Crisis Management
Analysis of COVID-19

November 03, 2023

Prepared for
S2 2023 - Business Analytics Project

Prepared by

Taylor Chu - 47484039

Manasbi Poudel - 47552794

Ngoc Hanh Dung (Daisy) Le - 47525088

Jenna (Cam Tu) Pham - 46864598

List of Tables and Figures

AI Use Acknowledgement

We acknowledge the use of chat GPT to help find models that are best for predicting the number of hospitalized patients, and for finding out how to do sentiment analysis and topic modeling effectively.

Executive summary

The client briefing report focuses on the crisis management of COVID-19. The pandemic's impact was analyzed through historical data, ongoing trends, and outlook based on the Twitter sentiments. Exploratory data analysis (EDA) was conducted on global data sets as well as Australia and New South Wales to check the impact of COVID-19. The purpose of this EDA was to shed light on ways in which the pandemic has spread, such as the time lag between waves of cases and increases in the death toll, and the potential reduction in the impact of subsequent waves as a result of vaccines and better treatments. The trend analysis indicated that the burden on healthcare systems varied over time and that the severity decreased in subsequent waves. Hospitalized patients were predicted using predictive models. More COVID-19 cases were reported in areas with more hospital beds, maybe as a result of improved testing. Higher case numbers were seen in older populations, which may indicate increased susceptibility or reporting. Stricter health regulations are associated with fewer cases, but complex, inverse correlations are seen with behaviors like smoking and handwashing.

The report also focused on sentiment analysis to determine how the public felt about the situation. Though the limitation of the unavailability of the Twitter data set made it challenging for the proper analysis of the report, some key themes were discovered that opened the door for further research. Key themes that emerged from the analysis of tweets during the pandemic included conversations about vaccinations, the use of masks, political reactions, government assistance, difficulties faced by healthcare workers, public health directives, news about research, testing procedures, and the effects of lockdowns and social distancing measures.

A Crisis Management Dashboard was created at the end to support NSW Health in making informed decisions. The dashboard offers real-time data, geographic insights, trending topics, and an hourly activity monitor. It serves as a central hub that promotes transparency and compliance while improving communication, situational awareness, and resource allocation in a variety of crisis situations.

Slides for presentation: https://prezi.com/view/TXNa8Jte83ArYYFZ2l2B/
Github: https://github.com/navasikya06/business-analytics-project-2023

1. Introduction

As of 12th October, there have been 771,191,203 confirmed cases of COVID-19, including 6,961,014 deaths, reported to WHO (World Health Organization, 2023). COVID was declared a public health emergency in a matter of months of first contraction. But after administering more than 13,516,282,548 (World Health Organization, 2023) vaccine doses, it is no longer declared as a global health emergency (Williams, 2023). While the immediate threat has been minimized, there are potential threats of COVID to break again or some virus turning lethal leading to a global health outcry in no time.

The COVID pandemic taught us how adaptable people can be, and how eager they were to adjust their conduct to keep themselves and others safe. Despite the difficulties, most people followed the guidelines during the peak of the pandemic. COVID has demonstrated how resilient humans can be (Kaye-Kauderer, Feingold, Feder, Southwick & Charney, 2021). These pandemic adaptations, as well as the fact that our pre-pandemic behaviors returned so fast, demonstrate the importance of social cues and social norms in behavior. Putting on a mask or keeping our distance from others were habits - reflexive responses activated in reaction to contextual cues such as viewing signs with images of people socially separating themselves (Neal, Wood, Labrecque & Lally, 2012). As a social being needing social cues, many people took the support of social media sites to express their opinions and views on the COVID. There have been many social media cues and words used to express what they think on the topic.

The WHO conducted a worldwide study on the use of digital technologies during a health crisis in partnership with Wunderman Thompson, the University of Melbourne, and Pollfish. The study included around 23,500 people aged 18 to 40 from 24 countries spread over five continents. The study's key findings were that, contrary to popular belief, scientific news and content were regarded to be the most share-worthy of material when compared to personal information, photographs, other articles, and other types of potentially worrying information (World Health Organization, 2021). In the aftermath of enormous global issues such as the COVID-19 epidemic, digital technologies provide us with both benefits and drawbacks. Increasing our understanding of the possible threats that social media may pose can help us navigate the usage of these platforms in a constructive and useful way in the future.

1.1. Twitter Sentiment Analysis

Twitter is one of the most popular social media sites, and it had a tremendous rise in tweets on coronavirus, including positive, negative, and neutral tweets, in a short period of time. People express their feelings on social media during the isolation process; nevertheless, social media

contains real-time and valuable information regarding COVID-19. However, data from social media may be useless or deceptive at times. Suffering is exacerbated if they come across inaccurate and gloomy material on social media. With the new normal of "staying at home", "working from home", and "isolation time", social networking sites have become popular for sharing news, ideas, emotions, and advice (Jalil, et al., 2021). Online social media sites such as Twitter and Facebook include a lot of noisy data, making it difficult to discover relevant material from them. However, once cleaned, this noisy data captures human feelings and emotions, expressions, and thoughts. When thoroughly examined, it reveals a great deal about a vast human group's current mood, attitude, and nature (Nandwani & Verma, 2021). Online social media platforms such as Twitter have become critical sources of health-related information offered by healthcare experts and citizens. People have been sharing their thoughts, opinions, and feelings about the COVID-19 pandemic on social media platforms (Garcia & Berton, 2021).

Sentiment analysis, often known as opinion mining, is a technique for determining whether a user's perspective on a topic is positive or negative. Sentiment analysis is described as the process of extracting relevant information and semantics from text utilizing natural processing techniques in order to determine the writer's attitude, which might be positive, negative, or neutral (Onyenwe, Mbeledogu, Onyedinma & Nwagbo, 2020).

1.2. Problem Statement

As part of our client briefing report on the crisis management, we chose COVID as our crisis and worked on to understand the crisis through three different aspects. First, we looked at the past records to see what happened during the crisis and used exploratory data analysis to look at the situation. Then we analyzed the ongoing situations through trend analysis and prediction of the hospitalized patients. And then we tried to see what the future holds for the crisis through social media sentiment analysis to check the public opinion and sentiments towards the crisis.

2. Exploratory Data Analysis

We performed exploratory data analysis based on the COVID-19 dataset worldwide (Our World In Data) and also explored data of COVID-19 cases from Australia and New South Wales. We analyzed to understand the patterns, trends, and relationship between the daily new COVID-19 cases and other variables globally. More information about the OWID COVID-19 dataset: https://github.com/owid/COVID-19-data/tree/master/public/data

2.1. Trend analysis of COVID-19 cases, deaths, and hospitalized patients worldwide

The pandemic has had several waves, each with varying impacts in terms of cases, deaths, and potential strain on healthcare systems. There is a noticeable lag between spikes in cases and spikes in deaths as shown in figure 1, which is expected as it takes time for the disease to progress. In the later part of the chart, it appears that the impact of each wave in terms of deaths has been decreasing, possibly due to vaccination, the development of better treatments, or the spread of less severe variants.

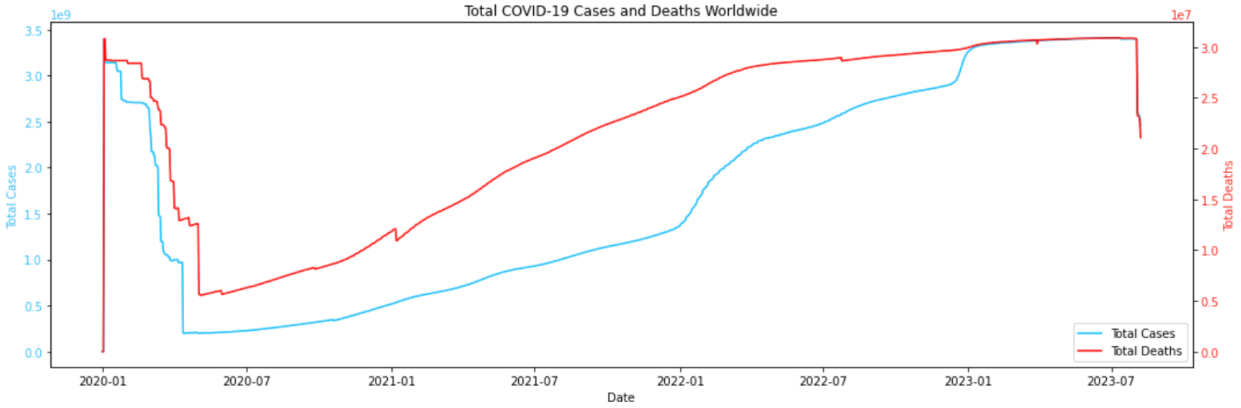

Figure 1: Trend of COVID-19 cases and deaths worldwide

The trend for hospitalized patients below would provide insights into the burden on healthcare systems during different waves of the pandemic.

