Mental health analysis during COVID-19:
A comparison before and during the pandemic

Gurshaan Singh Bajaj	2017UCO1627
Harsh Yadav	2017UCO1597
Sudhanshu Sah	2017UCO1637
Harshdeep Singh Sahdev	2017UCO1640

INTRODUCTION

The coronavirus disease 2019 (COVID-19) pandemic has turned out to be stressful for people. Fear and anxiety about a new disease.

Stress during the pandemic caused fear and worry.

Frequent use of online social network (OSN) such as Twitter, Facebook and Reddit, to express opinions and emotions.

Use of Reddit data to examine and analyse the rise of variety of mental health conditions during COVID-19.

WHY REDDIT?

- Reddit communities are referred to as subreddits.
- Each subreddit is devoted to a specific topic.
- Communities promote sharing recovery experiences and emotional support.
- Users stick to the topic or risk banning of their posts.
- Easy availability of historical data.

SUBREDDITS

The following subreddits have been chosen for our project.

r/ADHD

r/Anxiety

r/Autism

r/BrosOnToes

r/Neurodiversity

r/PTSD

r/SocialSkills

r/SuicideWatch

r/Depression

r/Stress

r/Offmychest

RELATED WORKS

Binary Classification to detect whether a user suffers from one of nine mental health disorders using ‘Hierarchical Attention Network’, and using SHMD Dataset. - Adapting Deep Learning Methods for Mental Health Prediction on Social Media

A large scale dataset WU3D (Weibo User Depression Detection Dataset) was published. Some statistical features were also determined for better classification. - A Multitask Deep Learning Approach For User Depression Detection On Sina Weibo

Detecting topics of interest for people suffering from Anorexia Nervosa who are reddit users and declared themselves as the patient of AN. Provided a dataset of annotated posts labelling the relevant topics for the post. - Anorexia Topical Trends in Self-declared Reddit Users

RELATED WORKS (Cont'd)

Investigate the impact of using emotional patterns identified by the clinical practitioners and computational linguists to enhance the prediction capabilities. - Multi-Task, Multi-Channel, Multi-Input Learning for Mental Illness Detection using Social Media Text

They address the challenge of identifying depressed users of the Reddit social platform. Use combination of a tf-idf weighting scheme for bag of words features and a rule-based estimator, that takes into account several psycholinguistic features that characterize depressed users. - Understanding Depression from Psycholinguistic Patterns in Social Media Texts

Methodology

- Data Collection and labelling
- Model Training and predicting
- Visual Depiction of Results

DATA COLLECTION: User Extraction

1. Filter out the depressed posts which were mentioning COVID using keywords like ‘covid’, ‘coronavirus’, ‘covid19’, ’pandemic’ etc. using API ‘PUSHSHIFT’ .

   (The duration for the extracted posts was taken from Feb 15, 2020 - Sep 30, 2020)

2. We extracted the users who authored those posts and created a set of those unique users.

Now, we have all the users who are depressed and are affected by the pandemic.

DATA COLLECTION: User Extraction (Cont'd)

3. Select only those with sufficient Reddit activity to get an insight.

4. Identified the users which have some posts before our prescribed pandemic timeline

5. Calculated the number of posts each user had before 15 February 2020, and decided a threshold the minimum number of posts a user must have before our timeline.

After speculation ,**we decided to put the threshold at 4**.

Consequently there were nearly **32000 users having 4 or more posts pre-covid**.

Now we have our set of relevant users.

DATA COLLECTION: Final Steps

Extracted all the posts on Reddit using the API ‘PRAW’.

For 32000 users we extracted around 2.9 million posts.

All these posts were later passed through a classifier to get some insight into the mental conditions of their authors before the pandemic.

