**Project Phase 2 – Finalized Topic**

**Group Members: Michael Liu, Genfu Liu**

**Topic: Natural Language Processing**

**Track 2:** Research on a paper with code

**Description:**

Natural language processing (NLP) has seen a tremendous advancement in recent years. With newer and better models such as GPT 4, Bard, and many others just coming out in the recent month, I think it would be beneficial for us to have a deeper understanding of this field.

In this project, we will explore the capabilities of Natural Language Processing models, specifically transformers, by fine-tuning the pretrained BERT model developed by Google Research, we aim to fine-tune and test the model based on different novel datasets provided by different databases, for instance the IMDB movie reviews dataset, wikitext, and more. Moreover, we will also suggest and test improvements to the approach used in the original BERT research paper.

**Methodology and Plan:**

Here are the steps in order which we will be creating out fine-tuned BERT model;

1. **Data Preparation**

* In this step we will be downloading, preprocessing and preparing the datasets that will be used to fine-tune the BERT model. We will preprocess the data so it matches the inputs of the BERT model and split the data into training, validation, and testing sets to prepare for fine-tuning.

1. **Fine-tuning the BERT Model**

Here are the general steps we will be taking to fine-tune the model:

* 1. Downloading the BERT model
  2. Decide the different hyperparameters that we will be using to optimize the model and modify the original model to fit our datasets
  3. Training the model
  4. Validate the model

1. **Evaluating and Testing the Model**

* Lastly, we will be testing out model and evaluating the performance of our model and also compare the results with other models to see if there are any improvements that can be made.

After creating the model, we will create the 10-15 minute video, slides for the presentation and also the research paper presenting our model and the work we have done.

* 10-15 minute video – In this presentation video, we will most likely demonstrate the training process of the model as well as the results of the final model.
* Slides – In this presentation we will also demonstrate the training process and also explaining how the fine-tuning was done as well as showing the results of the model.
* Paper format:
  + Abstract
  + Introduction: problem definition, summary of contribution
  + Approach, Challenges and/or improvements
  + Results
  + Conclusions

**Resource List:**

* **BERT** - Deep Bidirectional Transformers by Google Research:

Introduction: <https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html>

Paper: <https://arxiv.org/abs/1810.04805>

Code: <https://colab.research.google.com/github/tensorflow/tpu/blob/master/tools/colab/bert_finetuning_with_cloud_tpus.ipynb>

Pre-trained BERT model on huggingface:

https://huggingface.co/docs/transformers/model\_doc/bert

* **Datasets**

**IMDB movie reviews** - A dataset of 50,000 movie reviews from IMDB, labeled as positive or negative: <https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews>

**WikiText** – a dataset that consists of a subset of articles from Wikipedia, there are a few different version of this but they all contain millions of words from thousands of articles: <https://huggingface.co/datasets/wikitext>

**And any other dataset that we might come across along the way**

* **Others**

We will be using PyTorch on Google Colab to create the model.