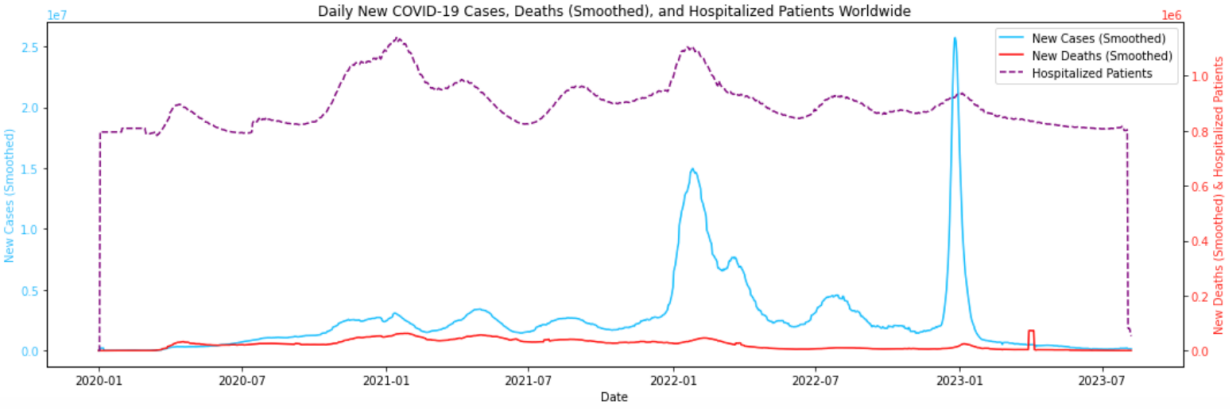

Figure 2: Daily new COVID-19 cases vs. hospitalized patients worldwide

2.2. Correlation analysis of COVID-19 cases

We perform the correlation analysis of COVID-19 cases vs. all other factors. We observe that the factors which have strongest positive correlations to COVID-19 outcome include population, new tests, and the age of people. New vaccinations cases show a moderate positive correlation to COVID-19 outcome. This might seem counterintuitive at first since vaccinations are meant to

reduce the spread of the virus. However, this correlation might be capturing the fact that countries with higher numbers of COVID-19 cases also ramp up their vaccination efforts.

Factors have strongest negative correlations to COVID-19 outcome: hospital beds and stringency index, handwashing facilities. Some factors,such as female and male smokers, show a weak negative correlation to COVID-19 outcome, which are counterintuitive.

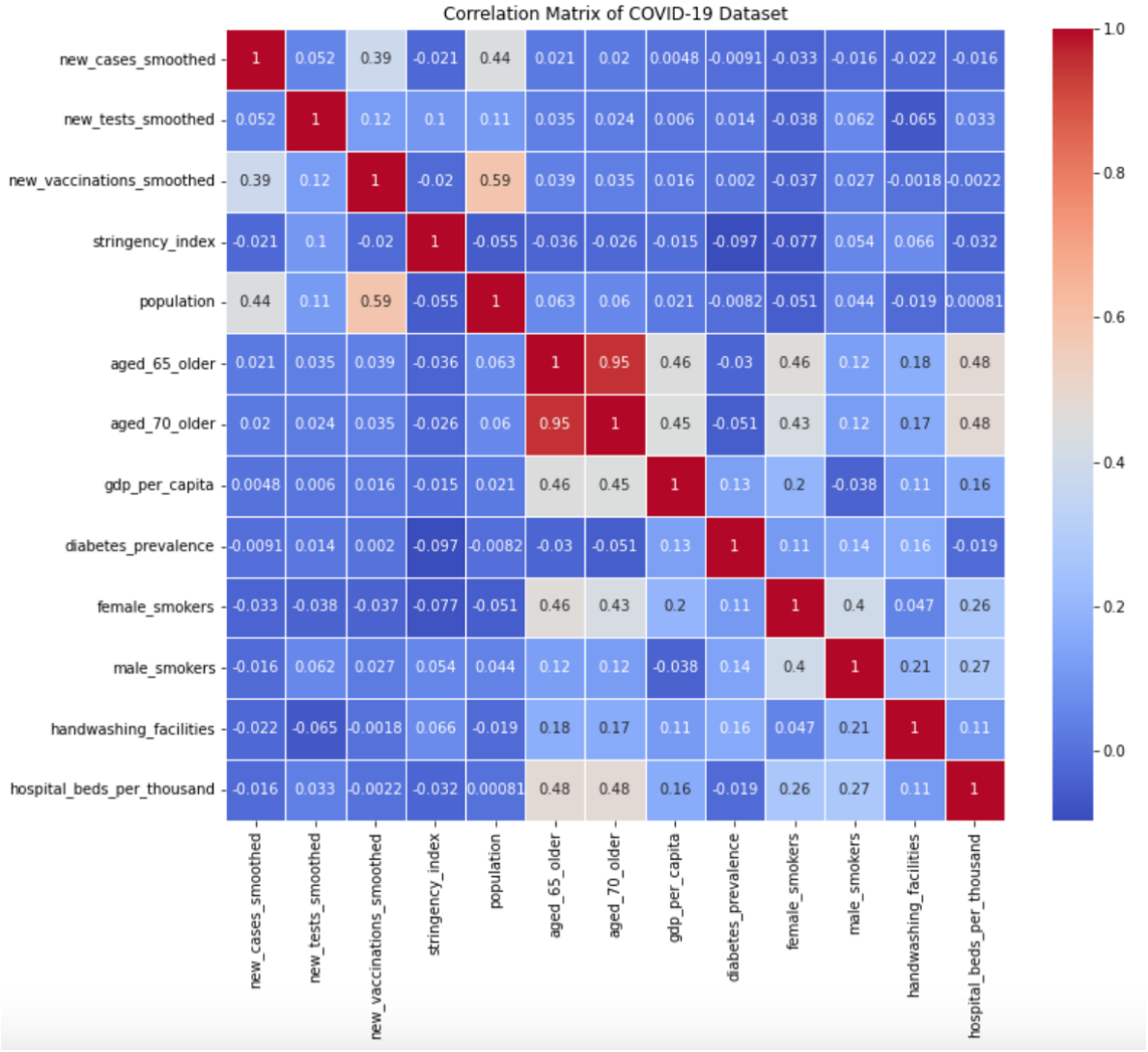

Figure 3: Correlation matrix of COVID-19 dataset

3. Predictive Model for Hospitalized Patients in Australia

Our target:

We aim to analyze and predict the future number of hospitalized patients by the key features such as a number of new COVID-19 cases, the new test cases, and healthcare systems

Our Approach:

- Data Preprocessing: We proceed data cleansing, especially focusing on the Australian subset for hospitalization prediction.
- Developing and selecting model: We develop and validate the predictive models to estimate the number of hospitalized patients in Australia. The model with the highest performance metrics has been selected.
- Interpretation and Reporting: we interpret the results from exploratory analysis and predictive modeling, to provide actionable insights and recommendations supporting healthcare system monitoring in Australia.

3.1. Data Preprocessing

Our key findings during the data pre-processing steps:

- All columns have missing values.
- We also noticed that the (OWID) COVID-19 dataset was combined from different digital data sources so the way of capturing the data may also be different.
- We also acknowledge that the number of COVID-19 cases also depends on the testing policies of each country or there is a time lag between when a case or death occurs and when it's reported in the dataset.
- Variables definition is also different among countries.
- Human Error: given the vast amount of data and the rapid pace at which it's been collected, human errors in data entry or aggregation are possible.
- Political and Social Influences: In some regions, political or social pressures might influence the reporting of cases or deaths, leading to underreporting or misreporting.

3.2. Developing and selecting predictive model

3.2.1. Linear regression model

Due to the unexpected correlation that we observed from the correlation matrix above, we use a linear regression model to understand the relationship between each independent factor and the COVID-19 new cases.

The key results of the linear regression model are below:

- Hospital infrastructure (represented by 'hospital_beds_per_thousand') seems to have a strong positive correlation with 'new_cases_smoothed'. This might be because regions with more hospital beds are better equipped to test and report cases, or it could be a result of more cases leading to a need for more hospital beds.

- Age demographics ('aged_70_older' and 'aged_65_older') are positively correlated with 'new_cases_smoothed', suggesting that older populations might be more susceptible or that these areas are more vigilant in reporting cases.
- The negative relationship with 'stringency_index' suggests that stricter measures (lockdowns, social distancing) might be effective in reducing new cases. However, this is a complex relationship that could be influenced by how well these measures are enforced and adhered to.
- Surprisingly, 'female_smokers' and 'handwashing_facilities' show a negative correlation. It's important to note that correlation does not imply causation, and these relationships might be influenced by other confounding factors.