Data Visualisation

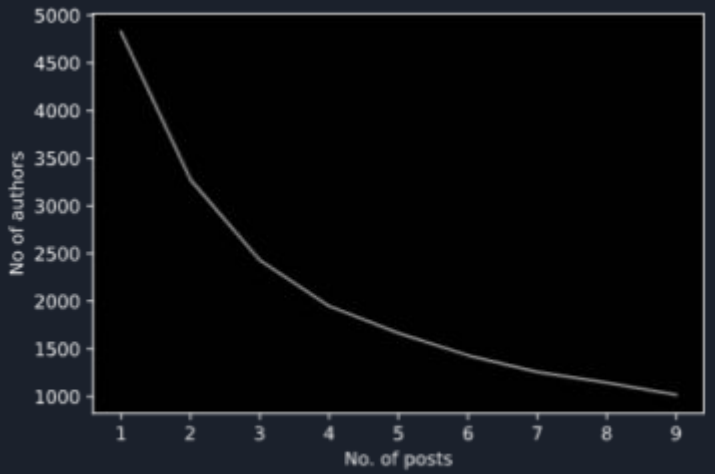


Figure 1 (a): Plot of number of authors with given number of posts

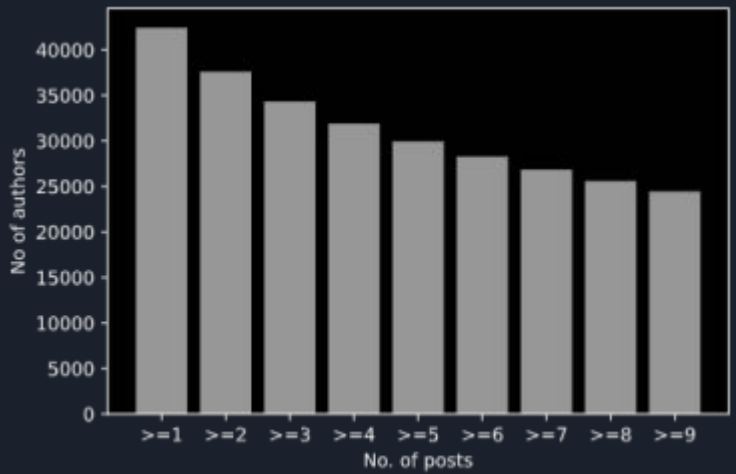


Figure 1(b): Plot of number of authors with greater than or equal to a given number of posts

CLASSIFICATION: Transformer vs RNN/LSTM

RNN/LSTM limitations:

- Dealing with long-range dependencies is still challenging.
- The sequential nature of the model architecture prevents parallelization.

Why Transformer?

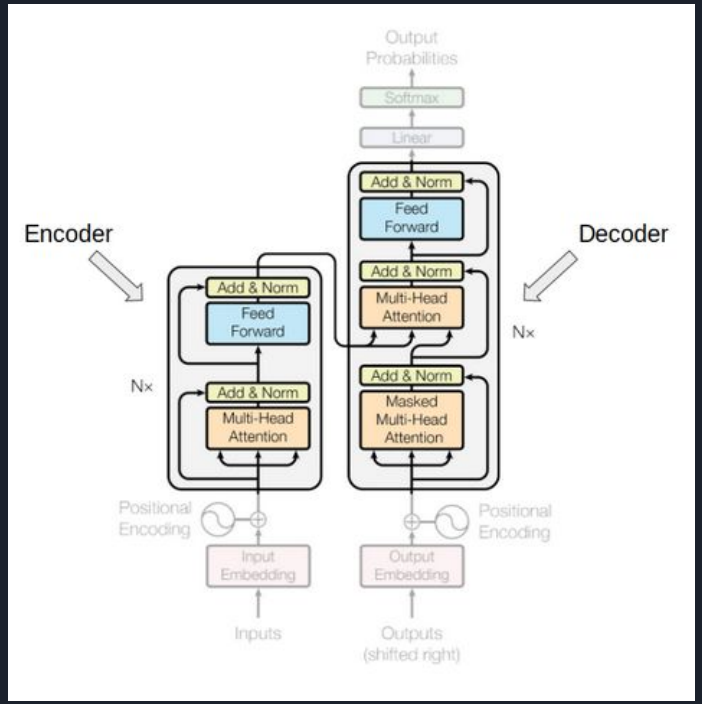
Transformers solve sequence-to-sequence tasks while handling long-range dependencies with ease.

Transformers avoid recursion, by processing sentences as a whole and by learning relationships between words, thanks to multi-head attention mechanisms and positional embeddings.

