Defending Against A Pandemic - The Fight for Survival

Sanskar Sehgal, Akhilesh Boppana, Nikhil Deshakulkarni, Vivek Golani

1. Introduction :

The COVID-19 pandemic has had a significant impact on global public health, with many people experiencing increased levels of stress and anxiety, social isolation, and economic hardship. As a result, there has been a growing concern about the potential increase in substance abuse and mental health issues during the pandemic. To address these concerns, our project aims to provide evidence of a surge in substance abuse during the pandemic and quantify its impact on mental health. We also strive to develop a forecasting model to predict hospital bed and staffing deficits with a one-week lead time, which could help enhance preparedness for future pandemics.

The United Nations Sustainable Development Goal 3 aims to ensure healthy lives and promote well-being for all, and our project on COVID-19 is aligned with this goal. We aim to provide evidence of a surge in narcotic drug abuse and harmful use of alcohol during the pandemic and show how the UN goal of reducing premature mortality from non-communicable diseases was violated. By forecasting hospital bed and staffing deficits with a one-week lead time, we aim to enhance preparedness for future pandemics, which is critical for achieving universal health coverage, including financial risk protection and access to affordable essential medicines and vaccines for all. Through our work, we hope to provide policymakers and healthcare professionals with the tools they need to prepare for future pandemics and minimize their impact on society. Ultimately, our goal is to ensure that everyone can lead healthy and fulfilling lives, regardless of external circumstances beyond their control.

2. Background :

Our project is built on the idea of assessing the impact of the COVID-19 pandemic on public health and is aligned with the United Nations Sustainable Development Goal 3, which aims to ensure healthy lives and promote well-being for all. Through our analysis of the correlation of the rise in substance abuse with the onset of the pandemic, we show that there was a significant impact on the mental health of people. We also provide a model to strategize bed allocations in hospitals to fight against the pandemic. To understand the application of our project, one might need some knowledge of public health issues, the impacts of pandemics on public health, and the United Nations Sustainable Development Goals.

Methodologies	Frameworks	Models
Sentiment Analysis: Classified posts into severity classes.	Pyspark, Pytorch	XLNet(Transformer)
Frequency Analysis: Quantified posts based on severity	Pytorch	
Pearson's correlation: covid cases and severity scores	Pandas	
Time Series: forecasted normal and ICU bed availability	sklearn, stat_models	ARIMA

Understanding the Sustainable Development Goals (SDGs) related to healthcare, particularly SDG 3, which aims to ensure healthy lives and promote well-being for all, is crucial to understanding the significance of our project. By applying time series analysis to COVID-19 hospital bed occupancy data, we hope to contribute to the larger goal of improving healthcare outcomes during pandemics. Similarly, our sentiment analysis on the correlation between substance abuse and COVID-19 cases can provide insights into ways to address the pandemic and mitigate its impact on mental health, which is an important aspect of SDG 3.

3. Data :

Dataset for sentiment analysis: Reddit Covid Dataset and Weekly Prevelance of Covid in the US

Size:

Covid Dataset - 5MB

Reddit Dataset - 14.62GB (Kaggle) with 17 million rows and 11 columns + 3.02 GB (scrapped) contains 13050 rows and 11 columns

Descriptive Statistics: Reddit dataset provides information about the posts posted by users with timestamps, along with the subreddit posted in. Columns include metadata such as engagement scores for that post. The Covid data is used for correlating severity of posts to prevalence of covid cases.

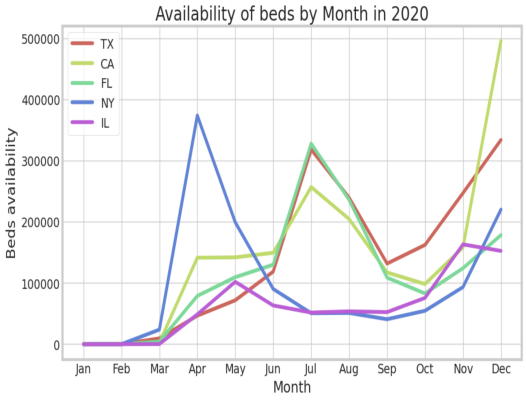
Labels: Normalized engagement score for each Reddit post