	Feature	Coefficient
9	aged_70_older	5.208781e+01
11	diabetes_prevalence	1.767798e+01
6	stringency_index	8.868870e+00
12	female_smokers	6.583044e+00
1	new_deaths_smoothed	9.303210e-01
10	gdp_per_capita	3.158808e-03
5	total_deaths	1.203578e-03
0	new_cases_smoothed	1.000691e-03
2	new_tests_smoothed	5.294236e-04
7	population	-6.974392e-07
4	total_cases	-3.098196e-06
3	new_vaccinations_smoothed	-1.847456e-04
14	handwashing_facilities	-2.582933e+00
13	male_smokers	-7.790845e+00
15	hospital_beds_per_thousand	-2.688836e+01
8	aged_65_older	-3.076760e+01

Figure 4: Coefficient table from the Linear regression model

Checking on multicollinearity of features vs. the target variables:

- The Variance Inflation Factor (VIF) is used to detect the presence of multicollinearity in regression models. Multicollinearity occurs when two or more independent variables are highly correlated, which can lead to unstable estimates of the coefficients and affect the interpretability of the model.
- VIF less than 5 is often considered acceptable.

	feature	VIF
6	stringency_index	3.777206e+00
1	new_deaths_smoothed	2.747282e+00
0	new_cases_smoothed	2.388047e+00
3	total_deaths	2.360713e+00
2	total_cases	2.156445e+00
5	new_vaccinations_smoothed	1.850429e+00
4	new_tests_smoothed	1.317734e+00
7	aged_70_older	1.645650e-06
11	hospital_beds_per_thousand	3.682626e-07
10	handwashing_facilities	3.438482e-09
8	diabetes_prevalence	0.000000e+00
9	male_smokers	0.000000e+00

Figure 5: Variance Inflation Factor result of features:

Performance of Linear Regression model:

An R-squared value explains the % of the variation in a number of hospitalized patients can be explained by the model. The linear regression model has a high value of R Squared, suggesting that the model has a good fit to the data. However, the MSE is quite large, which might be due to outliers, high variability in the number of hospitalized patients, or the scale of the number of hospitalized patients itself. Therefore, we consider using another stronger model - XGBoost.

```
# Initializing the Linear Regression model
linear_reg_model = LinearRegression()

# Training the model
linear_reg_model.fit(X_train, y_train)

# Making predictions on the test set
y_pred = linear_reg_model.predict(X_test)

# Evaluating the model
RMSE = np.sqrt(mean_squared_error(y_test, y_pred))
r2 = r2_score(y_test, y_pred)

RMSE, r2
```

```
(632.4852423180217, 0.8148892395747923)
```

Figure 6: Linear regression R squared and MSE

3.2.2. XGBoost Model

We observe that the performance metrics of the XGBoost model are absolutely high. In which, MSE value of 11510.0 suggests that the model's predictions are quite close to the actual values.

In addition, the R-squared score of 0.9946 means that the model explains 99.44% of the variance in the number of hospitalized patients, indicating a very good fit of the model to the data.

These results suggest that the model is highly effective in predicting the number of hospitalized patients with the selected features and parameters. The high R Squared score, in particular, indicates that the model captures nearly all the variability in the number of hospitalized patients, making it a powerful tool for forecasting and planning purposes.

```
# Initialize the XGBRegressor with the best parameters
xgb_best = xgb.XGBRegressor(
    objective='reg:squarederror',
    colsample_bytree=0.9,
    learning_rate=0.1,
    max_depth=5,
    n_estimators=200,
    subsample=0.9,
    random_state=42
)

# Train the model on the training data
xgb_best.fit(X_train, y_train)

# Make predictions on the test data
y_pred_xgb = xgb_best.predict(X_test)

# Evaluate the model's performance
rmse_xgb = np.sqrt(mean_squared_error(y_test, y_pred_xgb))
r2_xgb = r2_score(y_test, y_pred_xgb)

# Print the evaluation metrics
print("Root Mean Squared Error (XGBoost):", rmse_xgb)
print("R² Score (XGBoost):", r2_xgb)

# Save the model to a file
xgb_best.save_model('prediction/xgb_model_v3.json')
```

```
Root Mean Squared Error (XGBoost): 107.2851261908167
R² Score (XGBoost): 0.9946738927880472
```

Figure 7: XGBoost R squared and MSE

One of the advantages of the XGBoost model is that it can select the most important features for the model. The most important features are in the graph below:

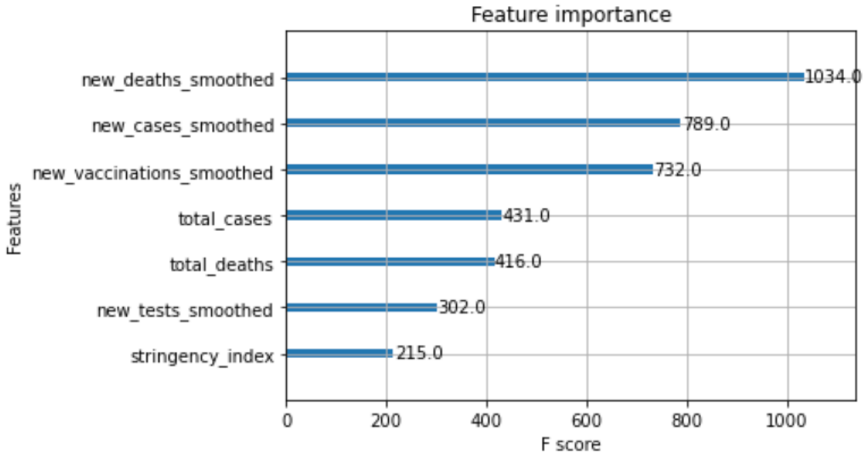

Figure 8: Feature importance in XGBoost model

Visualization of predicted value vs. actual value of number of hospitalized patients in XGBoost model:

The XGBoost model achieves the optimal performance which can be seen via the alignment of predicted value and the actual value of number of hospitalized patients in Australia in the visualized chart below.

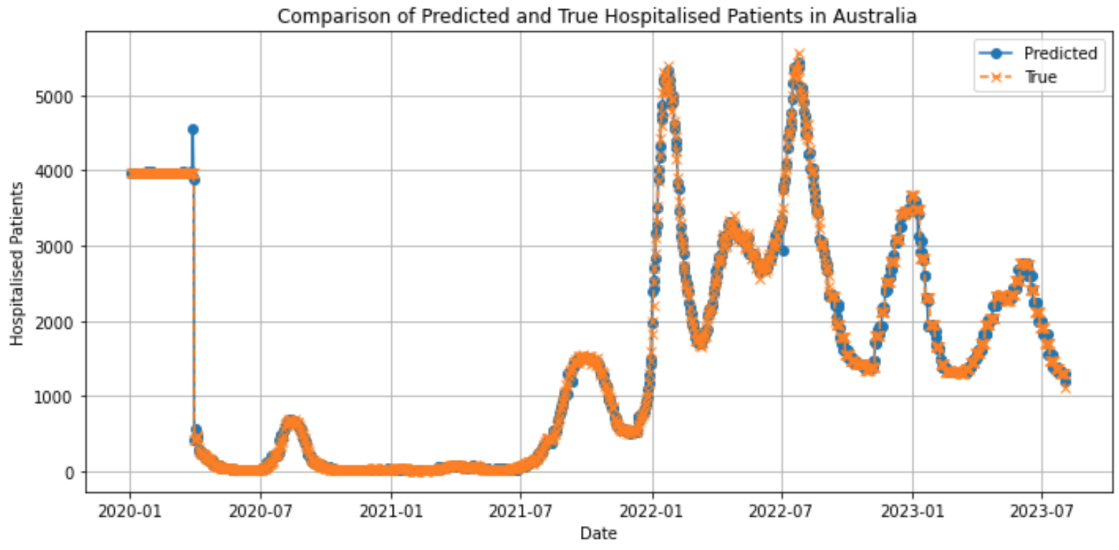

Figure 9: Comparison of predicted and true hospitalized patients in Australia

In Jan 2020, the model overestimates the hospitalized patients significantly. The model may not have enough data from the early stages of the pandemic to make accurate predictions, or it could be influenced by other features that were not adjusted properly.

Towards the later dates (July and August 2023), the predicted values are much closer to the true values, indicating that the model is performing better at this stage. This might be due to having more data available for the model to learn from, resulting in more accurate predictions.

The predicted values seem to be quite stable and do not show large fluctuations, which is a good sign. However, this could also mean that the model is not sensitive enough to changes in the input features, which could be a point of investigation. Further investigation and possibly model adjustment might be needed to improve the model's performance across all time periods.

We also mention in this report the further modeling methodology to help optimize the performance of the prediction model.

4. Social Media Sentiment Analysis

4.1. Data Exploration

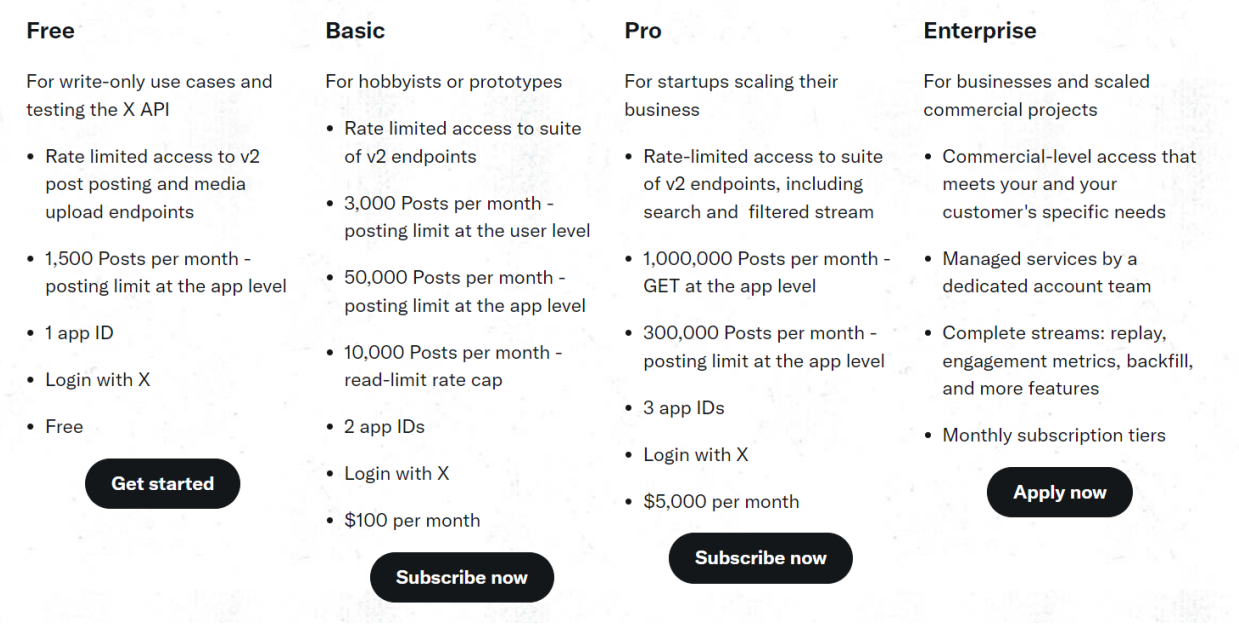

Figure 10: Twitter Data Purchase Package

Getting tweets directly from Twitter API and hydrating Twitter IDs both don't work anymore (change since March 2023 when Twitter was acquired). The minimum cost for reading tweets directly is $100 a month. Instead, we used a COVID-19 related dataset from Kaggle.