TRANSFORMERS

The main characteristics of transformers are:

- **Non sequential**: sentences are processed as a whole rather than word by word.
- **Self Attention**: this is the newly introduced 'unit' used to compute similarity scores between words in a sentence.
- **Positional embeddings**: another innovation introduced to replace recurrence. The idea is to use fixed or learned weights which encode information related to a specific position of a token in a sentence.



TRANSFER LEARNING: BERT

BERT - **B**idirectional **E**ncoder **R**epresentations from **T**ransformers.

BERT's model architecture - multi-layer bidirectional Transformer encoder.

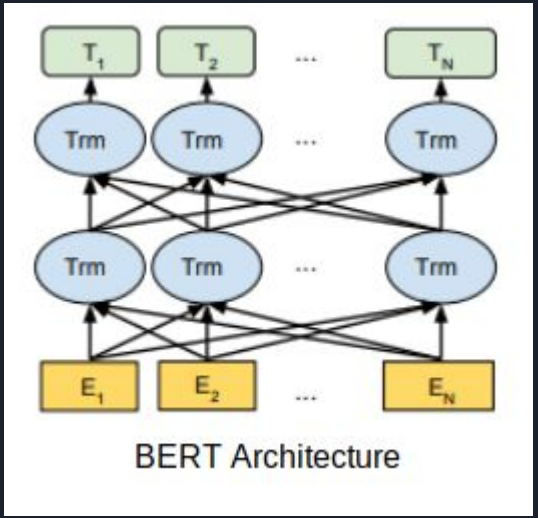
BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. BERT uses "masked language model" (MLM).

BERT outperforms all systems on all GLUE tasks by a substantial margin with BERT-base achieving an average 4.5% accuracy improvements and BERT-large, 7.8% improvements over previous *state of the arts*.

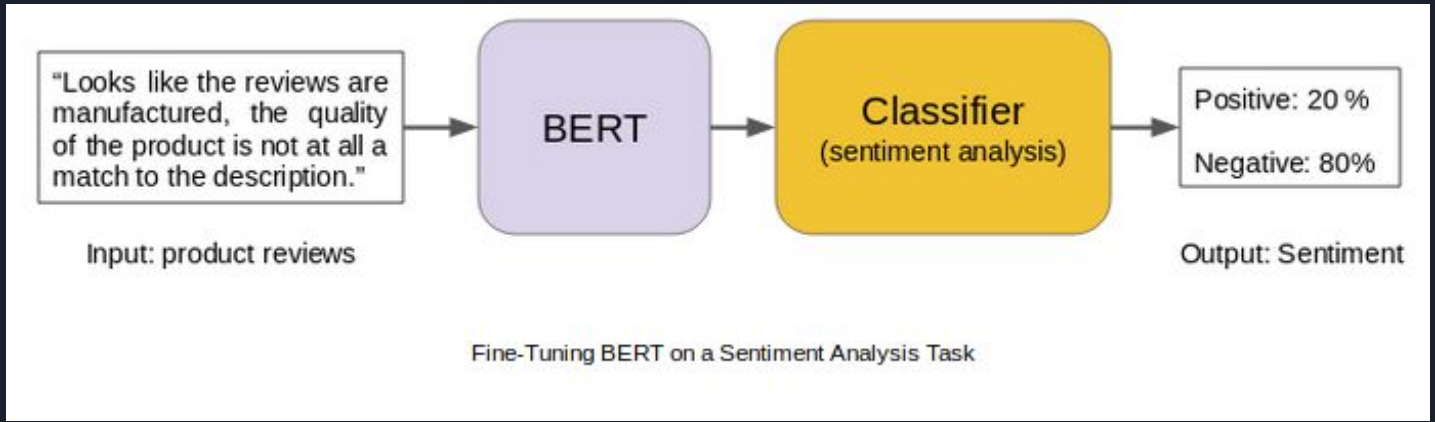
BERT: Architecture

BERT Base: Number of Transformers layers = 12, Total Parameters = 110M

BERT Large: Number of Transformers layers = 24, Total Parameters = 340M



BERT Architecture



Fine-Tuning BERT on a Sentiment Analysis Task

Transformer based Transfer Learning

The **BERT** framework is one such example, which is a transfer leaning transformer based model from Google AI.

