CSCE 5290: Natural Language Processing
Project Proposal

Project Title: Text Summarization of COVID-19 Articles using various NLP methods.

Team Members:

1. Harsha Vardhana Buddana [11521994]
2. Roshan Sah [11574385]

Motivation:

Covid is a one of the kind new age viruses that shocked the world. We had very few resources that could tell us the sources, effects and handling this crisis. But our researchers were on it and started digging more about the treatment of this dangerous virus. As the researcher increases the covid related analysis published number also increases. But one cannot go through all these papers for his solution. Hence, we came up with the summarization model which summarizes the COVID-19 related articles which can answer the questions by summarizing the abstracts of the papers.

Objective:

- To gather articles that include text summarization terms that focus on the medical domain of Covid-19 words.
- May boost diagnosis and save a doctor's time during a saturated workload situation like the COVID-19 pandemic

Significance:

Not everybody has time to go through the complete research paper for their questions, Hence, this model would be a great tool for people who look for fast and effective solutions to their questions.

Also, under developing or developing countries doesn't have the resources like hospitals and doctors. Hence, this would be a great tool to know effective answers for their questions without any misinformation.

Features:

Approach:

Unsupervised system for comprehending scientific literature that accepts questions in natural language with a focus keyword and retrieves precise responses from the CORD19 corpus of scientific papers.

Dataset:

The COVID-19 Open Scientific Dataset was created in response to the COVID-19 epidemic by the White House and a consortium of top research organizations (CORD-19). The CORD-19 database contains more than a million research publications regarding COVID-19, SARS-CoV-2, and similar coronaviruses, including more than 400,000 full-text articles. With the help of current developments in natural language processing and other AI approaches, we may use this freely accessible information to provide fresh insights that will aid in the ongoing battle against this contagious illness. The CORD-19 dataset is the largest machine-readable coronavirus literature collection that is currently accessible for data mining. To help the continuing COVID-19 response activities globally, we have now given the chance to employ text and data mining algorithms to uncover answers inside and link insights throughout this information. We have taken only few attributes out of 19 attribute of the dataset for the text summarization. The attribute are: title, doi, abstract, publish_time, authors and url.

Pre-processing:

As part of cleaning the data, we drop Null value columns, duplicate titles, convert the text into lowercase and consider research papers from the year 2020. We create a data frame in which we try to hold the abstracts of the papers which contains terms related to the COVID. Removed the stop words and tokenized the text. Created a Data Frame such that, the data frame contains the abstracts of the paper which focuses on the given focus words. Later we use sentence similarity from the SpaCy library to calculate the similarity of the given sentence to that of the summarized answer.

Model:

We will use the NLTK and other NLP libraries and for the model we will use the BERT, GPT-2, RNN, Transformer, Seq2seq, GAN based Models. We will also use the cosine similarity for the performance measure of our model.

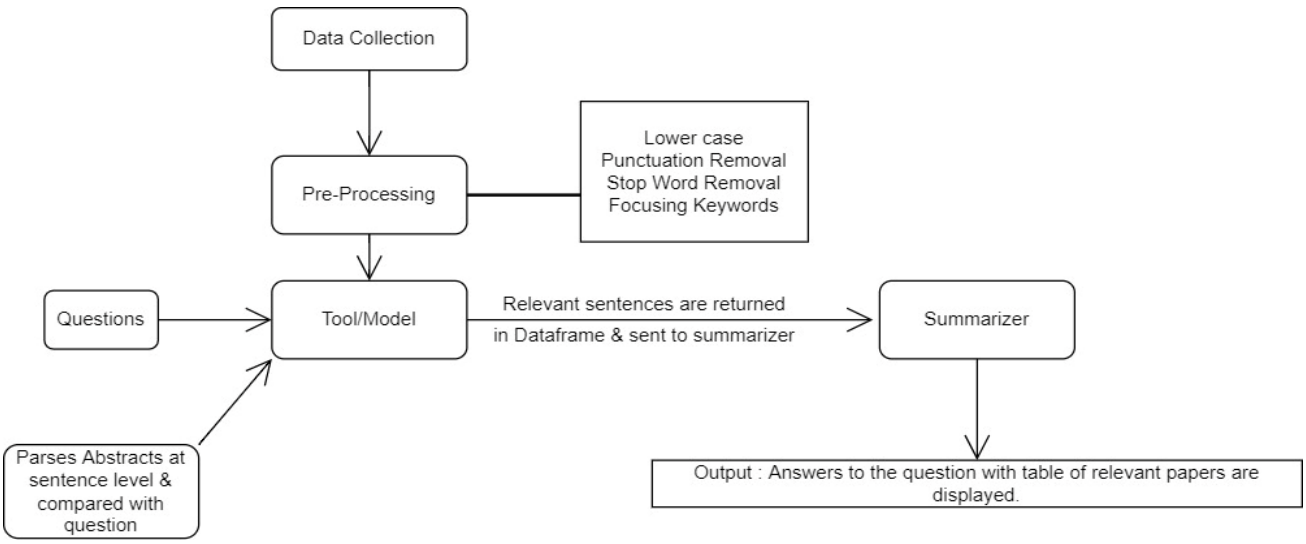

Figure: Purposed Flow-diagram of our Project.

References:

1. Guohui Song and Yongbing Wang. 2020. A Hybrid Model for Medical Paper Summarization Based on COVID-19 Open Research Dataset. https://doi.org/10.1145/3445815.3445824
2. Nemes, L.; Kiss, A. Information Extraction and Named Entity Recognition Supported Social Media Sentiment Analysis during the COVID-19 Pandemic. Appl. Sci. **2021**, 11, 11017. https://doi.org/10.3390/app112211017
3. Virapat Kieuvongngam, Bowen Tan & Yiming Niu (3 Jun 2020). Automatic Text Summarization of COVID-19 Medical Research Articles using BERT and GPT-2. https://doi.org/10.48550/arXiv.2006.01997
4. D. S, L. K. N and S. S, "Extractive Text Summarization for COVID-19 Medical Records," 2021 Innovations in Power and Advanced Computing Technologies (i-PACT), 2021. https://doi.org/10.1109/i-PACT52855.2021.9697019