**DATA SCIENCE PROJECT:**

**INTRODUCTION**

The volumes of text data in the medical industry is growing. Research by data bricks shows that in 2015, HIMSS estimated that the healthcare industry in the U.S. produced 1.2 billion clinical documents. The PubMed database contains more than 34 million citations and abstracts of biomedical literature, presenting its own set of challenges in cataloguing, distribution, standardising, and qualifying of medical text data. As medical texts become more voluminous, it becomes important for information to be summarised for those who read it. Given the paucity of time and the growing volume of information, text summarization becomes an important tool to provide information to the consumers of this data, whether it be doctors, researchers, or even patients who seek readable summaries of their electronic health records.

The need for research on biomedical text summarization is two fold. (1) With the increasing volume and variety of data (Medical texts are usually accompanied by images), along with growing regulation, it becomes important to interpret and summarise data in a fast, secure, and error free way (2) The medical domain has its own distinct language models, jargon, vocabulary and abbreviations compared to other domains or general English models. In this context, text summarization tools can help researchers and clinicians save time and resources by automatically identifying and presenting the most important ideas within long documents.

*In this project, I study the text summarization capabilities of BERT variants such as BioBERT in the medical domain to understand their working. I aim to demonstrate how the pretrained Biobert model can be used to summarize scientific papers from the Pubmed database and compare that it outperforms the vanilla BERT model.*

**BERT VARIENTS**

Pretrained language models have achieved improved performance over the past few years. Xiaozhi Wang and Zhengyan Zhang have compiled and put together a list of pretrained models visually. One of the most significant developments in text summarization has been the development of BERT and other methods based on transformers. In their paper titled ”Attention is all you need” , the authors described the ”Transformer” as first transduction model relying entirely on self-attention to compute representations of its input and output without using sequence- aligned RNNs or convolution. They mentioned that in earlier sequence-sequence models, the computational intensity of the summarization task grows with the distance between the input and output positions, while in transformers this is reduced to a fixed number of operations. The authors of BERT argued that the limitation of the earlier models was that they are unidirectional, thereby limiting the choice of architectures used in pretraining.

**BERT Architecture**

BERT which is an improved model based on transformer, has helped bring SOTA accuracy to at least 11 natural language problems such as NER, question answering, translation, and summarization. BERT summarizes text in 2 steps like other models: Pretraining and fine tuning. In pretraining, the model is trained on unlabelled data from different pretraining tasks. In fine tuning, the model weights are initialized using the pretrained data and then the whole model is fine tuned using labelled data. This is shown in the image below of BERT Architecture.

Graphical user interface, diagram, application

Description automatically generated

In BERT every sentence starts with a special CLS token and ends with an SEP token. In addition, the model contains learned weights to distinguish between words from different sentences. BERT uses masked language modelling and next sentence prediction methods for pretraining, which allows it to be bidirectional, unlike earlier models. In masked language modelling, some 15% of the words are masked at random and the system is made to predict them, thereby balancing the trade off between masking data and knowing the full data during fine tuning, while in next sentence prediction, the model learns the long term relationship between sentences. The figure below shows the architecture of the BERT model.

A screenshot of a computer

Description automatically generated with low confidence

**Adaption of BioBERT**

Also, as argued by the authors of BioBERT, directly applying BERT embeddings to biomedical text summarization often produces unsatisfactory results due to a word distribution difference between general and biomedical corpora. The authors clearly called out that BioBERT basically has the same structure as BERT, and the only difference lies in pretraining. Unlike BERT which is trained only on general english corpora, Biobert is also trained on PubMed abstracts and full papers. The authors found that Biobert outperformed BERT.

**Training data collection and challenges**

For BERT models, fine tuning requires a training dataset with a large number of samples. We faced the following challenges while attempting to train and fine tune our model.

• Large size of the sample data making it difficult to download The training dataset from different repositories is of the tune of 8 GB and it was a challenge downloading the samples from there. Hence we directly pulled data from the huggingface repository.

• Summarization not perfectly working for each sample necessiating manual cleaning of data We are using supervise methods to fine tune the model, and hence every sample needs to be summarized. In our case, we found that the summarizer was giving errors for certain passages, and working for others. Hence we found the need to introduce error handling and ignore or correct the samples where the summarization had an index out of bounds error.

• Difficulty in generating BERT embeddings for the records, getting IOPub data rate exceeded. When we did download 100 samples for training and tried to get BERT embeddings directly, we were not able to get the embeddings to use input to the fine tuning model.

• Multidimensionality of bert embeddings We got the bert embeddings and tried to use and train a 2 layer sequential model, and we found the embedding to be a 4 dimensional vector. Each word has a bert embedding as a 768 length vector and hence the number of embeddings increases in size for large documents. Using space as a delimiter to tokenize the document produced a vector of nearly 7000\*768 for each document. The running time just to generate the embeddings for 10 documents was more than 25 min when running in google colab. The other challenge we faced was in how to pass a multidimensional vector to our sequential model, given that the embeddings are not entirely in the same format.

• The embedding depends on the delimiter that we use to tokenize the data The number of dimensions of the bert embeddings depends on whether we split the documents into tokens using a space or using a newline character. This made it very difficult for us to find the shape of the embedding vector to pass into the sequential model.

• Model losses mounting during fine tuning

**RESULTS**

One of the challenges of biomedical text summarization is the presence of medical jargon which is significantly different from other domains. While using medical datasets for pretraining is known to help with this, there has been some criticism about the usage of conventional Rouge scores to measure the effectiveness of medical text summarization tasks. In my code when BERT embedding and BIO BERT embedding are compared with a gold summary they generate cosine similarity scores of -0.42 and 0.84 respectively and when compared against each other generate a score of -0.51.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*