Other transfer learnings and transformer-based frameworks that are available are **XLnet, Longformer, DistilBERT, and Roberta**

CLASSIFIER: Training and Testing Data-set

To train our classifier we needed an annotated data-set containing two field:

1. The text
2. A label - depressed or not.

Depressed Posts - Post from the subreddits listed in earlier slides

Non-depressed posts - Posts from many different subreddits like r/Thoughts, r/ReviewThis, r/culture, r/positivepsychology, r/love, r/travel and r/MovieSuggestion

The final data-set had **50,242 sample** containing equal number of post from the two labels.

CLASSIFIER: Testing Results

Model Name	Accuracy	Macro-AVG	Weighted-AVG	Label “Depressed”			Label “Non-depressed”		
				Precision	Recall	F1-score	Precision	Recall	F1-score
BERT	0.96	0.96	0.96	0.95	0.95	0.95	0.95	0.95	0.95
ROBERTa	**0.97**	**0.97**	**0.97**	**0.96**	**0.97**	**0.97**	**0.97**	**0.96**	**0.96**
XLnet	0.94	0.94	0.94	0.95	0.94	0.94	0.94	0.95	0.94
Longformer	0.96	0.96	0.96	0.95	0.96	0.96	0.96	0.95	0.96
DistilBERT	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95

Testing and Obtaining Results

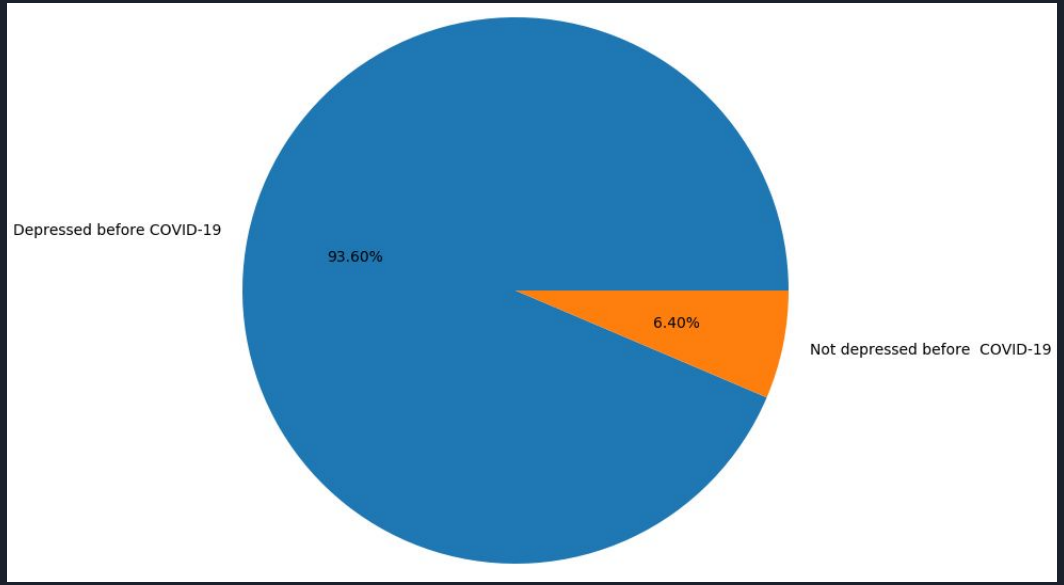
We chose Roberta because it had the best accuracy on our depression classification dataset.

All the shortlisted users' precovid posts were tested using the Roberta classifier and the results are as follows

