BERT Fine-Tuning for Narration Classification

Vincent Gu
W266 - Natural Language Processing with Deep Learning
UC Berkeley School of Information

June 19, 2023

Abstract
This paper aims to assess the performance of BERT fine-tuning for a text classification task on novel fictional text that deviates significantly from BERT's pre-training corpus. Specifically, how does a fine-tuned BERT model handle an author's unique writing style and novel vocabulary for a text classification task? To shed light on this question, this paper analyzes a science fiction novel 'Dark Age' by Pierce Brown which is told from five fictional perspectives from vastly different socio-economic backgrounds. Findings indicate that the fine-tuned model was able to learn a significant amount by retraining BERT parameters and achieved a test accuracy of roughly 80% from a baseline of 20%. However, analysis indicated significant diminishing returns gained from retraining all BERT layers as opposed to only a fraction of them. The slight improvements of retraining over half of the BERT layers in the fine-tuned model must be weighed against the additional time and computational resources required to do so.

I. Introduction

As a model framework, BERT's pre-training on existing data including Wikipedia and the Books corpus provides an enormous repository of context and position based word embeddings suitable as a starting point for a myriad of language tasks. Tuning this pre-trained model can lead to increased task accuracy for specific tasks and datasets.

However, how does BERT perform when presented with novel fictional text that differs substantially from the data it was trained on? Specifically, how does BERT perform when exposed to a unique author's writing style and vocabulary outside of its training corpus? Furthermore, how well can BERT be fine-tuned with this novel fictional text to increase performance at a classification text? Finally, how well can this fine-tuned BERT model pick up subtle variations in word meanings based on the surrounding fictional context?

To begin answering these questions, I leveraged the base BERT model with the science fiction novel 'Dark Age' by Pierce Brown. In the novel Brown crafts a dystopia in which mankind is genetically engineered into twelve distinct castes (denoted as colors in the book). Each color holds a certain function in society and inter-color marriage is forbidden. For example, Gold as the ruling class are genetically engineered to be taller, stronger and smarter than Reds who toil as menial slave laborers.

The novel contains narration from 5 different individuals of different colors and perspectives and the corresponding language task involves classifying an excerpt of the text based on which narrator the model believes the text came from. I then looked to improve model performance by

fine-tuning. Finally, I analyzed various out-of-context words specific to the world Pierce Brown created to determine if the fine-tuned BERT could pick up nuanced word meanings, as represented by word embeddings, based on the narrator.

The five narrators (classification buckets) are:

1. Darrow: A Red genetically 'carved' into a Gold. A revolutionary warlord
2. Virginia: The Gold Sovereign (ruler) of a more egalitarian/reformed society
3. Ephraim: A Gray (former foot soldier)
4. Lyria: A Red (former slave) disillusioned with Virginia's society
5. Lysander: A Gold, the last grandchild of the former Sovereign, a fascist slaver

II. Background

This study drew heavily on Goter's "Comparison of Unsupervised Data Augmentation with BERT and XLNet" which came to the conclusion that unsupervised data augmentation did not significantly outperform fine-tuned BERT for a text classification task on fictional excerpts. While unsupervised data augmentation achieved the same performance as a fine-tuned BERT model with as much as a third less data, the increased complexity and time required for unsupervised data augmentation must be weighed against the time required to simply label additional data.

Related studies including Jiang et al. (2021) and Meng et al. (2020), indicate, as expected, that fine-tuning BERT on text of interest significantly improves performance in classification tasks in comparison to the base pre-trained models. All in all, leveraging labeled data to fine-tune a BERT model provides a solid base framework for text classification tasks.

III. Methods

i. Baseline Model

The baseline for this study was set to 20% for this multi-class text classification task into five categories. Naively, anyone could just randomly guess one of the five labels and expect to be correct one in five times. This baseline can serve as a simple heuristic as any improvements above it can be attributed to learning done by the model.

ii. Fine-Tuned BERT Model

Fine-tuning the BERT model involved tweaking the pre-trained model parameters via back propagation for each epoch. These learned token embeddings for the [CLS] token are then fed into a few fully connected dense layers which outputs into a final softmax classification layer into the five classes. The [CLS] embedding was chosen as an excerpt-level embedding which can serve as an input that represents the meaning of the entire excerpt. As we are attempting to classify each excerpt, we need the [CLS] embedding as a representation of the excerpt as a whole and not an embedding of individual words in the excerpt.

