Sentiment Analysis:
Variations in the anxiety and depression levels during lockdown

Group members: Anne Mercado(aam9952), Fan Fan(ff2249), Sridevi Turaga (st4282), Sonali Mhatre (ssm10093), Sneha Trichy Shekar (skt9904)

Abstract
This study examines the fluctuations in anxiety and depression levels during the COVID-19 lockdown, focusing on how these changes are influenced by external events and news, and reflected through language used on social media. Utilizing datasets from Twitter and Yelp, which are widely used by people in daily life and provide genuine insights into public sentiment, this research employs a variety of analytical methods. Techniques such as decision trees, multinomial Naive Bayes, Support Vector Machines (SVM), clustering, time series analysis, and correlation analysis were implemented to explore the dynamics of mental health trends. Findings indicate that rates of anxiety and depression increased from early 2020 and decreased in 2021. People's negative feelings were exacerbated by the rising death rate and number of new cases, though this effect gradually diminished over time. Conversely, immunization efforts uplifted public emotion. Furthermore, the study found a correlation between people's anxiety and depression levels and political, news, and presidential information based on the frequency of words that appear on Twitter. These effects, however, are found to vary in intensity, suggesting complex interactions between public health dynamics and information dissemination.

Keywords
Covid-19, Anxiety and Depression Levels, Twitter, Yelp

Introduction
The COVID-19 pandemic has undeniably and unprecedentedly disrupted various aspects of daily life, ranging from professional work environments to personal social activities. This profound disruption has precipitated a spectrum of adverse mental health outcomes, as evidenced by numerous studies documenting the global mental health burden associated with the pandemic. Turna et al. (2021) highlight that the psychological impact of COVID-19 has been profound, contributing to a significant increase in anxiety, depression, and traumatic stress across diverse populations. Years after the initial outbreak, the lingering effects of the pandemic continue to manifest as persistent negative mental health outcomes. Boden et al. (2021b) suggest that these impacts are likely to persist, affecting mental health trends worldwide for the foreseeable future. This ongoing situation underscores the necessity for continued research and intervention to mitigate the long-term psychological effects of such a global health crisis.

With lockdowns and social distancing measures in force, social media has emerged as the

primary communication conduit for people across all demographics. It has swiftly evolved into an indispensable platform for the generation, dissemination, and consumption of information, utilized not only by individuals but also extensively by governments, organizations, and universities to relay critical updates to the public (Tsao et al., 2021). For many individuals, these virtual spaces have become essential in maintaining social connections with family and friends amidst physical isolation. However, this increased reliance on social media has also had unintended consequences. As users continually encounter updates on COVID-19, such as case counts and mortality rates, alongside the palpable emotional distress shared by others, a pervasive negative perception of the pandemic has taken shape. The barrage of distressing information and shared anxiety has, in turn, contributed to heightened levels of anxiety and depression among the population (Hong et al., 2021).

This study concentrates on the demographic of New Jersey residents, employing data harvested from online sources to analyze the sentiments expressed by individuals on social media. The objective is to delineate users' emotional responses throughout the pandemic, specifically comparing their mental states during and after the COVID-19 lockdown periods. By utilizing the data and methodologies to be discussed later, this research aims to provide a deeper understanding on how anxiety and depression levels have evolved over the course of the pandemic. Additionally, the study will identify specific patterns and characteristics observable in the data, offering insights into the broader psychological impacts of the pandemic on this particular demographic.

Literature Review

According to the World Health Organization (2022), the global prevalence of anxiety and depression surged by an alarming 25% during the first year of the COVID-19 pandemic. In the United States, nearly half of the participants surveyed in a National Institutes of Health study reported symptoms indicative of anxiety and depression disorders ('COVID-19 Mental Health Information and Resources,' n.d.). The impact was particularly pronounced among U.S. adults aged 18-29, where research conducted by Boston College found that rates of anxiety and depression escalated to 65% and 61%, respectively, within this demographic ('COVID-19's Toll on Mental Health,' n.d.). Further emphasizing the pandemic's deep psychological repercussions, a study by Kecojevic et al. (2020) on college students in New Jersey found that COVID-19 significantly exacerbated pre-existing mental health distress among the student population. This distress was largely attributed to increased academic pressures and life challenges intensified by the pandemic conditions. Additionally, data from the New Jersey Hospital Association's Center for Health Analytics, Research and Transformation (CHART, 2021) underscores the persistent impact on mental and behavioral health. The research indicates that, even during periods of stay-at-home orders, individuals continued to seek emergency department care for mental and behavioral health issues at rates comparable to or exceeding those before the pandemic. These

findings collectively illustrate the deep and ongoing psychological strains imposed by the pandemic, necessitating sustained attention and intervention.