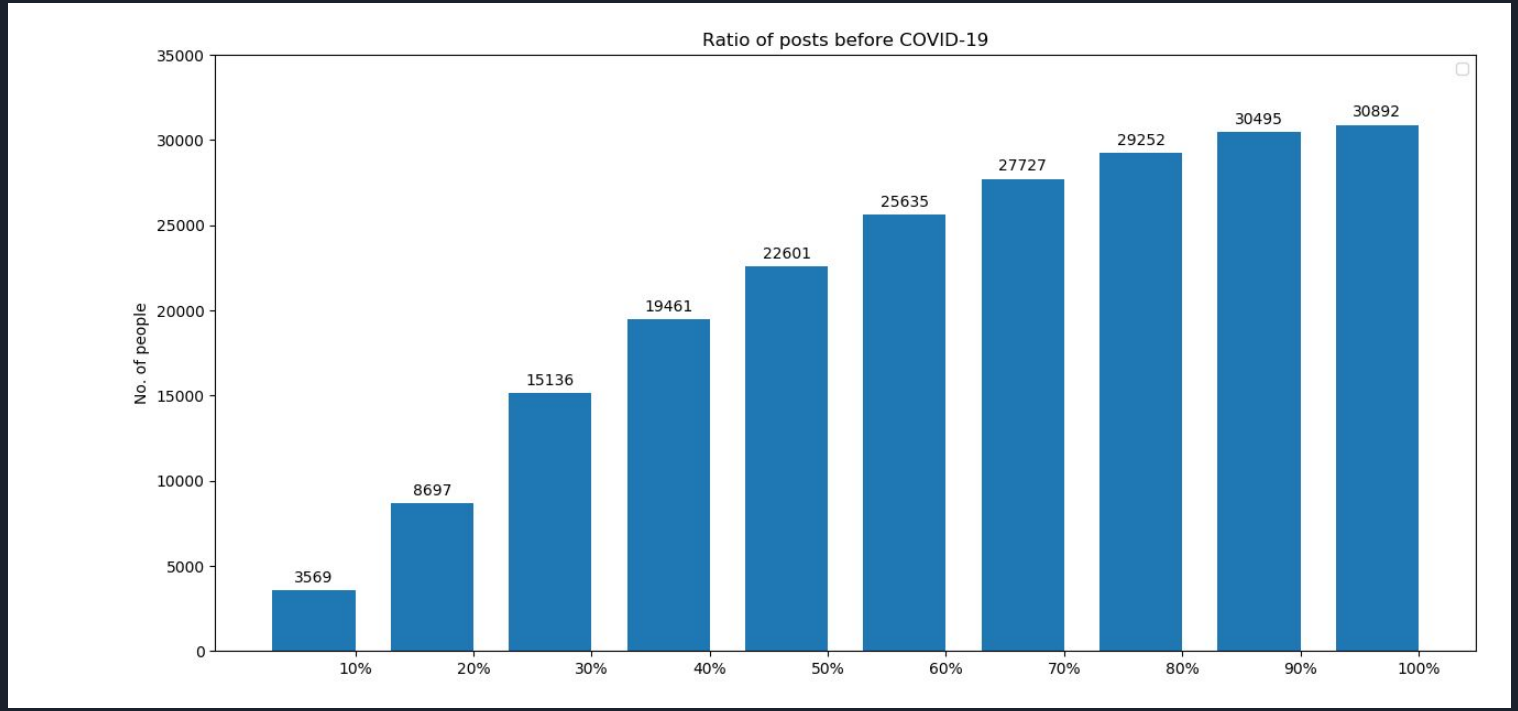
RESULTS

Out of 30892 total users, 1987 users were found having no depression classified post before the pandemic.

6.4% of our total users were not depressed before COVID-19.