For the pre-trained model, the BERT-base-cased model (12-layer, 768 hidden layer dimension, 12 attention heads and 110M model parameters) was chosen as the base model over the large BERT models due to computational constraints. The learning rate was set low, at 0.00005, to prevent divergence. The hyperparameters of batch size, excerpt size and number of layers retrained were manually varied during the fine-tuning process.

iii. Hardware and Software Versions

As BERT's underlying transformer architecture requires significant computational resources, models were trained using GPU (NVIDIA

GeForce RTX 3070 GPU) as opposed to CPU (12th Gen Intel(R) Core i7-12700H). Attempting to train on CPU took approximately two hours per epoch which was cut to between 2-15 minutes per epoch with GPU. Models were trained on Python 3 and Tensorflow 2.12.0.

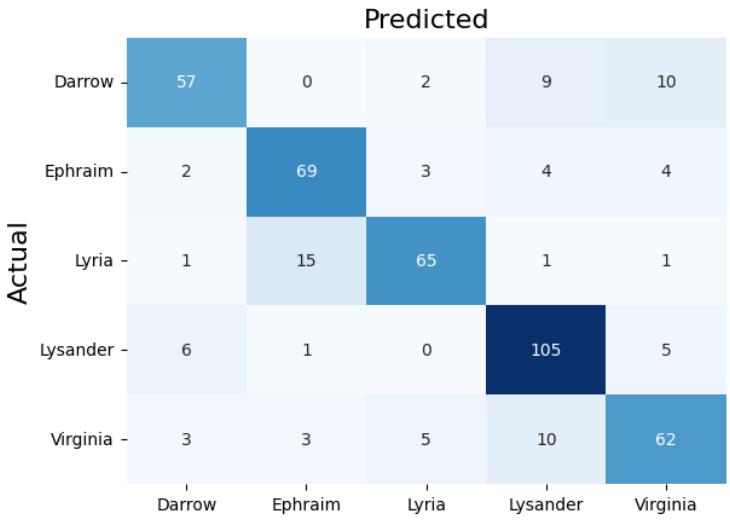
IV. Results

Following obvious intuition, we can clearly see how important retraining BERT layers is for fine-tuning for the classification language task. Models that did not retrain any BERT layers ended up with only a very slight increase in test accuracy over each epoch and after ten epochs barely crested 35% test accuracy. Without retraining BERT layers the only parameters updated in back-propagation are those in the fully connected dense layers which is clearly insufficient. While this is a substantial increase over the baseline heuristic of 20%, this demonstrates that solely retraining parameters in the dense layers is wholly insufficient.

However, models that retrained BERT parameters were able to learn a significant amount during training with final epoch validation accuracies for the best performing models around 80%. We see in initial epochs many models start off with about 20% validation accuracy - which indicates that at this point the model isn't performing any better at sorting into the five bins than it would if it was just randomly guessing. However, the substantial increases in validation accuracy over epochs demonstrates BERT's ability to learn. We can clearly see the improvements in test accuracy reaped from the fine-tuning process.

Despite this, we do see significant diminishing returns of training all twelve BERT layers as opposed to only half of them. For the model with batch size 16, excerpt size 128 and all 12 retrained BERT layers we had a validation accuracy of 80.81% but for the model with same batch size, excerpt size but only 5 retrained BERT layers we had a validation accuracy of 79.01%. This insignificant increase in accuracy was also accompanied by an increase of roughly 100 seconds per training epoch which indicates a significantly increased strain on computational resources for training all BERT layers over only half of them.

The best performing model used batch size 16, excerpt size 128 and 12 retrained BERT layers with a validation accuracy of 80.81%.



The above confusion matrix details for which narrators the model performed best and worst on. The model performs very well at identifying the 'Lysander' perspective and performs moderately well on most of the others. Lysander is the last scion of the former, slaver regime and thus has an extremely different perspective on the events of the novel in comparison to the other narrators.

Note that the most substantial misclassification involved fifteen 'Lyria' excerpts identified as 'Ephraim'. Ephraim, alongside Lyria, is the only other non-Gold narrator in the novel, as Darrow is a de-facto Gold, who doesn't lead armies or govern.

Alse note that for each of 'Darrow', 'Lysander' and 'Virginia' (the Gold or de-facto Gold rulers) there is substantially more overlap with each other's excerpts than with the others.

Finally, note that only a single ‘Lyria’ excerpt was identified as ‘Lysander’ and no ‘Lysander’ excerpts were identified as ‘Lyria’. From the novel’s perspective, Lysander is a fascist slaver and Lyria is a common former slave so intuition dictates that their perspectives and passages would differ greatly - which is demonstrated by the model’s superb performance in distinguishing between these two narrators.

The following table displays summary statistics for model performance.