Over the past decade, the impact of social media on human lives has escalated to extraordinary levels, achieving unprecedented reach and influence (Goel & Gupta, 2020). In today's internet-driven world, social media has become a pivotal channel for the dissemination and exchange of information. An increasing number of individuals now depend on these platforms not just for social interaction but also for accessing a diverse array of information, including critical health-related data (Park et al., 2024). The advent of the COVID-19 pandemic has further magnified this influence, solidifying the role of social media as an essential tool across various sectors. During the crisis, social media platforms have taken on crucial roles: they have been instrumental in spreading vital public health messages, facilitating educational continuity through distance learning, enabling remote work and monitoring, and significantly boosting healthcare and research capabilities. However, these benefits were not without challenges, notably the spread of disinformation. Open data and social media also expedited data collection and sharing among researchers, broadening the scope and accelerating the pace of studies related to COVID-19. These studies encompassed online surveys and the application of artificial intelligence for mining and synthesizing information from social media discussions about the pandemic (Goel & Gupta, 2020). Social media sentiment analysis emerged as a pivotal method for capturing the perspectives of various users and individuals (Barnhart, 2024). This technique has been extensively employed during the pandemic and continues to be utilized, providing invaluable insights into the public's reaction to this unprecedented global crisis.

Various studies have employed social media sentiment analysis to explore the impacts of the pandemic, differing in geographical scope, demographics, and platforms analyzed. Examples include Nemes and Kiss (2020) study on “Social Media Sentiment Analysis Based on COVID-19” which focus was primarily on the content of tweets related to COVID-19, without specifying or analyzing a specific demographic characteristic, “Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA” by Garcia and Berton (2021), which examines a large demographic across two countries and there is the “COVID-19 pandemic: a sentiment analysis: A short review of the emotional effects produced by social media posts during this global crisis” by Kumar et al. (2020b) that focused on analyzing tweets, retweets, and replies that included the #COVID19 hashtag over a specific period, from March 17 to 30, 2020. Building on these foundational studies, our research focuses on a more localized demographic, New Jersey, to understand how it compares with the broader landscape. We aim to analyze how specific events, such as vaccine rollouts and new public policy implementations, have influenced the local population. Additionally, our study conducts a temporal analysis of sentiments from the pre-lockdown phase through the year 2021, a year after the pandemic began, providing a nuanced view of the evolving emotional landscape.

Methods

Decision Tree:

Decision Trees are highly interpretable, as they provide a clear visualization of the decision-making process. A decision tree can illustrate the importance of certain features and how they lead to a particular classification, which can be valuable for understanding the factors driving sentiment or category assignment. We were focusing on interpretability and therefore preferred Decision Tree over Random Forest.

Multinomial Naive Bayes:

Firstly, Naive Bayes classifiers are known for their efficacy with text classification tasks, particularly due to their foundation in Bayes' theorem which calculates the probability of an event based on prior knowledge of conditions related to the event. Multinomial Naive Bayes is particularly apt for handling classification problems where features are represented as frequency counts; this is a natural fit for text data where we count how often each word appears. With our textual data from social media posts, the frequency of specific terms often reflects the intensity of sentiment, something that Multinomial Naive Bayes can leverage.

SVM:

SVM is widely used for classification and regression tasks for supervised learning. In the domain of text analysis, this technique is a powerful classifier, displaying a high degree of accuracy when handling high-dimensional data. For different sentiment categories SVMs give insights into features that contribute most to the separation of data points. The importance of certain specific words or phrases required in determining sentiment is understood while analyzing the weights assigned to the features. This study employed SVM to categorize text data, where it excels due to its ability to manage the extensive feature spaces that are typical in text applications, such as TF-IDF(term frequency-inverse document frequency) representations. The flexibility of tweaking the SVM kernel allows the model to capture complex patterns within the text, a capability that is crucial for discerning the subtle nuances of sentiment and intent.

