* What were the goals and focuses of this project?

This project is a part f IBM hackathon. This project is done by in a group of 3 people (including myself) where 1 person was from psychology background and the other person was from computer science.

The goal of the project was to “Analyse and provide an insight to predict Mental Stress during the pandemic”. Here we made use of Reddit as our primary data source. We web scrapped the data from reddit from 2019 December to current date. Based on this, we made use of NLP to identify trends in our data.

Some facts about our curated dataset:

We crawled down 25024 unemployment subreddit members’ names. The users distribution is as below:

* 29 of them made the first unemployment post in January
* 40 of them made the first unemployment post in February
* 947 of them made the first unemployment post in March
* 5782 of them made the of them made the first unemployment post in April
* 7687 of them made the first unemployment post in May
* 6258 of them made the first unemployment post in June
* 4287 of them made the first unemployment post in July

We also crawled down 1156 work-from-home subreddit members' names. The users distribution is as below:

* 0 of them made the first WFH post in January and February
* 209 of them made the first WFH post in March
* 364 of them made the first WFH post in April
* 260 of them made the first WFH post in May
* 197 of them made the first WFH post in June
* 126 of them made the first WFH post in July

Due to the limited time for hackathon and time constraints, we just crawled 30k historical posts of 998 April-unemployed users and 138k historical posts of all the 1156 WFH users.

NLP along with other techniques- sentiment analysis, LIWC (Linguistic Inquiry and Word Count), Topic Modelling and Emotion analysis helped us we to generate key insights by comparing the posts made by the user after the COVID pandemic with the posts made by users before the COVID 19 situation. This led to very interesting results which shows us the significant trends and the key factors that might were the reason of rise in mental stress levels among the two classes of people – Unemployed and WFH (Work From Home) taken into consideration.

* Describe the technical details related to the project, such as: the input features; model architectures; algorithms; metrics; optimizers; performance evaluation; etc

In this project we made use of Transfer Learning on BERT model. As training BERT network is a huge task, we made use of pre-trained network. We just played with last few layers of the model to incorporate our own sentiment analysis use case into it.

Input features was the posted message by the user, which was further cleaned by removing stop words, special symbols, URLs etc. They were then tokenized and processed in our transfer learning model. The code was running for 3000 epochs with batch size=10 for our new model. The optimizer selected was Adam optimizer. Learning Rate=0.001(default value). All this was running on IBM cloud as we were having access to it. There were in total 12 BERT layers of which the last two layers were modified.

* What were some novel approaches that you employed while solving the problem?

Novelty in our project was the 1) Dataset that we curated and 2) Transfer learning on BERT which saved us a lot of time.

1) **Our source of data:** Stressed people usually try to divert their attention to something else. Here comes the social media which is one of the biggest helps that helps the depressed person to divert their attention or ask their problems. This is the place where a person can disclose their secrets (anonymously on Reddit) and share their thoughts and problems.

2) **What type of data:** We were mostly interested in data that was mainly capturing the depressed people on Reddit. So, we did our research and identified that depressed person is more likely to make a post at late night. Using this and along with LIWC and search filters we were able to collect posts of around 8000 unemployed people and around 400 unique WFH people.

What kinds of results did you produce?

* From the analysis and insights, we confidently concluded that both the class of people are having different reasons of stress inducing factors. The prolonged sadness, anxiety, work pressure, no job/money, the agitation among different groups of people, deaths, crimes etc are some of the factors that are playing a major role in controlling either both the classes (in some of the cases) or one of the either classes in inducing the stress level among the people.
* These data/insights can be used by the local health administrations, healthcare providers and even the government to focus on the issues concerning the general public affected by the pandemic. Using these insight data, the government or the healthcare provider can identify which state needs more resource investment to help keep the people more motivated and encouraged. This will also reduce down the agitations among the general people and will help in reducing down the death rates (maybe due to suicide) and also the crime rates significantly.
* Our model’s accuracy rates were high (84%). We were definitely able to see some trends but initially we were seeing some sudden bumps in our results/insights. Later on our investigation of our data we identified that because of in between protests like- BLM and other reasons like sad demise of Kobe, family loss due to covid etc were some additional regions we saw the bumps in our insights.

One of the insights that we got from the analysis

Important Key Insights:



Y-axis: i\_ppron\_prob = total number of 1st person pronoun word / total number of personal pronoun word

X-axis: day 0 is the day when a user made the first WFH/unemployment post

From the references, [1] and [2] we can see that depressed people tend to use more “I” keywords. From the above graphs, we see that unemployed people (left graph) are still very depressed even after the COVID-19 impact. There is no significant change seen here. On the other hand, for the WFH people (right graph) it got little better and it may be because of the family support and time the person gets staying and working from home

What would you change about this project?

* This work can further be extended to more root level like getting into districts or even to county wise.
* This work can further class consider the other classes of people which were not taken into account in this project due to time constraints.
* Also, if somehow, we are able to identify the location from where the post was made, we would definitely have a crisper level of level where we can evaluate our results county/city/state or even country wise.

***[NOTE] : I am really sorry but the GitHub repository for this project cannot be shared by me till mid Feb. We have qualified for the next round of Hackathon where we will improve and work on our future works. I hope you understand.***