**\*Task was to use gemma-2b-it / gemma-7b-it but my GPU ram was insufficient so I used MT5 base\***

**Objective**

The main objective of this project was to fine-tune the T5 model for generating medical report impressions based on the given history and observations. Additionally, we conducted a text analysis to explore the top 100 word pairs based on embedding similarity and developed interactive visualizations for these word pairs.

**Findings**

* **Fine-tuning T5 Model:**

After preprocessing the dataset and tokenizing the text inputs and labels, the T5 model was fine-tuned using 300 medical reports.

The model was trained for 3 epochs with a learning rate of 2e-5, and batch sizes of 8 for both training and evaluation sets.

* **Perplexity**: The evaluation phase resulted in a perplexity score of ‘perplexity‘, which measures how well the model predicts the next word in a sequence. A lower perplexity indicates a better predictive ability.
* **ROUGE Score Calculation:** ROUGE is a standard metric for evaluating text summarization and generation models. The generated impressions from the model were evaluated against the true impressions from the dataset using the ROUGE metric. This provided insight into the overlap between the model-generated text and the reference impressions.
* **Text Analysis:**

Stopword Removal, Stemming, and Lemmatization:

We removed stopwords and applied stemming and lemmatization to normalize the text across all reports.

This was essential to focus the analysis on meaningful terms and avoid redundant or filler words.

* **Word Embedding and Similarity Calculation:**

The processed text was converted into embeddings using a pre-trained model.

We computed cosine similarity between word pairs and identified the top 100 most similar word pairs based on their embedding distances.

* **Interactive Visualization:**

Using Plotly, we developed an interactive scatter plot to visualize the top 100 word pairs based on their cosine similarity scores.

This visualization enables users to explore the relationships between words, helping to uncover semantic similarities within the medical reports.

**Challenges Encountered**

* **Empty Hypothesis in ROUGE Calculation:**

Initially, some prompts were empty or not generating any output, leading to an error during ROUGE score calculation. This was handled by adding a check to ensure that the generated text is non-empty before evaluating it.

* **Heatmap Visualization:**

We encountered a mismatch error in the dimensions of annotations and data when visualizing the similarity matrix using Seaborn. The issue was resolved by ensuring proper reshaping of both the similarity scores and their corresponding word pairs.

* **Handling Large Text Input:**

Some text inputs in the reports were too long for the T5 model's tokenization limit, leading to truncation. We ensured that inputs were truncated appropriately without losing essential context.

* **Computational Resources:**

Fine-tuning large models like T5 requires substantial computational resources, especially with large datasets. Utilizing GPUs was critical in speeding up the training and evaluation phases.

**Potential Areas for Further Improvement**

* **Fine-tuning on a Larger Dataset:**

The dataset used for fine-tuning consisted of only 300 reports. Fine-tuning on a larger dataset might lead to better generalization and improved performance on unseen data.

* **Evaluation Metrics:**

While ROUGE and perplexity provide a good starting point for evaluating text generation models, incorporating metrics such as BLEU, METEOR, and human evaluation could provide a more comprehensive evaluation of the model's performance.

* **Model Hyperparameter Tuning:**

Further tuning of the model's hyperparameters, such as learning rate, number of training epochs, and batch size, could potentially improve the quality of the generated impressions.

* **Improving Text Processing Techniques:**

Instead of traditional stopword removal, stemming, and lemmatization, using more advanced techniques such as named entity recognition (NER) could help retain valuable information in the text analysis phase.

* **Interactive Visualizations:** The interactive scatter plot could be enhanced by integrating filters to allow users to explore specific types of word relationships (e.g., medical terminology, symptoms, or diagnoses).

**Conclusion**

This project demonstrated the effective use of T5 for generating medical report impressions, as well as the usefulness of text analysis techniques in understanding word relationships within medical reports. With further dataset expansion, evaluation, and refinement, this approach could be applied in clinical settings to assist in automating report generation and uncovering insights from medical text.