**Will Armstrong Practicum Documentation**

**V2: Fine Tuning Sentence Transformers without Model Training**

**Description:**

NOT retain the Bert model but understand vocabulary in the dataset that are not pretrained in the Bert models, no model training or classification, how do we add domain knowledge without model training.

To incorporate domain knowledge into the BERT embeddings without retraining the model, I followed these steps:

**Identify Domain-Specific Vocabulary:**

Extract unique words or phrases from your dataset that are not well-represented or understood by the pretrained BERT tokenizer. You can use the tokenizer to tokenize your text data and identify tokens that are split in unnatural ways, which suggests they are not in the tokenizer’s vocabulary.

**Create Embeddings for New Vocabulary:**

Use a method to create embeddings for the new vocabulary words. You can average the embeddings of subword tokens that BERT assigns to these words or use context-based embeddings from BERT to represent them. For example, pass sentences containing the domain-specific terms through the BERT model and average the output embeddings from the relevant layers.

**Enhance the Embeddings with Domain Knowledge:**

Use techniques like post-processing or fusion with external embeddings (e.g., using domain-specific word embeddings) to enrich the representation of these terms. You can combine BERT embeddings with domain-specific embeddings by concatenating or averaging them.

**Preprocessing:**

Ensure that your tokenizer is appropriately configured to handle domain-specific terms. If certain terms are critical, consider adding special tokens in the preprocessing pipeline.

**Evaluation and Use:**

Check if the enhanced embeddings improve performance on any downstream tasks or provide better semantic representation for analysis.