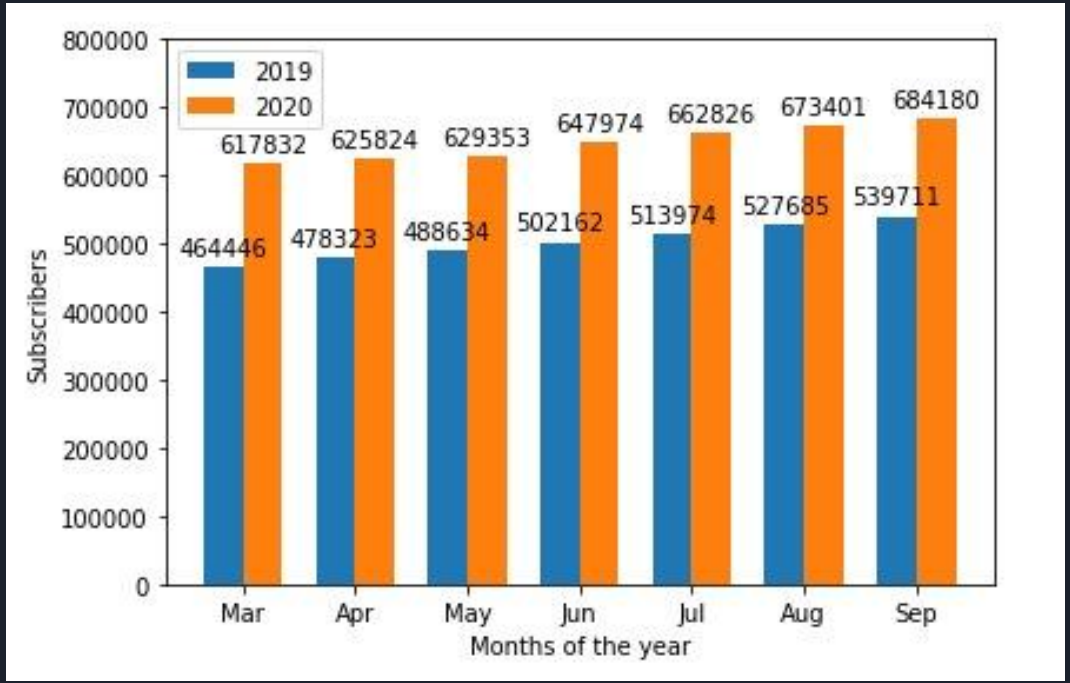


RESULTS



NUMBERS OF USERS vs PERCENTAGE OF DEPRESSED POST FOR A USER BEFORE COVID

RESULTS

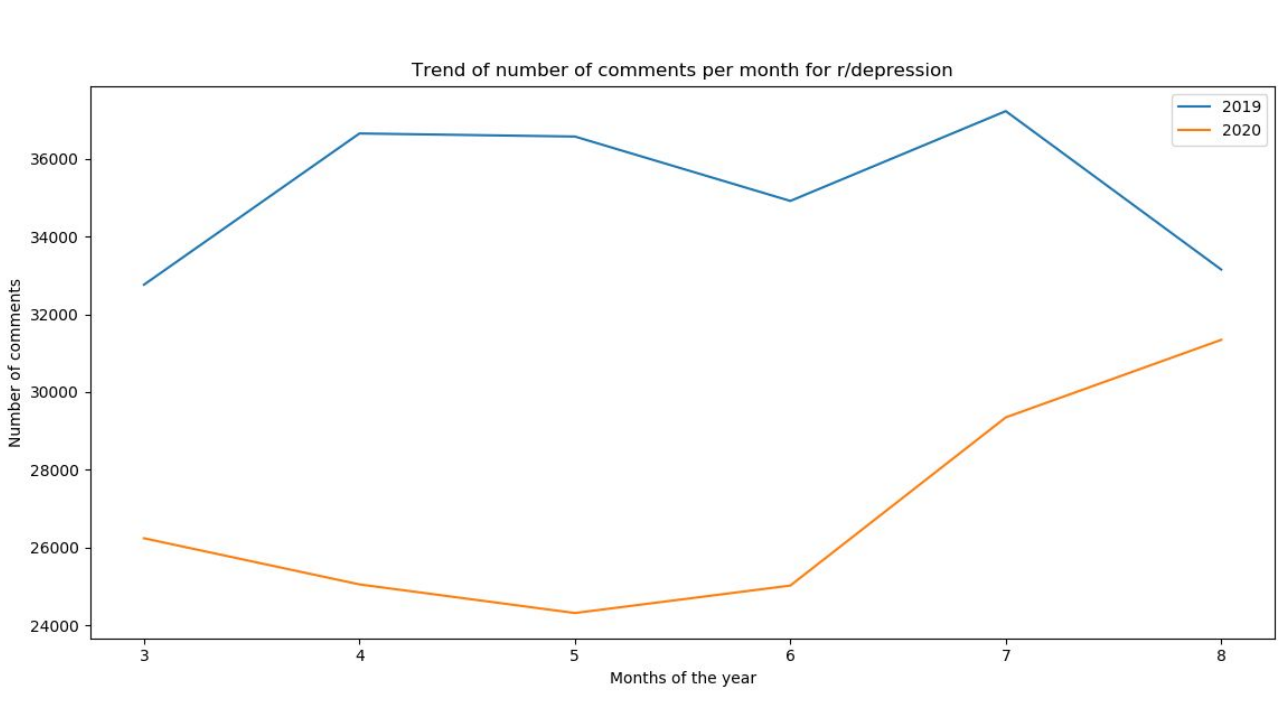


This bar graph shows the number of subscriptions for each month between March and September for both 2019 and 2020.