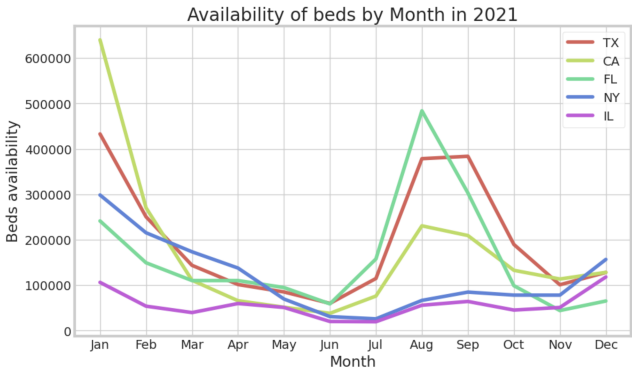
Dataset for time series analysis: COVID-19 Reported Patient Impact and Hospital Capacity

Size: 50MB containing 135 columns and 75000+ records.

Descriptive Statistics: It provides information on hospital utilization in different US states from Jan 2020-May 2023. Columns include critical staffing shortages, fully staffed pediatric ICU beds, etc.

Labels: Unlabeled dataset to forecast hospital infrastructure shortages.





4. Methodology:

4.1 Sentiment Analysis :

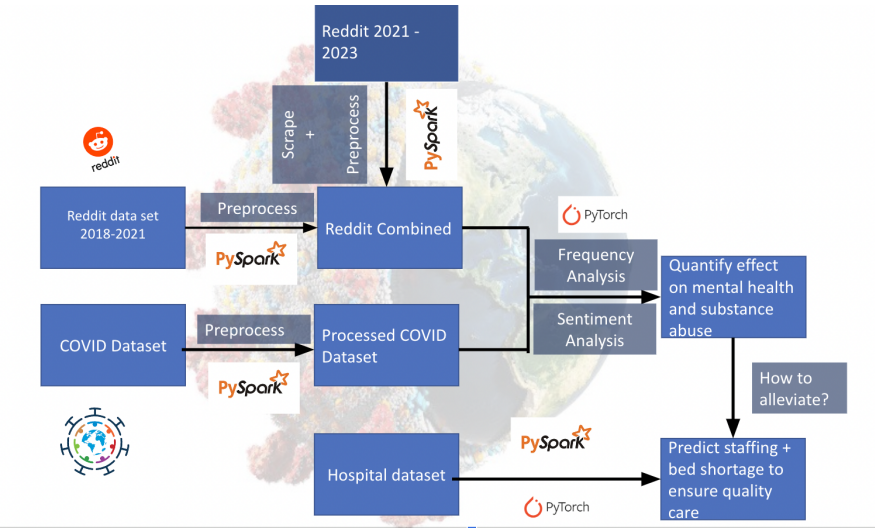
To gather our data for analysis, we utilized both existing COVID-19 Reddit datasets from 2018 to 2021 and also scraped data from 2021 to 2023 using Reddit's open-source APIs. The data was loaded into Spark RDDs and combined, and we performed preprocessing by filtering out posts not related to mental health or substance abuse. We also used engagement scores to filter the data for better results. The

XLNet model, a transformers model, was used to classify the posts into five severity score classes: critical, major, moderate, minor, and cosmetic, through a zero-shot classification pipeline.

We evaluated models like - AIMH/mental-bert-large-cased, AIMH/mental-longformer-base-4096, and AIMH/mental-roberta-large, on our dataset to determine the top-performing one. Our analysis revealed that the AIMH/mental-xlnet-base-cased model yielded the most precise outcomes. We further try to provide insights from the data by calculating the correlation with the number of covid cases as shown in section 5.

4.2 Frequency Analysis :

We analyze the frequency of certain keywords from subreddits related to mental health and substance abuse before and during covid. The keywords included - ‘isolation', 'anxiety', 'depression', 'stress', 'loneliness', 'fear', 'uncertainty', 'unemployment', 'economic', 'financial', 'recession', 'alcohol', 'beer', 'wine', 'liquor', 'whiskey', 'vodka', 'rum', 'tequila', 'gin', 'drugs', 'weed', 'marijuana', 'cannabis', 'cocaine', 'heroin', 'meth', 'crack', 'addiction', 'recover', 'sobriety'.