The dataset comprises tweets from all over the world related to COVID-19, including 13 distinct columns. These tweets were collected in three phases, namely April-June 2020, August-October 2020, and April-June 2021. The tweets are related to discussions and mentions of the COVID-19 pandemic, with hashtags such as #COVID-19, #coronavirus, #COVID, #covaccine, #lockdown, #homequarantine, #quarantinecenter, #socialdistancing, #stayhome, #staysafe, and more. The dataset's primary attributes are the Tweet ID, Creation Date & Time, Source Link, Original Tweet, Favorite Count, Retweet Count, Original Author, Hashtags, User Mentions, and Place.

4.2. Data Preprocessing

The data underwent a rigorous pre-processing regimen using a custom function rooted in the capabilities of the Natural Language Toolkit (NLTK) — a prominent Python library for Natural Language Processing. Initially, all tweets were converted to lowercase, followed by the elimination of superfluous white spaces, numbers, special and ASCII characters, URLs,

punctuations, and stopwords. Subsequently, every occurrence of the word 'COVID' was transformed into 'COVID19' due to the prior removal of numerical data. Stemming was then applied to truncate inflected words to their respective stems. With a clean dataset in place, sentiment analysis was conducted using NLTK's Sentiment Analyzer. This rendered sentiment scores for three categories: positive, negative, and neutral. These scores paved the way for the computation of a compound sentiment score for each tweet. Classification ensued based on these compound scores, bucketing tweets into Positive, Negative, or Neutral sentiments. A specific algorithm was devised to assign sentiment polarity ratings: tweets with a compound score below zero were marked as Negative (0.0), those above zero as Positive (1.0), and the rest, which displayed a neutral sentiment, were designated a score of 0.5.

4.3. Sentiment Analysis

1. Distribution of Sentiments

The bar chart shows the distribution of sentiments (positive, negative, and neutral) among the tweets in the dataset. From the chart, we can observe that:

- A significant number of tweets are categorized as neutral.
- There are also a considerable number of positive tweets.
- Negative tweets are the least common among the three categories.

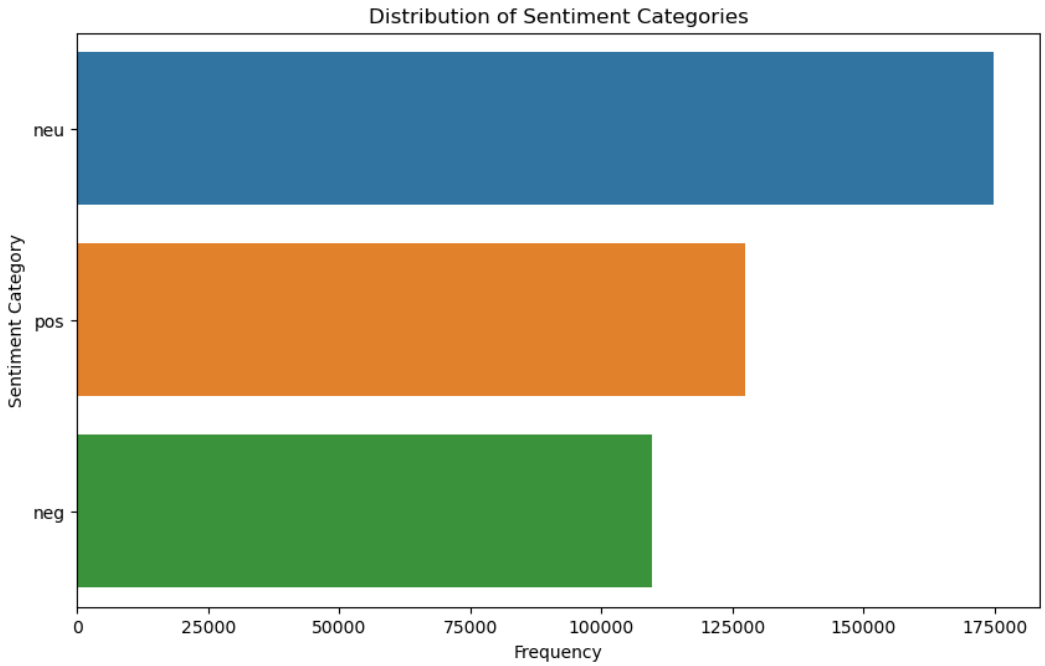

Figure 11: Distribution of sentiment categories

2. Sentiment Scores Over Time

The time series plot as shown in Figure 12 shows how the sentiment scores (compound, positive, and negative) have changed over time. Each subplot represents a different type of sentiment score:

