# BUAN 6342.S01 Applied Natural Language Processing

HW -7 Final Report

**SAI MANOJ REDDY VANCHA**

**SXV220035**

### This paper outlines methods for fine-tuning large-scale NLP models using parameter-efficient approaches such as LoRA (Low Rank Adaptation) and QLoRA (Quantized Low Rank Adaptation). These techniques aim to improve the performance of the models on named entity recognition (NER), sentiment analysis, and emotion identification tasks by lowering the number of parameters and computational load. For sentiment analysis, we used the Kaggle emotion detection hugging face dataset. This dataset has binary labels for the following emotions: anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, and trust. Each row in the dataset corresponds to a tweet. Every tweet has a unique identifier thanks to the 'NewID' column. Columns in dataset

In deep learning, fine-tuning—which comes after pre-training—is often used, especially in Natural Language Processing (NLP). By utilizing the pre-trained representations, this method helps the model effectively adjust to new tasks. Furthermore, fine-tuning allows users to customize pre-trained models to various tasks and domains, providing flexibility.

**Taks 1: Use the `google/gemma-1.1-2b-it` model  (or a similar size latest model like phi-3)**

This model is part of the Hugging Face model hub and was made by Google.The model architecture most likely rests on the transformer design, which has been widely used for NLP tasks.

It might have settings ideal for text classification, named entity recognition, machine translation, and other uses.

**Part A: Fine-tune using LoRA.**

[**wandb.ai/vmanoj-vancha**](https://cometmail-my.sharepoint.com/personal/sxv220035_utdallas_edu/Documents/Documents/Spring%202024/NLP/assignment/wandb.ai/vmanoj-vancha)

A method for efficiently adjusting large pre-trained language models with few parameters is called Low-Rank Adaptation (LoRA). LoRA fine-tuning reduces the number of parameters while preserving performance, allowing pre-trained language models to be efficiently adapted to particular downstream tasks. On the evaluation dataset, the model, which was trained using Low-Rank Adaptation (LoRA), performed moderately well. The model achieved a macro F1 score of approximately 55.89 % and a micro F1 score of approximately 66.77%, correctly classifying approximately one-fifth of the instances with an accuracy of approximately 21.52%. Higher F1 scores indicate better performance, and these scores show how well the model captures overall and class-specific performance. The loss value of roughly 0.72 indicates how well the model reduces the difference between the values that were predicted and those that were observed.

**A screenshot of a computer

Description automatically generated**

**Part B: Employ a parameter-efficient fine-tuning method other than LoRA.**

[**wandb.ai/vmanoj-vancha**](https://cometmail-my.sharepoint.com/personal/sxv220035_utdallas_edu/Documents/Documents/Spring%202024/NLP/assignment/wandb.ai/vmanoj-vancha)

This model is part of the Hugging Face model hub and was made by Google.The model architecture most likely rests on the transformer design, which has been widely used for NLP tasks. On the evaluation dataset, the model's accuracy was roughly 18.67%, meaning that only roughly 18.67% of the instances were correctly classified. The macro and micro F1 scores, which indicate a moderate performance in capturing overall and class-specific performance, were roughly 54.26% and 65.401%, respectively. The evaluation dataset's loss value, which measures how well the model reduces differences A screenshot of a computer

Description automatically generatedbetween expected and actual values, was roughly 0.62322.

**Task 2: Fine tuning model from the MTEB benchmark**

[**https://huggingface.co/settings/tokens**](https://huggingface.co/settings/tokens)**and fine-tune it using QLoRA.**

[**wandb.ai/vmanoj-vancha**](https://cometmail-my.sharepoint.com/personal/sxv220035_utdallas_edu/Documents/Documents/Spring%202024/NLP/assignment/wandb.ai/vmanoj-vancha)

On the evaluation dataset, the adjusted model obtained an accuracy of roughly 19.19%, showing an improvement in classification performance over the baseline model. The results indicate a moderate improvement in both overall and class-specific performance. The macro F1 score was around 56.21%, while the micro F1 score was roughly 66.90%. On the other hand, the evaluation dataset's loss value—roughly 1.04—was rather high, which might point to some difficulties in reducing differences between expected and actual values. Overall, the adjusted model exhibits encouraging gains in F1 scores and accuracy, suggesting that it may be useful for the intended goal.

****

All three models perform mediocrely, but the model optimized with QLoRA performs better in terms of classification accuracy than the other two. On the other hand, the relatively high loss value suggests that more optimization could be carried out to enhance performance. When all is said and done, the comparison shows how parameter-efficient fine-tuning techniques like LoRA and QLoRA can effectively adapt language models that have already been trained to specific tasks without sacrificing or improving performance. With relatively low macro and micro F1 scores of roughly 56 % and 66%, respectively.