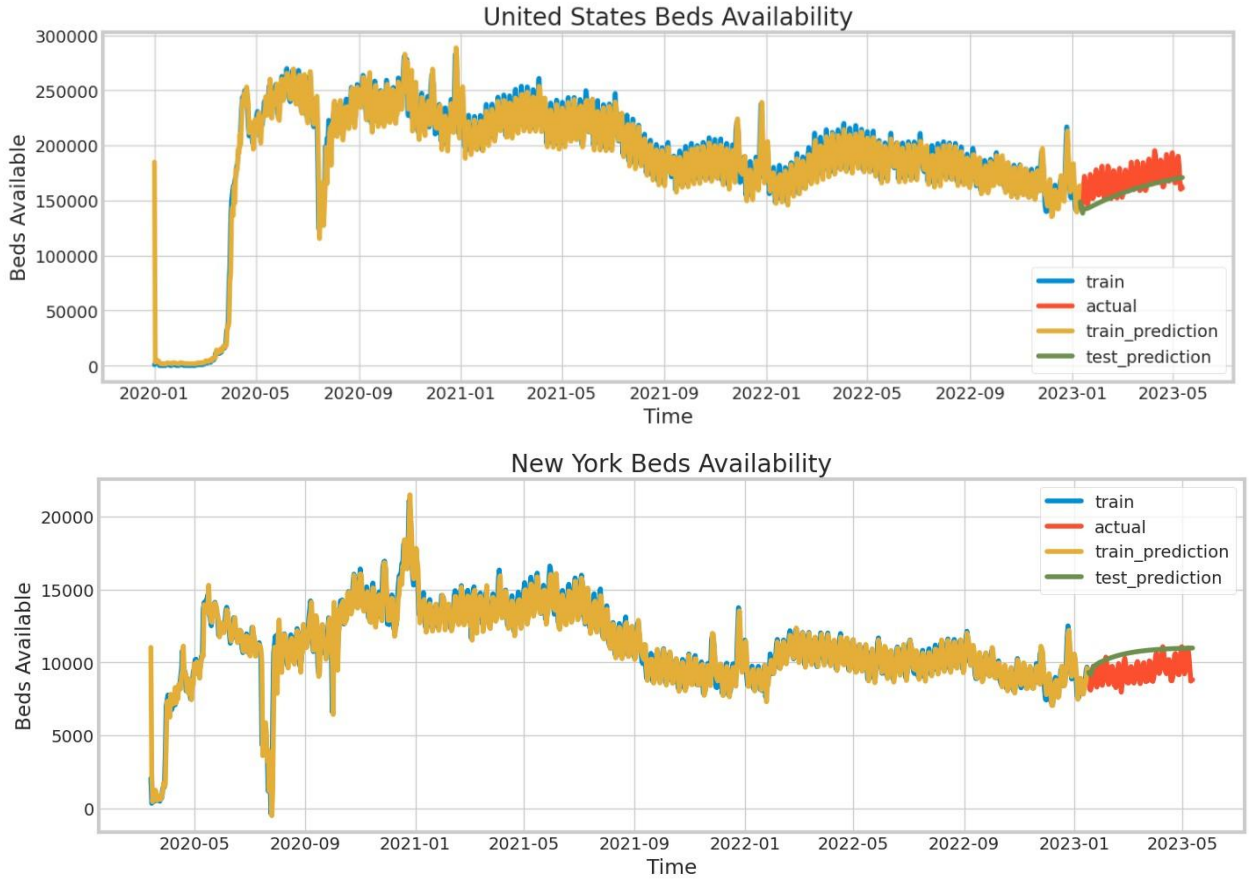
Methodology Chart

4.3 Time series analysis :

The code performs a time series analysis of COVID-19 hospital bed occupancy data for two granularities: New York state and United States as a whole. We filtered relevant columns from the data namely inpatient beds, beds used, and beds used for covid, and calculated beds available using these fields.

We then employ the **ARIMA** model, imported from statsmodels, using 90% of the data and test it against the remaining 10% for availability of inpatient and ICU beds for New York state and for the entire country. This model employs autoregressive and moving average components to detect trends and patterns in the time series data. We keep two 2 lagged values of the dependent variable for the autoregressive term and moving average term includes 6 lagged forecast errors, which are used to capture any seasonal patterns that may exist in the data.