Note: The graph shows the total subscribers during the particular month.

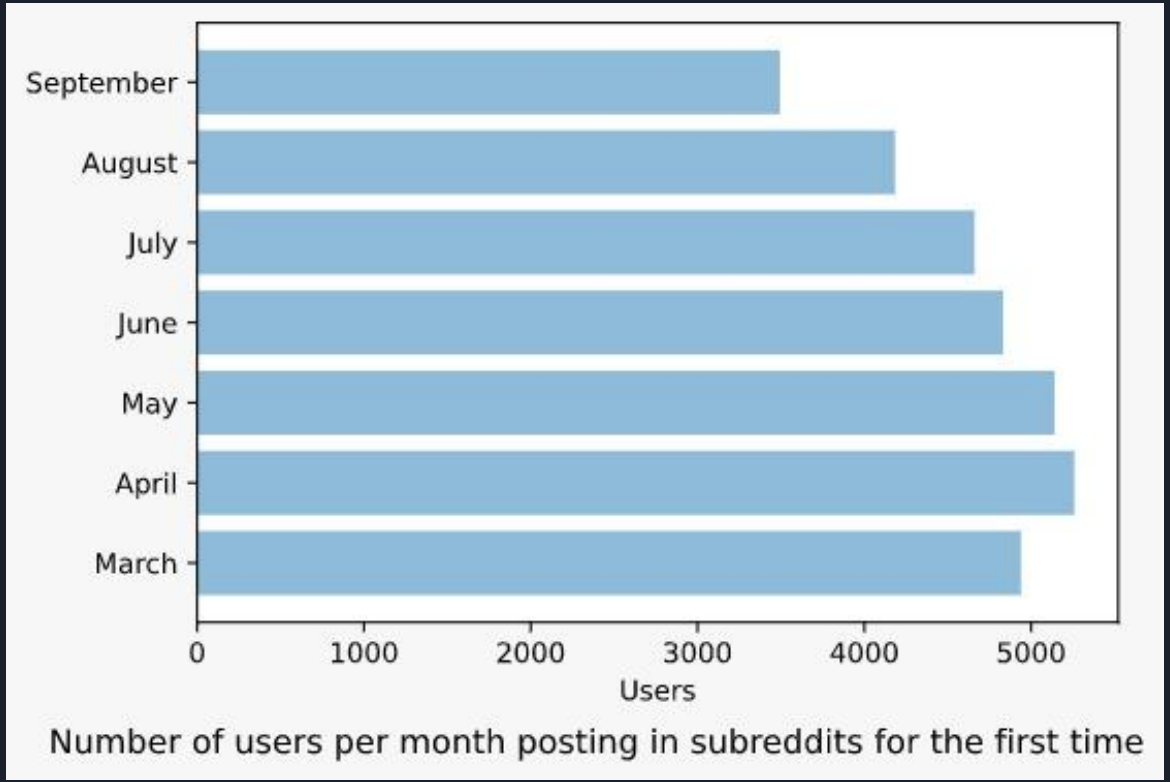
The graph shows that the subscriber growth in 2019 was much more than 2020 during the shown interval.

RESULTS



This graphs also shows a similar trend as the previous graph and also shows that the user engagement on the subreddit r/depression has been much higher during 2019 rather than during the pandemic.

RESULTS



Number of users per month posting in subreddits for the first time

This graph shows when during the pandemic the users who were impacted by COVID-19 posted for the first time on Reddit.

Though the distribution is seen to be much dominant during the quarter of April, May and June.

CONCLUSION

- 6.4% users showed normal behaviour before the pandemic but are now depressed.
- Also 27727 users have more than 70% of their activity during COVID-19.
- Till the month of April, the number of users who posted for the first time increased.
- Towards September, the number started to fall which shows that people started to adapt to this pandemic.
- Used Transformer based model to get better accuracy over RNN-LSTM based models.

