Natural Language Processing

Question Answering(Using BERT) for COVID-19

Atanu Guin (MT19AI002) **Vikanksh Nath (MT19AI024)**

Objective: We are required to build a question answering system for COVID-19 using pre trained Bidirectional Encoder Representations from Transformers. (BERT) model.

BERT: BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine- tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering,classification,inference etc.

BERT is pre-trained on a large corpus of unlabelled text including the entire Wikipedia(that's 2,500 million words!) and Book Corpus (800 million words). This pretraining step is really important for BERT's success. This is because as we train a model on a large text corpus, our model starts to pick up the deeper and intimate understandings of how the language works. This knowledge is the swiss army knife that is useful for almost any NLP task.

Pre-trained Bert Model:

We have used **BERT-Large, Uncased (Whole Word Masking)** which has 24-layers, 1024-hidden, 16-attention-heads, 340M parameters from the huggingface library .It is trained on Squad 2.0 dataset with whole word masking. The Stanford Question Answering Dataset(SQuAD) is a dataset for training and evaluation of the Question Answering task. SQuAD now has released two versions — v1 and v2. The main difference between the two datasets is that SQuAD v2 also considers samples where the questions have no answer in the given paragraph. When a model is trained on SQuAD v1, the model will return an answer even if one doesn't exist. To some extent you can use the probability of the answer to filter out unlikely answers but this doesn't always work. On the other hand, training on SQuAD v2 dataset is a challenging task requiring careful monitoring of precision and hyper parameter tuning.

Dataset for fine tuning:

To test our model, we compile a text using the given COVID-19 question-answering data set. In this way, we test how question answering works on articles that the model has never seen before. For preprocessing tasks, we have removed the hyperlinks from the texts and any additional punctuations which were present in the dataset.

We have merged all the answers for the COVID-19 question-answering dataset which was then used as the context for the Bert model during testing. The questions were used from the dataset itself. The model is predicting answers for the question to some extent.

	Questions	Answers
0	What is the incubation period of the corona vi...	The incubation period means the time between...
1	Who is most at risk for the corona virus disease?	People of all ages can be infected by the new ...
2	Is the corona virus disease the same as SARS?	No. The virus that causes COVID-19 and the one...
3	What is a corona virus?	corona viruses are a large family of viruses t...
4	Can humans become infected with a novel corona...	Detailed investigations found that SARS-CoV wa...
...	...	...
200	Is it safe to receive a package from any area ...	Yes. The likelihood of an infected person cont...
201	Do I need to wear a mask every time I step out...	Wearing masks does more to protect others agai...
202	What is the government doing to contain the sp...	The government is following the instructions o...
203	Where can I go to obtain accurate, scientific,...	The websites of the World Health Organization ...
204	How long is the incubation period for COVID-19?	The incubation period means the time between...

205 rows × 2 columns

Results:

We have few correct responses using this model. Here is the results:

```
Q1 : What is the incubation period of the corona virus disease?
100%|          | 1/1 [00:01<00:00,  1.51s/it]
Answer: Means the time between catching the virus and beginning to have symptoms of
Q2 : Who is most at risk for the corona virus disease?
100%|          | 1/1 [00:01<00:00,  1.25s/it]
Answer: . most estimates of the incubation period for covid
Q3 : Is the corona virus disease the same as SARS?
100%|          | 1/1 [00:01<00:00,  1.84s/it]
Answer: A large family of viruses that are known to cause illness ranging from the common cold to more severe disease.
Q4 : What is a corona virus?
100%|          | 1/1 [00:01<00:00,  1.27s/it]
Answer: A large family of viruses that are known to cause illness ranging
Q5 : Can humans become infected with a novel corona virus of animal source?
100%|          | 1/1 [00:01<00:00,  1.49s/it]
Answer: , many of the symptoms can be treated and therefore treatment based on the patients clinical
Q6 : What are the symptoms of someone infected with a corona virus?
100%|          | 1/1 [00:01<00:00,  1.54s/it]Answer: , many of the symptoms can be treated and therefore treatment ba.
```