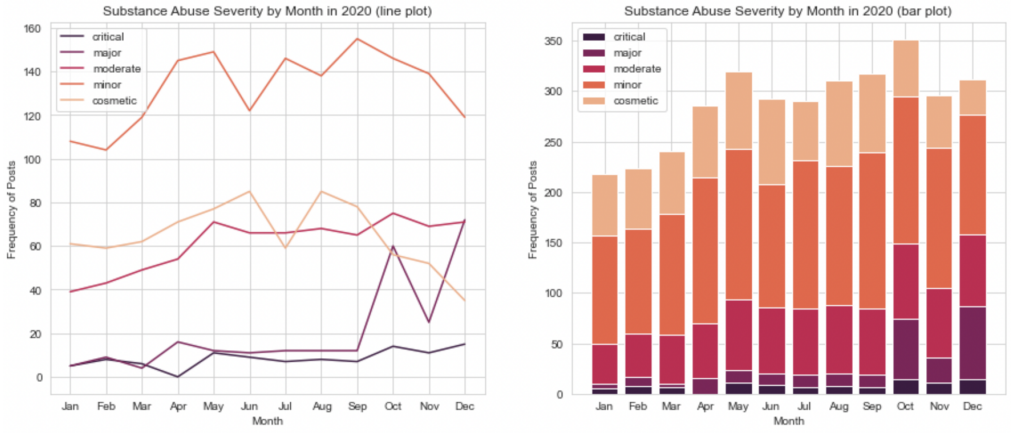
After training the model on a portion of the available data spanning from May 2020 to January 2023, we apply it to forecast future bed availability from January 2023 to May 2023. We also compared ARIMA, SARIMA and Holts models for forecasting bed availability and found ARIMA to be the best based on root mean squared error values.



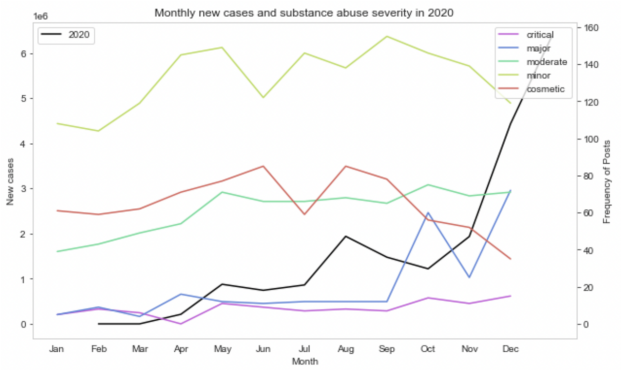
5. Results:

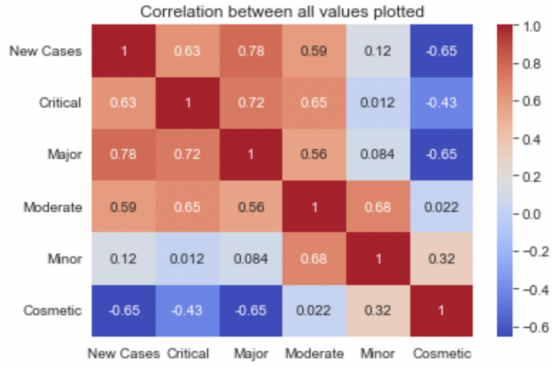
5.1 Sentiment Analysis Results:

Using sentiment analysis, we categorized posts into 5 severity classes and plotted them by year from 2018 to 2023 to track changes in each category over time. This approach provided us with a broad understanding of the distribution of post categories year by year.