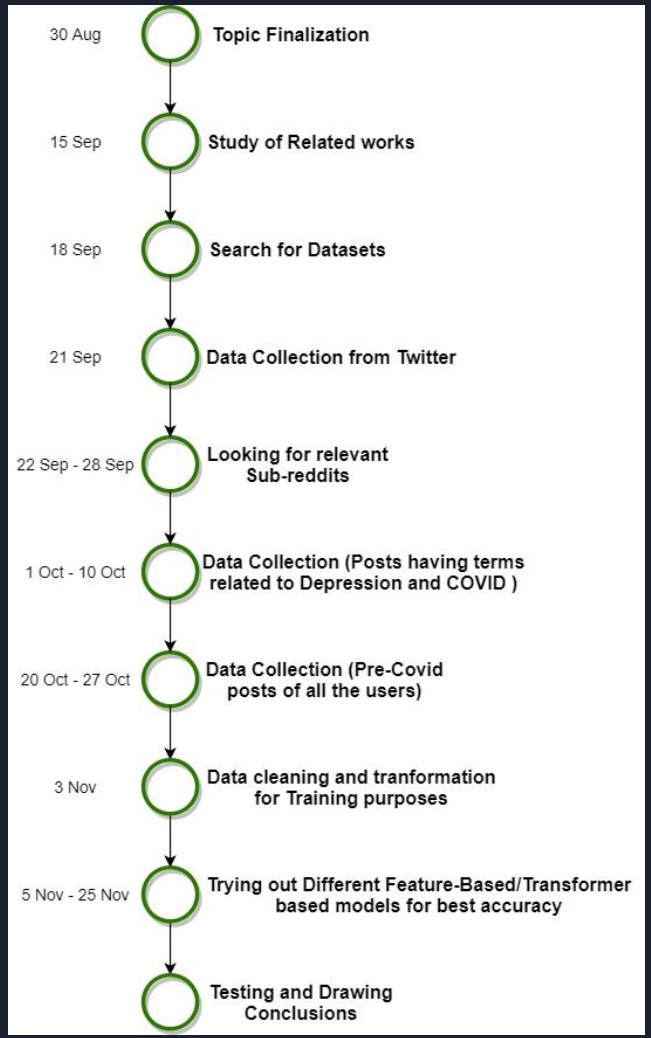
LIMITATIONS AND CHALLENGES

- The data collection relied heavily on the rules of posting in subreddits.
- Only limited personal information was available.
- Non-availability of GPU and short runtimes available on online servers.
- Due to lack of previous datasets available, we couldn’t get further insights about the major symptoms of depression shown by the people during the pandemic like anxiety, insomnia etc.
- Reddit doesn't provide API to get posts/comments data for a time frame directly. We had to use third-party APIs to make multiple API calls for fragments of time.
- Reddit doesn’t allow us to get a particular user's activity in a specific subreddit, we have to collect the entire history and filter it accordingly.

FUTURE WORK

- We will gain more insights about the major symptoms using textual patterns.
- Find the reason behind falling subscription gains and activity of the users during COVID-19 which ideally should have increased.
- Compare the condition and severity of people with different mental disorders and find out, how they have been affected by COVID.
- Incorporate demographic data into our dataset.

Our Timeline



Mental health analysis during COVID

ORIGINALITY REPORT

%6	%5	%2	%4
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Netaji Subhas Institute of Technology Student Paper	%1
2	www.analyticsvidhya.com Internet Source	%1
3	github.com Internet Source	%1
4	fr.scribd.com Internet Source	%1
5	www.kdnuggets.com Internet Source	<%1
6	"Natural Language Processing and Chinese Computing", Springer Science and Business Media LLC, 2019 Publication	<%1
7	studentsrepo.um.edu.my Internet Source	<%1
8	Submitted to East Carolina University Student Paper	<%1

Originality Report

Project completed under the guidance and assistance of Prof. Preeti Kaur

Gurshaan Singh Bajaj	2017UCO1627
Harsh Yadav	2017UCO1597
Sudhanshu Sah	2017UCO1637
Harshdeep Singh Sahdev	2017UCO1640