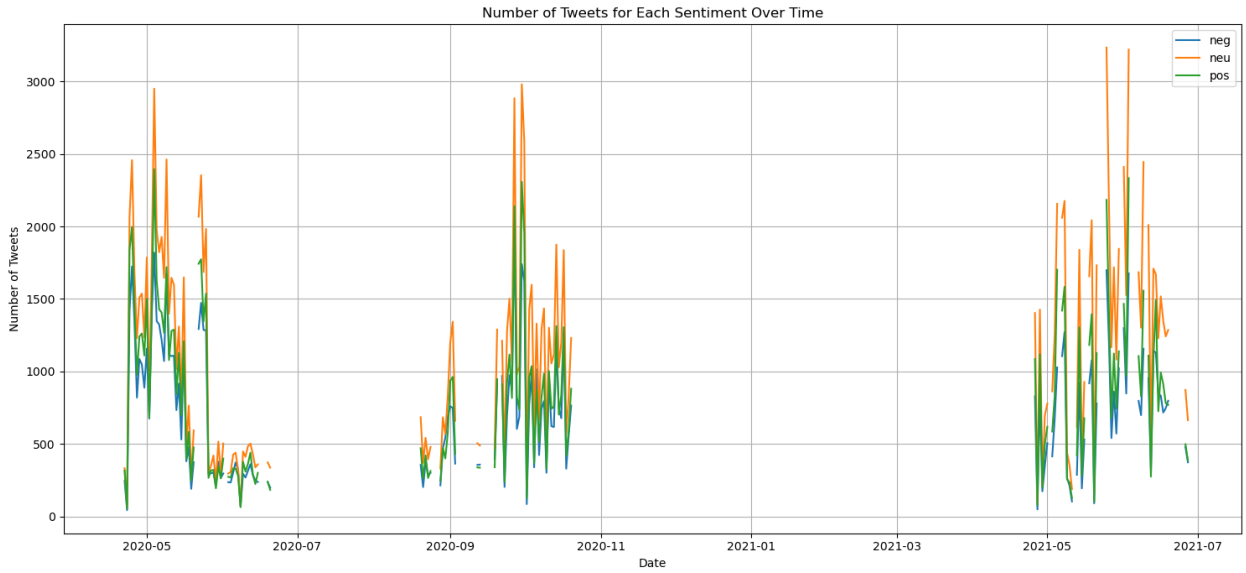

Figure 12: Number of tweets for each sentiment over time

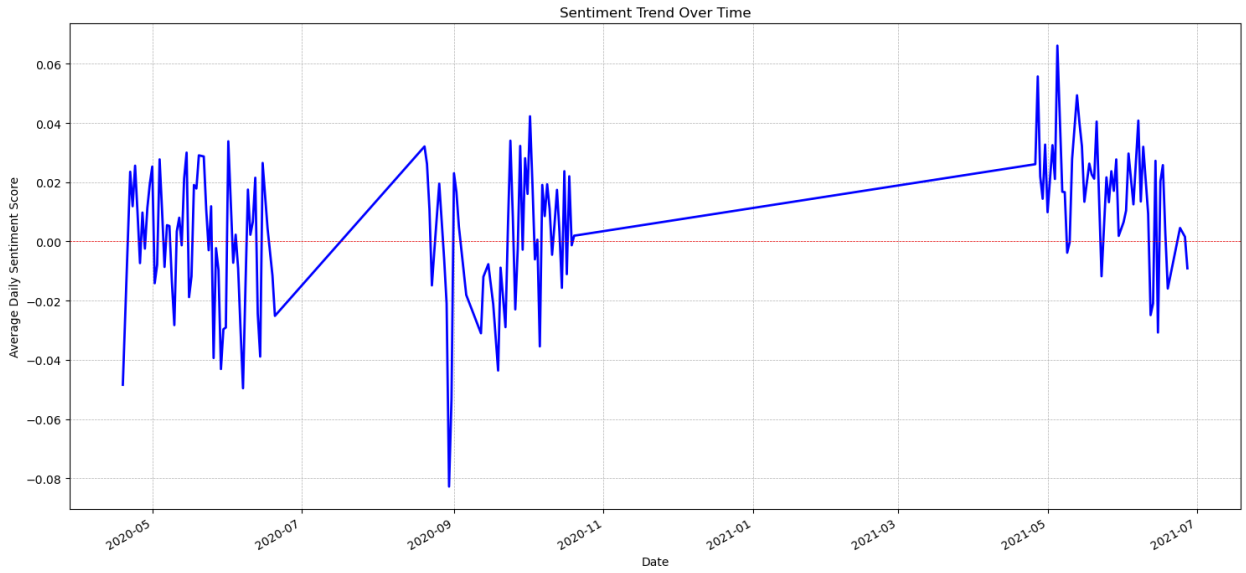

Figure 13: Sentiment Trend over Time by daily average sentiment

Compound Sentiment Score Over Time: This score is a composite score that calculates the sum of all the lexicon ratings and normalizes the result between -1 (most negative) and 1 (most positive).

Positive Sentiment Score Over Time: This score represents the proportion of positive words in the text.
Negative Sentiment Score Over Time: This score represents the proportion of negative words in the text.

From the plots, we can observe fluctuations in sentiment scores over time, which reflects a more positive sentiment from 2020 to 2021.

4.4. Correlation Analysis

When comparing the number of COVID-19 cases with the volume of tweets from April to June 2020, we observed a pattern: a rise in the number of tweets preceded a peak in COVID-19 cases. The correlation between daily cases and sentiment, and between daily cases and tweet count, is -0.5 and -0.47, respectively. Our correlation analysis revealed a negative relationship, indicating that as the situation worsened, the sentiment on Twitter became more negative and the activity of tweeting decreased.

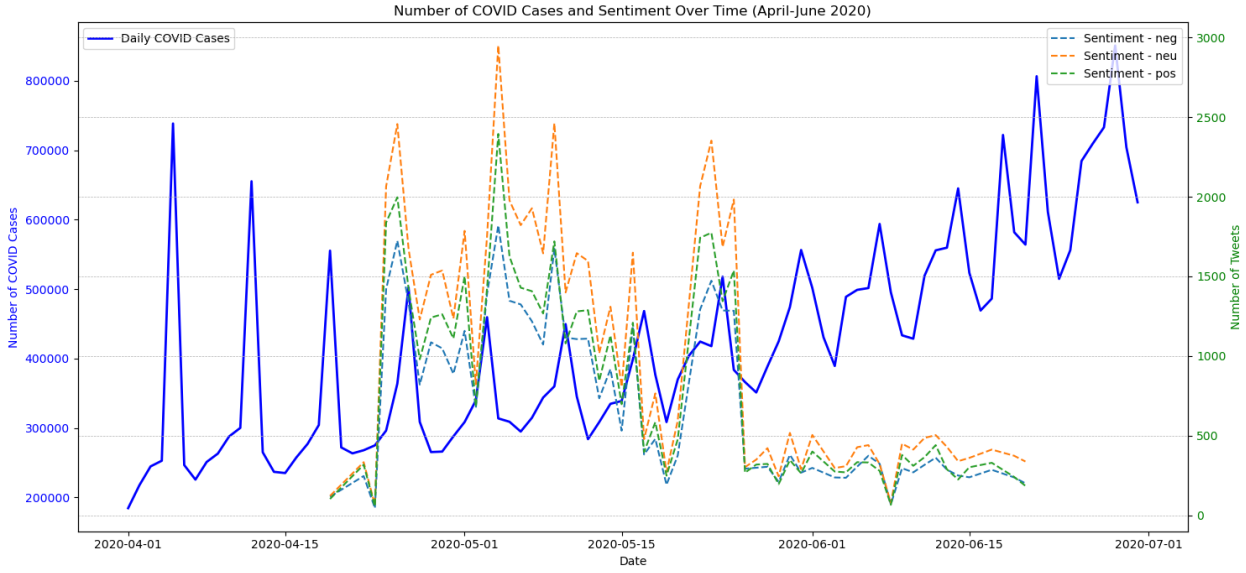

Figure 14: Number of COVID cases vs number of tweets by sentiment type (April-June 2020)

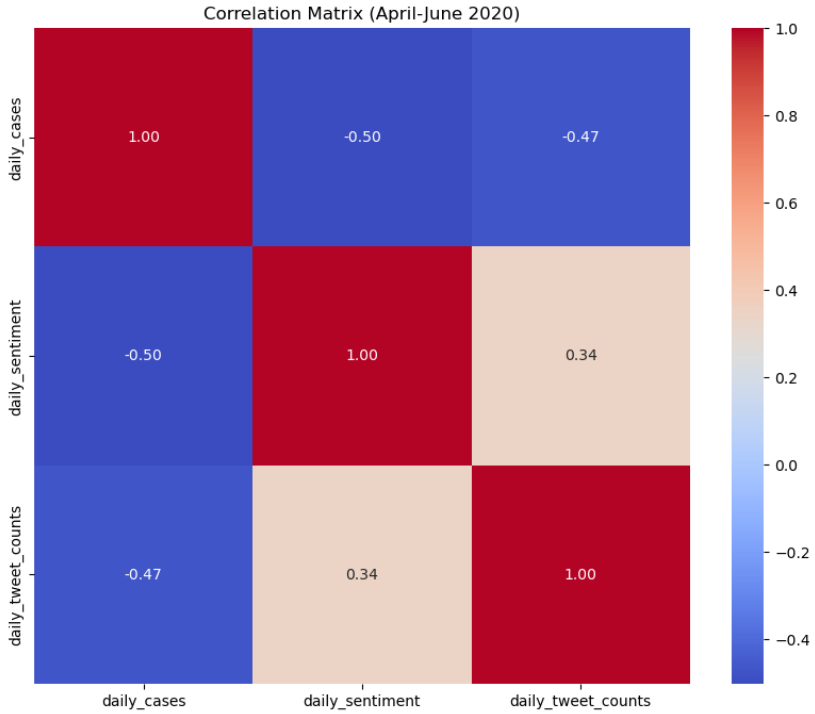

Figure 15: Correlation matrix for daily number of tweets, average sentiment, and number of cases (April-June 2020)

For the period from August to October 2020 and April to June 2021, the correlation coefficients are very close to 0, indicating a very weak or no linear relationship between daily COVID-19 cases and Twitter sentiment scores during these periods for the "World" location. This could mean that the sentiments expressed in tweets are not strongly influenced by the number of new COVID-19 cases at a global level, or it could indicate that other factors are influencing the sentiments expressed on Twitter.

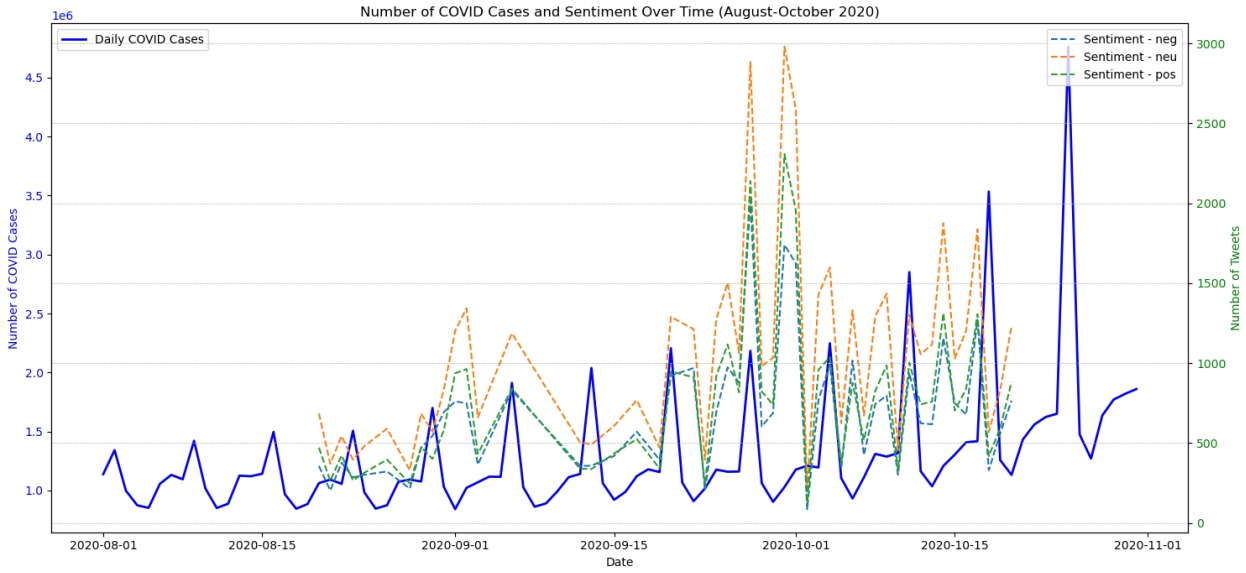

Figure 16: Number of COVID cases vs number of tweets by sentiment type (August-October 2020)

