# README

**(All code files are .ipynb files so to run them simply run the jupyter notebook)**

**DataCollection/SubredditDataCapture.ipynb**: Contains code for querying pushshift API in order to obtain posts in all our selected subreddits during the time period 15'Feb 2020 to 30'Sep 2020 having keywords 'pandemic', 'coronavirus', 'covid' , etc. Also contains code for outputting the results in a csv file. We get the list of all the unique authors who are found to be depressed because of covid here.

Libraries used:

- Pushshift
- Pandas
- Requests

**DataCollection/UserDataCapture.ipynb**: Contains code for extracting the entire history of the authors obtained from the previous file using PRAW API. The result was then combined in a JSON file using pandas and json python packages.

**DataCleaning/AnalyseData.ipynb:** Contains code for analysing collected posts from our selected users. Also has logic behind selecting users with more than or equal to 4 posts before COVID-19 i.e. 15 February, for our selected authors. We finally get the list of our selected authors who have sufficient amount of reddit activity.

**TrainingDataCollection/TrainingDataCollection.ipynb:** Contains code to collect data for the creation of our training Dataset. Here we have compiled a list of Non-Depressing Subreddits and depressing subreddits and annotated our collected data accordingly from the subreddit the data belongs

**DataCleaning/TrainingDataCleaning.ipynb:** Used for cleaning the training and testing data, removing the posts containing null, or less than 20 words, title and text combined as for classification title and text were concatenated.
Also there were some posts containing only emojis or not text in the post which were also removed. After cleaning there were 25141 non depressed posts and 34053 depressed data. The data was cleaned and converted to a json file.

**Classifier/Classifier.ipynb**: Contains the code to train various transformer based transfer learning models on the datasets listed above.

**Testing/Testing_batchWise.ipynb**: The code for testing the pre-Covid post collected of the selected users. The testing was done batchwise to predict whether a post is depressed or not individually.

**ResultAnalysis/Result_Analysis_Finding_depresses_users.ipynb**: Contains code to analyse the results found using the classifier on the pre-covid data obtained of selected users. We find the count of users who do not have any depressed post before COVID and other analytic information as well.

**ResultAnalysis/Graphs.ipynb**
Graphs.ipynb contains code written using python to generate graphs which graphically represent our analysis done on the data collected from Reddit.
Libraries used:

- matplotlib
- numpy

The data to plot the graphs was taken from
First graph is a pie chart to show the ratio of the number of people having posts showing signs of depression before & after COVID-19.
Second graph shows the ratio of posts before COVID-19.
Third graph shows the trend of the number of comments per month for r/depression.
The last graph shows the trend of the number of subscribers for r/depression subreddit.
Code can be executed by opening the python notebook using Jupyter notebook and running all the cells.

**ResultAnalysis/Result_Analysis_Finding_depresses_users.ipynb**: Contains code to analyse the results found using the classifier on the pre-covid data obtained of selected users. We find the count of users who do not have any depressed post before COVID and other analytic information as well.

**Links to Dataset (Uploaded links because of huge size)**

Unique Authors who were found to be depressed during COVID (99334 users)
https://drive.google.com/file/d/1C_D8Gc8nuQMSknfz8y9Quf_-D9ZptgU4/view?usp=sharing

Entire Reddit Activity of the authors found above
https://drive.google.com/file/d/1FLEFjmfu1MAGn967_PwuCWFZdc6x9PIg/view?usp=sharing

Authors with more than 4 post before COVID
https://drive.google.com/file/d/1BPeO0Fmakgs1bW3uINsK-uDcDsk90T_T/view?usp=sharing

Reddit Activity of the above selected users (31878 users)
https://drive.google.com/file/d/1bKrsf5LX0cwepk9XhQNZ46XoYdJSJtvE/view?usp=sharing

Reddit Activity of the above selected users (Cleaned)
https://drive.google.com/file/d/1MNWAgEFc3fiBWZpZBiemBzHzA8aJhtFs/view?usp=sharing

**Training Dataset**
Depressed Posts data
https://drive.google.com/file/d/1--EGw_EG2M0U7RrWvu0BNBYmg6Ou8VFj/view?usp=sharing

Non-Depressed Posts data
https://drive.google.com/file/d/1vWt84Lbhr1yPUn8mfmyI3itBm4_LgbTX/view?usp=sharing

Testing Results of 31878 selected users
https://drive.google.com/drive/folders/1M4ClBDOi_Hu2ew6MsbgFA9cS05cr5_c0?usp=sharing