Clustering:

Clustering is the unsupervised learning technique used here, to group data points that are together based on their characteristics or their features. In sentiment analysis, to identify patterns clustering algorithms are applied to clusters of sentiments with text data. Grouping of sentiments that are similar together, enables clustering of common themes and topics that are expressed by the users. This approach aids and analyzes emotional responses in uncovering underlying patterns during the COVID-19 lockdown period. Clustering algorithms such as K-means, DBSCAN, or hierarchical clustering are used to analyze the evolution of sentiment clusters over time.

Dimensionality Reduction:

For our SVM model and clustering analysis, we applied Singular Value Decomposition (SVD) to reduce the dimensionality of our TF-IDF vectors. We used TruncatedSVD over PCA, as it is good at dealing with large sparse datasets, commonly found in text data applications such as term frequency-inverse document frequency (TF-IDF) matrices used in natural language processing. Unlike PCA, TruncatedSVD does not center the data before computing the decomposition, which means it does not require the entire dataset to fit into memory and therefore can be applied to datasets that are too large to handle with PCA.

Other methods that we used : Time Series Analysis, Correlation Analysis

Data

Data collection:

To perform a detailed sentiment analysis and gauge public sentiment over time, we used various datasets. Publicly available yelp reviews dataset, _top1000trigrams from GitHub repository of tweets and Kaggle tweets dataset.

Data Exploration & Pre-processing:

Online yelp reviews contains 6,990,280 reviews ranging from 2005-08-29 to 2022-01-19. Of these, 260,897 reviews are related to New Jersey. Trigrams were fetched from GitHub between the dates 2020-03-23 and 2022-12-26. Additionally, a Kaggle tweets dataset was used, which is available between the dates 2020-07-24 and 2020-08-30. After merging these sources, we obtained 46,860 rows of text data for analysis, covering the period from January 1, 2020, to December 26, 2022.

We used the following dates as gathered from official announced websites to analyze different lockdown phases:

- pre_lockdown: 8th March 2020, Before the first state of emergency declaration
- during_lockdown: 9th March 2020 to 1st May 2020, From NJ state of emergency to the start of reopening
- post_lockdown: 2nd May 2020 to 10th Dec 2020, After reopening begins
- vaccine_rollout: 11th Dec 2020 to 18th Dec 2020, Vaccine dates
- new_normal: 19th Dec 2020

Preprocessing steps including removal of punctuations, special characters, url's, stop words, lemmatization and vectorization.

Labeling for dataset:

We used the NLTK library to derive polarity scores, which allowed us to categorize the reviews and obtain sentiment scores for our analysis. We used the lockdown period dates to label the

rows based on the dates to analyze trends. In addition to this, to enhance the granularity of our sentiment analysis across various stages of the COVID-19 pandemic - pre-lockdown, during lockdown, post-lockdown, and vaccine rollout, we developed a custom dictionary (refer Appendix) tailored to capture the distinctive lexicon that could have been related to these periods. The dictionary encompasses a curated collection of terms and phrases that are uniquely indicative of specific sentiments such as anxiety, depression, recovery, isolation, work and financial.

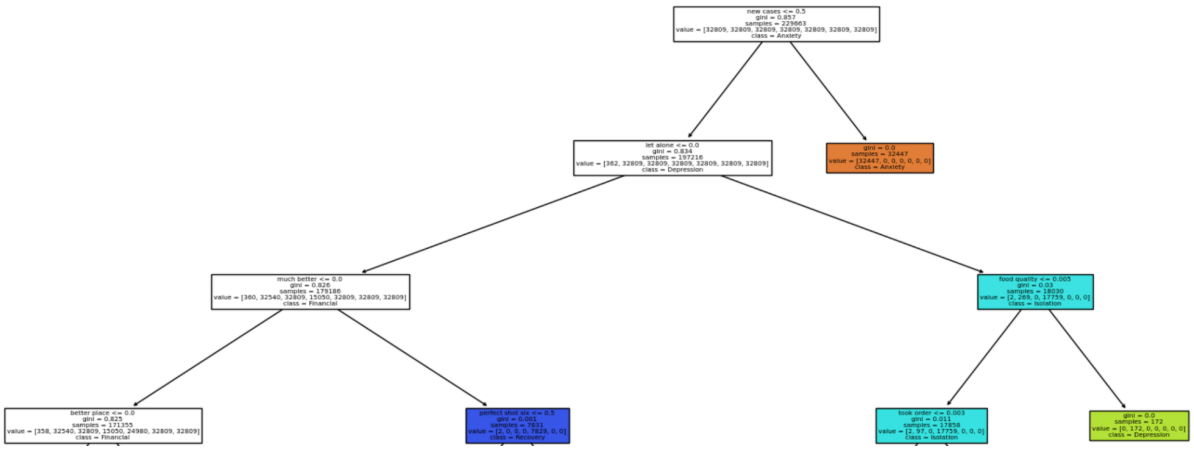
After categorizing the data into three distinct sentiment classes namely positive, negative and neutral, we found the distribution as follows: 'negative' sentiments totaling 23,985 instances, 'positive' sentiments encompassing 22,639 instances, and 'neutral' sentiments noticeably underrepresented at 236 instances. Acknowledging the potential biases an imbalanced dataset could introduce into our predictive models, we employed the Synthetic Minority Over-sampling Technique (SMOTE) to adequately resample the 'neutral' class and used this for our modeling purposes.