	Precision	Recall	F1-Score	Support
Darrow	0.83	0.73	0.78	78
Ephraim	0.78	0.84	0.81	82
Lyria	0.87	0.78	0.82	83
Lysander	0.81	0.90	0.85	117
Virginia	0.76	0.75	0.75	83
accuracy			0.81	443
macro avg	0.81	0.80	0.80	443
weighted avg	0.81	0.81	0.81	443

V. Discussion

i. Analysis of Misclassifications

An insignificant number, fifteen, of ‘Lyria’ excerpts were misidentified as ‘Ephraim’. In the novel, Lyria and Ephraim's stories interweave significantly which could have contributed to the misclassifications. In these 'Lyria' passages we see direct references to Ephraim's character as well as to Volga, another character central to Ephraim's story. We also see references to both drug use and firearms which mirrors Ephraim's drug abuse and his life as an ex-soldier. Both narrators also frequently refer to individuals of the ruling class as ‘the Gold’. This out-of-vocabulary phrase encompasses a unique meaning only understood in the context of the novel.

Furthermore, looking at the five ‘Lysander’ excerpts that were misidentified as ‘Darrow’ we see similar trends hold. Again, Lysander and Darrow’s stories interweave as they interact with each other. In these passages we see a direct reference to 'Darrow' as well as multiple references to 'Alexander' who is another key character in both Lysander's and Darrow's stories.

ii. Analysis of Out-of-Context Words

To determine if and how the BERT fine-tuning process learned contextual word meanings for select out-of-context words, the last hidden-state vector for the [CLS] token for key words were taken from both the original pre-trained BERT and my fine-tuned BERT layers. Each BERT token embedding has a dimensionality of 768.

To empirically calculate the similarities between these token embeddings these last hidden-state vectors were fed into a cosine similarity function. Two identical vectors would have a cosine similarity of one and two opposite vectors, normal to each other in 768 dimensional space, would have a cosine similarity of zero.

First, a comparison of ‘Gold’ embeddings between the pre-trained and fine-tuned BERT model yielded a cosine_similarity of 0.52. This indicates that the model’s learned meaning for the ‘Gold’ changed significantly during the fine-tuning process. While in normal conversation ‘Gold’ would likely refer only to the metal, in the novel ‘Gold’ is a nuanced word indicative of the ruling class and therefore holds a significantly different meaning. Similarly, the cosine similarity of ‘Red’ between the

pre-trained and fine-tuned BERT model yielded a cosine similarity of only 0.41.

Furthermore, the learned embeddings from the fine-tuned model for ‘Red’ and ‘Gold’ have a cosine similarity of 0.82. We also see that pre-trained 'ruler' and fine-tuned 'Red' are very different from each other (0.40) but a bit closer between pre-trained 'ruler' and fine-tuned 'Gold' (0.53). This suggests that the model learned something about the novel’s caste system and about the intrinsic meaning of the out-of-context words.

VI. Conclusion

The above analysis demonstrates fine-tuning the BERT model over several epochs significantly improves the model’s performance at text classification for a novel text input. Performance far outstrips the baseline heuristic of randomly choosing labels and therefore demonstrates the model’s ability to learn.

Furthermore, retraining BERT layers provides significant improvements in performance over simply retraining fully connected dense parameters. It is essential to note that there are significant diminishing marginal returns to training more and more BERT layers as the model that retrained on all twelve layers only yielded a marginally higher test accuracy over the model that retrained on only six layers despite taking almost a third longer to train. This improvement must be weighed against the increased computational complexity of retraining additional layers.

The best performing model was the most successful at distinguishing between narrators that most readers would also determine to be the most different from each other. On the flip side, it struggled more at classifying narrators with more similar socio-economic backgrounds and those that shared similar interwoven stories. Misclassified excerpts often included the literal names of the narrator it was misclassified as. Additionally, they often included additional character names central to the stories of multiple narrators and key words and phrases shared between characters.

While this model does not serve any future application beyond the fictional universe created by Pierce Brown, it serves as an example of the importance of fine-tuning base models for specific text tasks and, more importantly, demonstrates the model’s ability to distinguish between narrators of varying socio-economic and political backgrounds. A similar study could be conducted on various news sources to determine author biases (authoritarian/democratic, dirigisme/libertarian, liberal/conservative) and assign labels to news sources to better inform readers of the potential skew encoded in whatever article they choose to read.

References

I. Thomas Goter. (2019). Comparison of Unsupervised Data Augmentation with BERT and XLNet, December 2019.

II. Ming Jiang, Yuerong Hu, Glen Worthey, Ryan C. Dubnicek, Ted Underwood and J Stephen Downie. (2021). Impact of OCR Quality on BERT Embeddings in the Domain Classification of Book Excerpts, November 2021.

III. Anna Glazkova. (2021). Exploring cross-genre performance for age-based fiction classification models, November 2021.

IV. Yu Meng, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, Jiawei Han. (2020). Text Classification Using Label Names Only: A Language Model Self-Training Approach, October 2020.