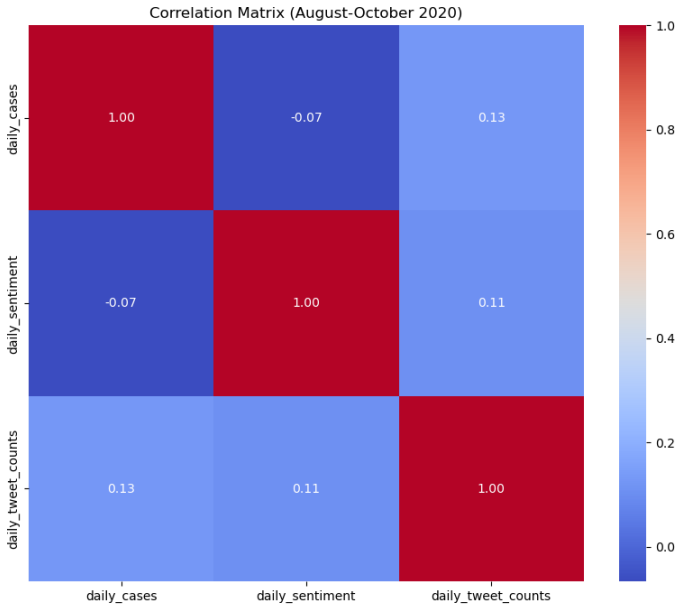

Figure 17: Correlation matrix for daily number of tweets, average sentiment, and number of cases (August-October 2020)

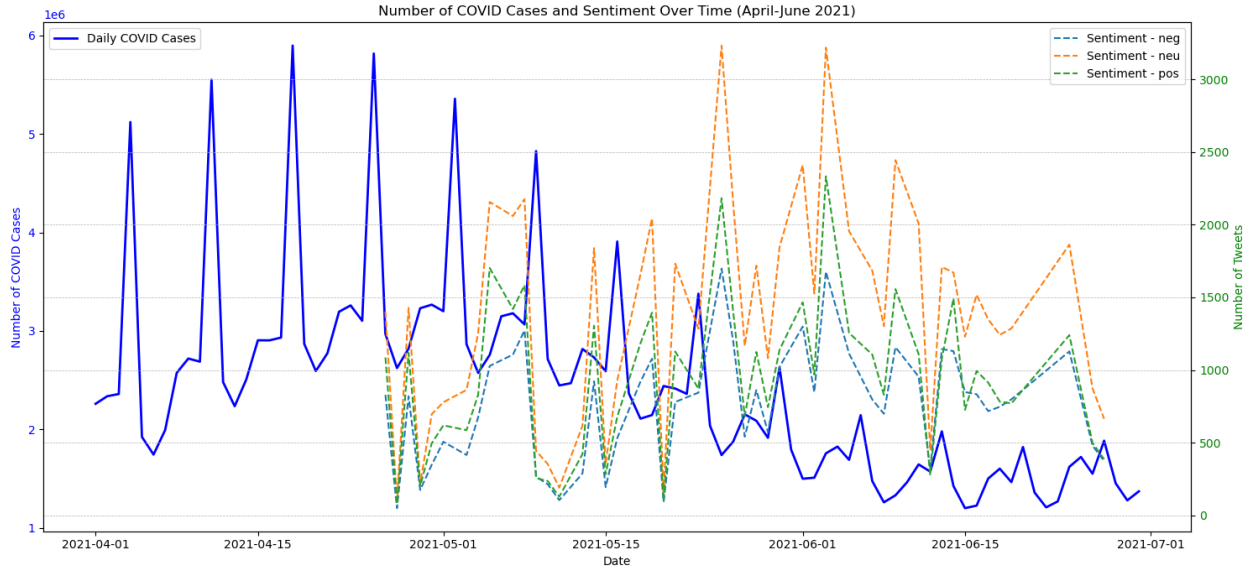

Figure 18: Number of COVID cases vs number of tweets by sentiment type (April-June 2021)

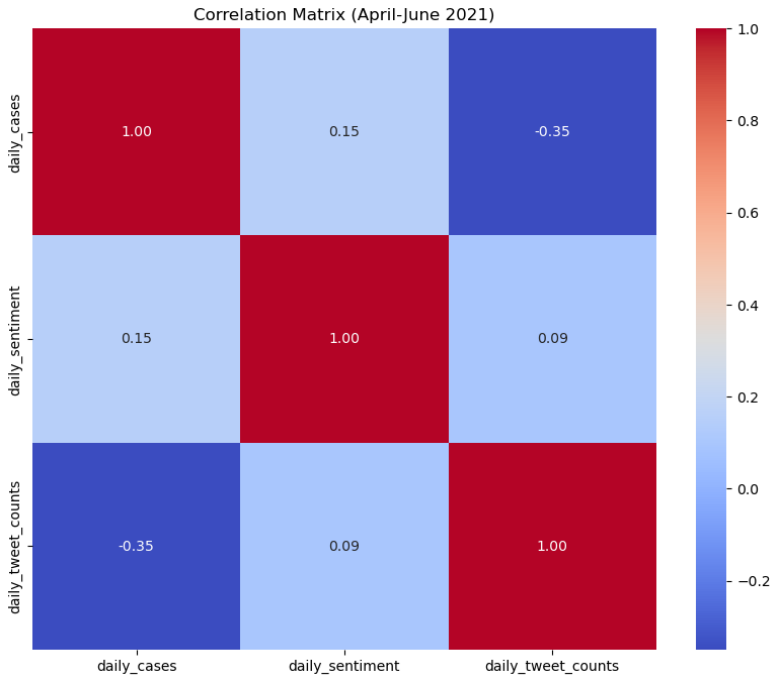

Figure 19: Correlation matrix for daily number of tweets, average sentiment, and number of cases (April-June 2021)

4.5. Topic Analysis

Furthermore, an analysis of the predominant topics from the tweets (using Latent Dirichlet Allocation or LDA) unveiled several key themes:

- Vaccination & Precautionary Measures: This theme revolved around vaccine discussions, data about vaccinations, the significance of face masks, and the pandemic's ramifications on families.
- Political Leadership & Response: This primarily concerned discussions about political figures, especially Trump, the international response to the pandemic, the UK's position, and leadership sentiments.
- Government Support & Vaccination Efforts: Here, the focus was on vaccination campaigns, governmental backing, crisis management, and calls for support.
- Healthcare Workers & Patient Care: This spotlighted the challenges healthcare workers faced, the care patients received, and the ambiance of hospitals during the pandemic.
- Public Health Guidelines: This covered dialogues about health advisories, the essence of staying home, and instructions from health officials.
- Pandemic News & Research: This involved the dissemination of news updates, findings from research, and the pandemic's broad impact on everyday life.
- Testing & Diagnosis: This emphasized discussions about COVID-19 testing, counts of positive cases, and the value of prompt diagnosis.

- Lockdowns & Restrictions: This detailed conversations about lockdowns, the principles of social distancing, the pace of vaccinations, and the potential for ensuing waves of the virus.

Topic 0	Vaccination & Precautionary Measures - Discussions about vaccines, data regarding vaccination, the importance of face masks, and the impact of the pandemic on families.	['rt', 'vaccine', 'data', 'face', 'mask', 'family', 'pandemic', 'lost', 'masks', 'years']
Topic 1	Political Leadership & Response - Discussions centered around political figures, notably Trump, the global response to the pandemic, the UK's stance, and opinions on leadership.	['rt', 'trump', 'world', 'president', 'response', 'said', 'great', 'uk', 'covid', 'think']
Topic 2	Government Support & Vaccination Efforts - Emphasis on vaccination drives, governmental support, crisis management, and rallying support.	['rt', 'vaccination', 'fight', 'free', 'students', 'crisis', 'government', 'know', 'covid', 'support']
Topic 3	Healthcare Workers & Patient Care - Focus on the challenges and needs of healthcare workers, patient care, and the hospital environment during the pandemic.	['rt', 'people', 'need', 'like', 'help', 'covid', 'workers', 'care', 'hospital', 'know']
Topic 4	Public Health Guidelines - Conversations about public health recommendations, staying at home, and directives from health officials.	['rt', 'health', 'public', 'say', 'home', 'impact', 'make', 'stay', 'officials', 'pm'],
Topic 5	Pandemic News & Research - Sharing of news updates, research findings, and the general impact of the pandemic on daily life.	['rt', 'just', 'news', 'good', 'days', 'study', 'house', 'use', 'research', 'life']
Topic 6	Testing & Diagnosis - Discussions about testing for COVID-19, the number of positive cases, and the importance of timely diagnosis.	['rt', 'positive', 'tested', 'test', 'patients', 'million', 'people', 'tests', 'testing', 'want']
Topic 7	Lockdowns & Restrictions - Conversations around lockdowns, social distancing measures, vaccination progress, and the potential for a second wave.	['rt', 'lockdown', 'going', 'social', 'global', 'vaccinated', 'restrictions', 'second', 'way', 'wave']
Topic 8	Statistics & Updates - Sharing of new COVID-19 cases, death statistics, and state-level data.	['new', 'rt', 'cases', 'covid', 'deaths', 'reported', 'virus', 'state', 'read', 'confirmed']
Topic 9	Global Impact & Statistics - Discussion about the global spread, focusing on countries like India, the overall number of cases, and daily updates.	['rt', 'time', 'cases', 'death', 'new', 'coronavirus', 'india', 'spread', 'number', 'daily']

Table 1: Topic Analysis by Latent Dirichlet Allocation (LDA)

We can see that Topic 2 has the most number of tweets, which is about government support and vaccination efforts. This is reasonable as people are probably most concerned about these policies.