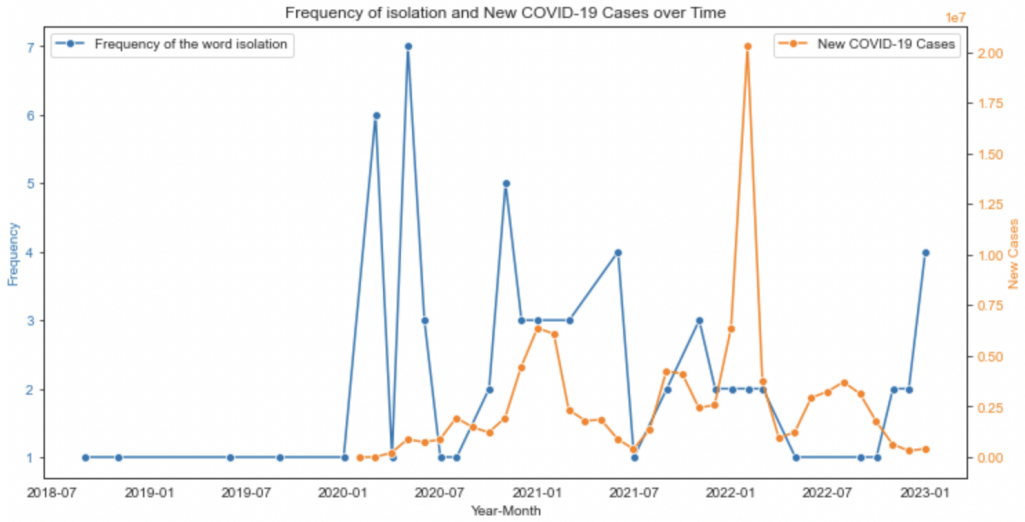
Next, we attempted to correlate the number of covid cases in a year with the number of posts classified as critical, major, moderate, minor, and cosmetic.





The chart displayed above depicts the number of COVID-19 cases in the year 2020 as represented by the black line. Based on the heatmap presented above, it is evident that there exists a significant correlation between the number of new COVID-19 cases and the frequency of posts classified as critical, major, and moderate. It is essential to note that the above conclusions are based on the data collected in the year 2020, and further insights from the years 2021 and 2022 can be found in our Jupyter Notebook.

5.2 Frequency Analysis Results:



The above is a chart of the usage of word ‘isolation’ compared to covid cases over time. More such plots can be found in our notebook.

5.3 ARIMA forecast output:

The result tables show our predictions for inpatient and ICU bed availability from March 2023 to mid-May 2023. The values are in line with the actual values with slightly poor performance on ICU bed availability at the country level. The proximity in train and test RMSE values indicated good generalization of the ARIMA model.

NY state:

week	Actual beds	Predicted beds	Actual icu beds	Predicted icu beds
10	64739	75024	8224	9083
11	65189	75450	8376	9251
12	64570	75794	8454	9406
13	66793	76070	8537	9549
14	69855	76293	8668	9681
15	68924	76473	8714	9801
16	66647	76618	8509	9913
17	69773	76735	8758	10015
18	70808	76829	8804	10109
19	46344	54925	5587	7274

USA:

week	Actual beds	Predicted beds	Actual icu beds	Predicted icu beds
10	1174492	1111296	142718	118584
11	1192448	1123361	144926	118851
12	1183742	1134631	145163	119110
13	1192904	1145158	143944	119361
14	1235239	1154990	148667	119604
15	1232948	1164173	148597	119839
16	1216311	1172751	148897	120067
17	1246315	1180763	152209	120288
18	1234436	1188247	151322	120502
19	835176	853055	100660	86200

6. Conclusion:

In conclusion, our project aimed to contribute to improving healthcare outcomes during pandemics, specifically by using sentiment analysis and time series analysis techniques. The sentiment analysis on Reddit data revealed a strong correlation between the volume of new COVID-19 cases and the frequency of critical, major, and moderate posts related to mental health and substance abuse. Our time series analysis accurately predicted hospital bed availability, providing a helpful tool for hospitals, public health officials, and policymakers in planning and responding to pandemics. Our project aligns strongly with the UN Goal 3 of ensuring healthy lives and promoting well-being for all at all ages. By applying advanced analytical techniques to healthcare data, we hope to contribute to the larger goal of preparing for future pandemics and improving mental health outcomes.

7. References:

1. **Impact of covid on mental health: https://www.covidminds.org/**
2. **WHO Data: https://www.who.int/data/gho**
3. **Mental health during the COVID-19 pandemic: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0277562**
4. **XLNET paper: https://arxiv.org/abs/1906.08237**
5. **Dataset:** https://www.kaggle.com/datasets/pavellexyr/the-reddit-covid-dataset
6. **Dataset:** COVID-19 Reported Patient Impact and Hospital Capacity
7. **Dataset:** Covid Prevelance in the US
8. **ARIMA Model used : https://www.researchgate.net/publication/328633706_Forecasting_of_demand_using_ARIMA_model**