**Assignment 4: Fine-Tuning a Transformer for Sentiment**

Jorge Castro – U58196429

Department of Computer Sc. and Eng.

University of South Florida, Tampa, Florida, USA

**Introduction**

This report explores the performance of the BERT model for sentiment analysis through various fine-tuning methods. The base BERT model was evaluated against modifications including added layers, increased batch size (2048), and a shift from the Adam to the Stochastic Gradient Descent (SGD) optimizer.

Additionally, the model was tested on inputs outside the typical IMDb corpus to assess its adaptability to different contexts, such as sentiments on the performance of a software application, personal relationships, and food criticism. The results detail how each configuration impacts the accuracy and efficiency of sentiment detection, offering insights into optimizing BERT for diverse and enhanced NLP applications.

**Small Bert - Base Model**

To establish a baseline sentiment classification performance, all models are compared against this baseline model from TensorFlow’s Bert Classification Tutorial. The model is built using

**small\_bert/bert\_en\_uncased\_L-4\_H-512\_A-8**

**small\_bert:** Indicates a smaller, more resource-efficient variant of BERT.

**bert\_en\_uncased:** The model uses lowercase English text.

**L-4:** The model has 4 layers, making it lighter and faster. The number of layers (L-4) in a neural network, particularly in a transformer model, influences both the complexity and the depth of learning

**H-512:** Each layer outputs embeddings of size 512, affecting learning capacity. A higher dimensionality of embeddings allows the model to encode more information into a denser representation, potentially capturing more complex relationships in the data. This can enhance the

model's ability to learn and generalize but also increases the computational load and memory usage.

**A-8:** The model uses 8 attention heads, allowing it to focus on different parts of the input simultaneously, making it more robust in capturing various dependencies (e.g., syntactic and semantic relationships in text). Each head can potentially focus on different "features" or "parts" of the data, providing a more comprehensive understanding.

**Base Model Plot**

Diagram

Description automatically generated

**Small Bert - Base Model Performance**

Please refer to Figure 1 in the Addendum to see the base model's performance. Notably, the model demonstrates declining performance when dealing with inputs containing 50/50 opinionated sentiments. Words like "meh" and "okish," which are highly opinionated, are poorly handled. For example, a comment like "the movie was okish" should not be equated with stronger sentiments such as "I totally hated this movie," yet the model scores them similarly. Additionally, the model fails to recognize certain vocabulary, such as "phenomenal," but can accurately extract sentiment from known words even when inputs are censored with "\*\*\*." In our class discussions, we observed that models pre-processed to handle misspellings and random typos generally perform better. This was empirically evident with this small BERT model.

Furthermore, when testing the model outside of its training corpus, it performed well, detecting sentiment accurately in contexts unrelated to movie reviews. This suggests the model's robustness in general sentiment analysis across different domains.

**Large-Bert Model Performance**

**bert\_en\_uncased\_L-12\_H-768\_A-12**

**bert\_en\_uncased:** The model uses lowercase English text.

**L-12:** The model has12 layers x3 small BERT

**H-512:** Same as small BERT

Refer to Figure 2 for a comparison of the larger BERT model vs the smaller base model. Here we observe some notable differences. The larger BERT model, is equipped with a broader vocabulary, can correctly identify words like "phenomenal." However, it has lost the capability to detect typos or censored words—a strength observed in the smaller BERT model. For example, the method of censoring part of a word with "\*\*\*" no longer results in successful word completion, even when the first and last letters are visible.

This change likely stems from the training scope and inherent trade-offs in model training. The smaller model may utilize data augmentation techniques like introducing typos, which helps it handle a variety of input errors effectively. This compensates for its limited data scope by enhancing its robustness. In contrast, the larger model, trained on a more extensive set of real-world data, tends to overfit, focusing more on well-formed text and losing flexibility in processing irregular inputs like censored words or typos. This demonstrates a common trade-off in machine learning: a model's complexity and its ability to generalize across different types of input.

**Other Comparisons**

Other alternative version of the small BERT was fine-tuned as well to make comparisons with our based model. Multiple iterations involved implementing a multi-hidden layer classifier, increasing the batch size, and replacing the Adam optimizer (Adaptive Moment Estimation) with SGD (Stochastic Gradient Descent). However, none of these modifications led to significant improvements in the small BERT model's performance.

It is noted that increasing the batch size somewhat improved the model's ability to classify moderately opinionated sentiments, like "meh" or "okish," approximately around the 50% mark. Yet, this observation could be biased, as different applications may require either highly precise or broadly generalized sentiment scores.

**Conclusion**

The most significant outcome from the post-processing analysis was the observation that the larger model possesses a broader vocabulary and was able to detect phenomenal as a positive sentiment word. However, this advantage is offset by its diminished capacity to infer intended meanings in the presence of censored text, typos, and grammatically incorrect sentences, which then become problematic inputs for the model. This suggests a trade-off between a model's lexical knowledge and its ability to handle fewer standard forms of input effectively. Another way of looking at this is saying the model loses its ability to reason outside it’s training data if the training data did not include some noise.  
  
Don’t overfit your models. Give them some noise.

A picture containing text, indoor, wall, floor

Description automatically generated

**Addendum**

**Fig1: Small BERT-Base Model**

Text

Description automatically generated

**Fig2: Large BERT**

Text

Description automatically generated

**Fig3: Scores**

**Additional Hidden Layers in the Classifier**

Text

Description automatically generated

**References**

**"Deep Learning" by Ian Goodfellow, Yoshua Bengio, and Aaron Courville**

**"Attention Is All You Need" by Vaswani et al.**