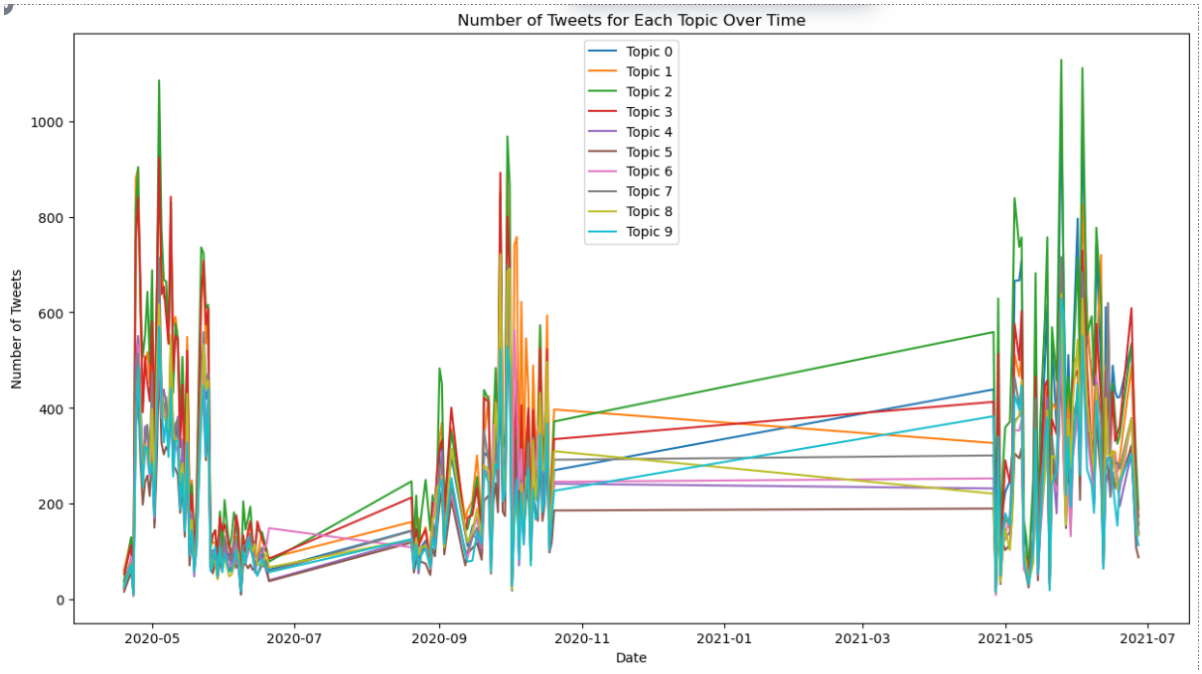

Figure 20: Number of tweets by topic over time

4.6. Specific Analysis for Australia

We do not have Twitter data filtered just for Australia directly, so we referenced a paper that looked into Twitter sentiment analysis for Australia (Lambsal, 2023). We can see that tweets correlate with rises in cases early in the pandemic (2020), but later in 2021 and 2022, even though the number of cases increased, there wasn't an increase in the number of tweets correspondingly. This shows that people stopped having immediate social reaction to rises in cases as the pandemic went on, and this might make sentiment time series not very helpful to predict the number of cases later on in the pandemic.

Lamsal et al. *A Twitter narrative of the COVID-19 pandemic in Australia*

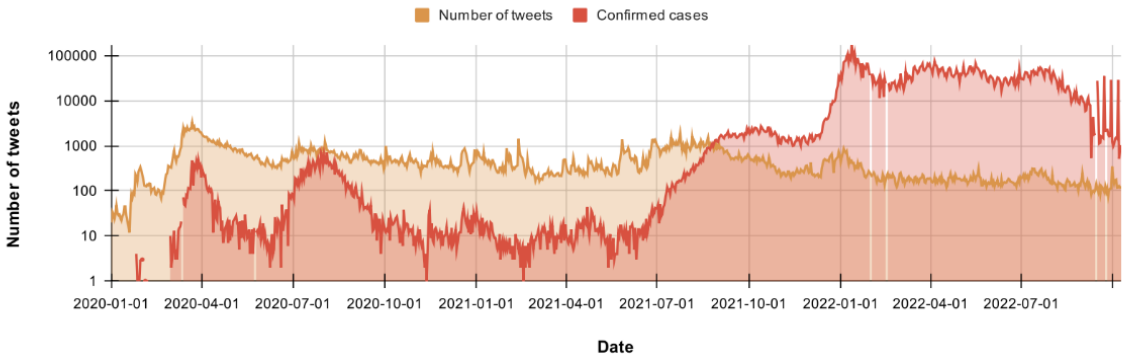

Figure 21: Number of Tweets vs number of COVID cases over time-Australia (Lambsal, 2023)

4.7. Challenges and Limitations

During our analysis, we encountered a significant issue: the inability to access real-time Twitter data, which led us to use a dataset from Kaggle. This dataset might not fully represent the diverse opinions and emotions of humans. Additionally, Twitter is not the most popular social media platform in Australia, suggesting that our findings may not accurately reflect the sentiments of the Australian public. We also relied heavily on numerical sentiment scores, which are unable to capture the complexities of human emotions fully. Last but not least, we did not take into account external factors that could influence tweeting frequency and sentiment, such as major news events or changes to Twitter's algorithms, potentially leading to less accurate conclusions.

4.8. Recommendation for future work

In order to do better forecasting for the number of hospitalized patients, we can use time series models like ARIMA, FB Prophet and XGBoost. A previous study has looked into these 3 models and found that ARIMA performed the best (Lambsal, 2022).

Table 8
Best forecasting model for y.

Approach	Avg. RMSE
Traditional model[a] (ARIMA of $p = 1$, $d = 1$, $q = 3$)[b]	135.387
Additive model (FB Prophet)	236.427
Machine learning model (XGBoost)	341.8

[a]also involves the participation of the models such as *AR*, *MA*, *ARIMA*, *SARIMA*.
[b]the traditional models and their mathematical structures are discussed later in Section 4.2.1.

Figure 22: Forecasting Models for number of cases by time series models (Lambsal, 2022)

To achieve a more holistic understanding of public sentiment regarding COVID-19, future studies could benefit from leveraging real-time data extraction from Twitter. This would allow for the capture of dynamic shifts in sentiment as they occur, providing a more granular perspective on public reaction to ongoing events. Additionally, expanding the data collection to include a broader range of social media platforms, such as Facebook, Instagram, and local platforms dominant in specific regions, can offer a more comprehensive view of global sentiment. Research by Lambsal also reviewed some ARIMAX models (ARIMA combined with explanatory variables) with search index and sick posts information as shown below.

Table 13
Results from fitting the exogenous variables listed in Table 12 and their respective 14 days' lags against 84 weeks of data (January 15, 2020, to August 26, 2021).

	Best fitted model	Exo. Variables count	AIC	RMSE
Baseline	ARIMA(6,2,7)	–	6118.50	37.78
Search index (dry cough)	ARIMAX(9,2,9)	1 and its 14 lags	6019.93	37.46
Search index (coronavirus)	ARIMAX(7,2,5)	1 and its 14 lags	6013.5	37.51
Search index (fever)	ARIMAX(5,2,8)	1 and its 14 lags	5993.47	37.55
Search index (pneumonia)	ARIMAX(6,2,9)	1 and its 14 lags	6001.28	37.52
Search indexes Combined	ARIMAX(7,2,8)	4 and respective 14 lags	6085.15	36.53
Sick posts	ARIMAX(8,2,7)	1 and its 14 lags	5989.78	37.12
Sick posts + Search indexes combined	ARIMAX(3,2,9)	5 and respective 14 lags	6069.28	35.77
Overall posts	ARIMAX(4,2,5)	1 and its 14 lags	5991.94	37.34
Latent variables search	ARIMAX(2,2,3)	14 and respective 14 lags	5941.08	32.97

Figure 23: ARIMAX Models with search index, sick posts and latent topic variables search (Lambsal, 2022)

This research showed that including sentiment score time series based on topics into their ARIMAX model produced better results than previous studies that used ARIMAX with search index and sick posts.

In addition, merging more explanatory data, such as vaccine distribution statistics, stringency index, and percentage of older population, into the ARIMAX models can potentially improve the accuracy of the prediction.

Lastly, there's also potential in exploring advanced Natural Language Processing techniques, like transformer-based models (BERT), to better understand the nuances in the tweets, not just positive, negative and neutral, but also irony, sarcasm, excitement, and other types of emotions.

