this project.

References

[1] Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *NIPS*, 2015.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.