5. Crisis Management Dashboard Pilot

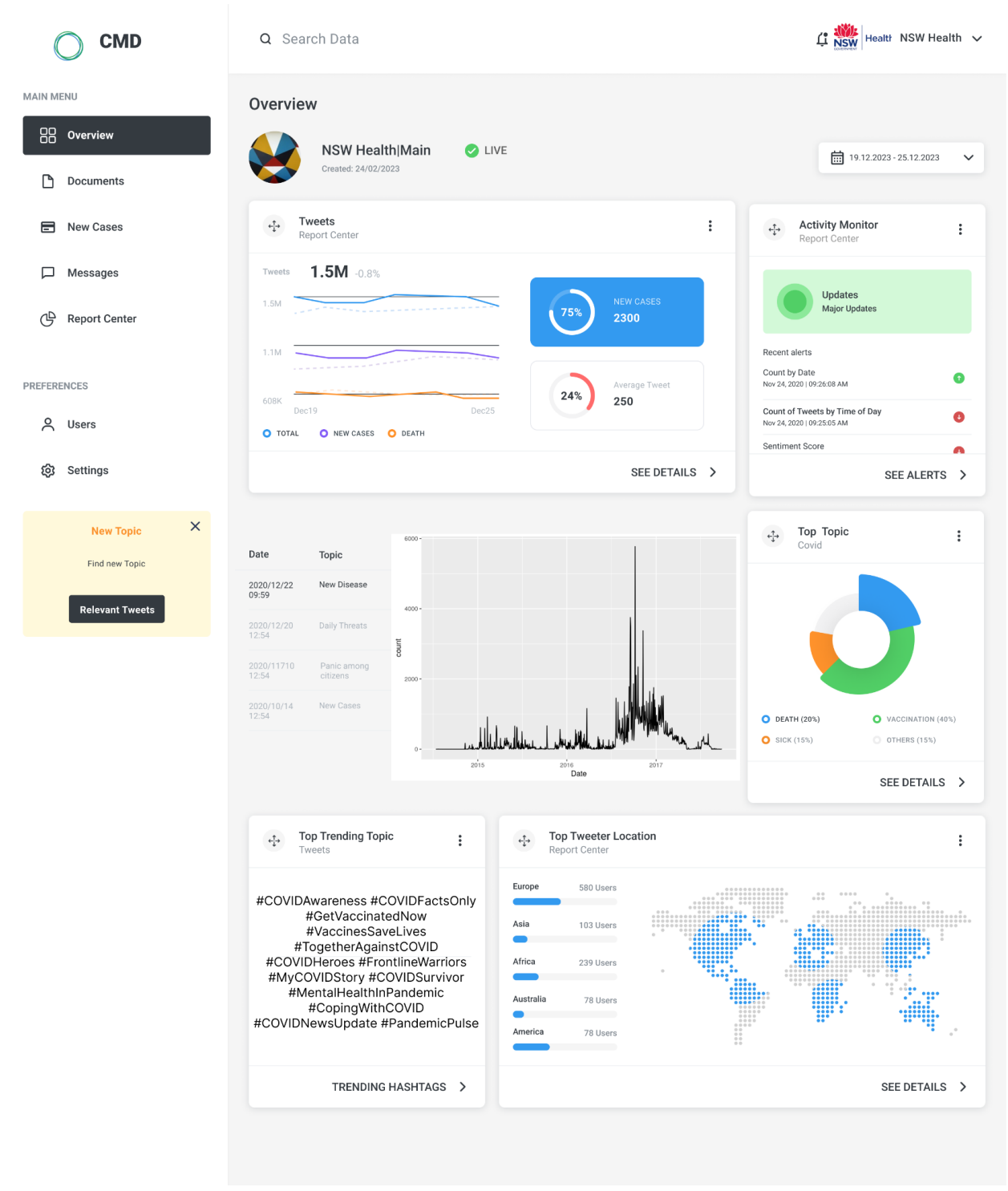

Figure 24: Crisis Management Dashboard

With the intention to help NSW Health have a real-time overview of the crisis, we designed a Crisis Management Dashboard in Figma. The dashboard contains an overview of different key components that would help NSW Health in analyzing the situation. The location mapping helps with the geographic knowledge and overview of the situation. The dashboard also contains a top trending topic which helps in analyzing the trending hashtags, hence understanding the public sentiments. The main feature of the dashboard includes an activity monitor which will help figure out the situation per hour, assisting the organization in making informed decisions.

By incorporating these features into the dashboard, stakeholders will gain a comprehensive understanding of public opinion, sentiment, and engagement during the crisis. This can inform communication strategies, policymaking, and crisis management efforts, ultimately contributing to a more informed and effective response.

For organizations and agencies handling a variety of crises, such as cyberattacks, natural disasters, pandemics, and public health emergencies, a crisis management dashboard as shown in Figure 24, provides several benefits. It serves as a central hub, combining information from various sources to provide situational awareness in real time. This helps decision-makers make quick, well-informed decisions in addition to helping them stay current on the most recent information. As the dashboard assists in identifying crisis hotspots, resource allocation becomes more effective because resources are allocated where they are most needed. Additionally, it improves coordination and communication, supports risk assessment, enables scenario planning, and permits ongoing performance monitoring, all of which help with crisis response.

Furthermore, the transparency it provides helps meet regulatory requirements and foster public trust, especially when combined with compliance and reporting features. A crisis management dashboard's customization options, remote access capabilities, and ability to store historical data make it an indispensable tool for managing crises, enhancing future readiness, and upholding effective public relations.

5.1. Further recommendations and updates

While the dashboard helps with the understanding of the situation in real time, further improvement could be done to ideate a more comprehensive data-oriented channel that helps with the better analysis of any crisis. Expanding data coverage to include emerging variants and post-pandemic recovery metrics can be done to ensure a comprehensive view when improving a COVID-19 dashboard. Similarly, user customization would be enabled for users to personalize their dashboard view according to interests. Accuracy and timeliness could be enhanced by real-time data updates, particularly in dynamic scenarios such as pandemics. The dashboard's functionality is improved through integration with external data sources, such as economic indicators and travel restrictions, which provide insights into broader societal impacts.

5.2. Long-term relevance of the dashboard

The COVID-19 dashboard has the potential to develop into a versatile long-term health surveillance tool that can track other infectious diseases and health trends in addition to COVID-19. It is essential to inform the public about pandemics and get them ready for any medical emergencies. Furthermore, even after the pandemic subsides, the rich data on the dashboard can guide future research and public health policies. Crucially, it can support global health initiatives by keeping an eye on and reacting to new health risks across the globe, providing a thorough and proactive strategy for preserving public health globally.

6. Conclusion

Our analysis of the COVID-19 crisis undertaken through data exploration, predictive modeling, and Twitter sentiment analysis has unveiled critical trends and public perceptions. The pandemic's severity has lessened over time, likely due to vaccinations and improved treatments. The XGBoost predictive model proved highly accurate in forecasting hospitalizations, offering a strategic asset for healthcare planning. Sentiment analysis on Twitter highlighted a tangible link between rising case numbers and declining public mood, though the inability to access real-time data and potential biases should temper conclusions drawn from this analysis.

The innovative Crisis Management Dashboard for NSW Health encapsulates these findings in a dynamic, real-time interface, enhancing decision-making and situational awareness for health authorities. For future enhancement, more diverse data integration, advanced NLP techniques, and enriched predictive models are recommended. The developed tools and methodologies will continue to be relevant for health monitoring and crisis management, marking a significant contribution to public health resilience.

7. References

- Jalil, Z., Abbasi, A., Javed, A. R., Khan, M. B., Hasanat, M. H., Malik, K. M., & Saudagar, A. K. (2021). COVID-19 Related Sentiment Analysis Using State-of-the-Art Machine Learning and Deep Learning Techniques. *Frontiers in Public Health, 9*.
- Nandwani, P., & Verma, R. (2021, August 28). A review on sentiment analysis and emotion detection from text. *PubMed Central, 11*(1).
- Neal , D. T., Wood, W., Labrecque , J. S., & Lally , P. (2012, March). How do habits guide behavior? Perceived and actual triggers of habits in daily life. *Journal of Experimental Social Psychology, 48*(2), 492-498.
- Garcia, K., & Berton, L. (2021). Topic detection and sentiment analysis in twitter content related to COVID-19 from Brazil and the USA. *Appl Soft Comput*.
- Kaye-Kauderer, H., Feingold, J. H., Feder, A., Southwick, S., & Charney, D. (2021). Resilience in the age of COVID-19. *Crossmark, 27*, 166-178.

- Onyenwe , I., Mbeledogu , N., Onyedinma , E., & Nwagbo , S. (2020). The impact of political party/candidate on the election results from a sentiment analysis perspective using# anambradecides2017 tweets. *Soc Netw Anal Min, 10*(1), 1-17.
- Williams, S. N. (2023, May 6). COVID is officially no longer a global health emergency – here's what that means (and what we've learned along the way). *The Conversation*.
- World Health Organization. (2021). *Social media & COVID-19: A global study of digital crisis interaction among Gen Z and Millennials.*
- World Health Organization. (2023). *WHO Coronavirus (COVID-19) Dashboard.* Retrieved from World Health Organization: https://COVID19.who.int
- Twitter conversations predict the daily confirmed covid-19 cases. Applied Soft Computing, 129, 109603. https://doi.org/10.1016/j.asoc.2022.109603
- A Twitter narrative of the COVID-19 pandemic in Australia. Proceedings of the 20th International Conference on Information Systems for Crisis Response and Management. https://doi.org/10.59297/gqed8281