Vectorization:
We tried running the models with multiple variations of vectors. We used CountVectorizer and TF-IDF vectorization of bigrams and trigrams and also tried a new approach. We used both textual sentiment features from review text and categorical data representing different lockdown periods as inputs (X-features) to our predictive models. This combination of features allowed us to capture the nuanced sentiment changes associated with each stage of the COVID-19 lockdown. Our target variable (y), label_sentiment, included sentiment classifications, which helped us evaluate how well our model predicted sentiments.

Result

Decision Tree:
Decision tree model can tell us what words more relate with anxiety and depression emotion for us. To help us to improve our model for choosing words for the rest steps. This is one of the decision tree models we have. We can know what kinds of words we can divide them into the parts that express anxiety or depression. This decision tree visualization illustrates the process of classifying words according to different words, such as "new cases," "girl," and "food quality," thereby illuminating the decision path of the word features. The data is divided into segments starting at the top root node based on the "new cases" feature's threshold, which establishes whether the data flows to the left or right branch. The data is then processed further by several intermediary nodes using attributes like "better place" and "let alone," before flowing to leaf nodes that reflect certain categories like "Anxiety,", "Depression" and "Financial." The number of samples, gini impurity, and dominant category for each node are displayed, reflecting the significance of the characteristics and the effectiveness of the data segmentation in the decision tree. For instance, the data category at the right leaf node is extremely pure, as seen by the Gini

coefficient of 0. This analysis is useful for better feature selection and further model improvement since it not only clarifies the functioning of the decision tree but also, highlights critical elements that significantly affect the final prediction and what words we should choose for the features of future models.



Multinomial Naive Bayes:

The multinomial naive bayes yielded distinct performance results when trained with and without the Synthetic Minority Over-sampling Technique (SMOTE). Without SMOTE, our model demonstrated high precision and recall for the predominant 'Anxiety' sentiment, but it struggled significantly with the minority classes ('Depression', 'Financial', 'Isolation', 'Recovery', 'Work'), reflecting the classifier's bias towards the majority class. This resulted in an overall accuracy of 0.87, misleadingly high due to the disproportionate influence of 'Anxiety' labels. On the other hand, employing SMOTE markedly improved the recognition of 'neutral' sentiment and reduced the misclassification of 'Anxiety', achieving better balance across classes. The improvement is evidenced by a broader distribution of recall and precision across varied sentiments, albeit with persistence of underperformance in some minority classes.

	precision	recall	f1-score	support
Anxiety	1.00	0.99	0.99	8176
Depression	0.00	0.00	0.00	68
Financial	0.43	0.06	0.10	52
Isolation	0.00	0.00	0.00	14
Recovery	0.48	0.10	0.17	127
Work	0.21	0.06	0.09	83
neutral	0.67	0.97	0.79	852
accuracy			0.95	9372
macro avg	0.40	0.31	0.31	9372
weighted avg	0.94	0.95	0.94	9372

Precision: 0.943911283628918
Recall: 0.9523047375160051

With SMOTE + lockdown_period features

	precision	recall	f1-score	support
Anxiety	0.88	1.00	0.93	8176
Depression	0.00	0.00	0.00	68
Financial	0.00	0.00	0.00	52
Isolation	0.00	0.00	0.00	14
Recovery	0.00	0.00	0.00	127
Work	0.00	0.00	0.00	83
neutral	0.32	0.03	0.05	852
accuracy			0.87	9372
macro avg	0.17	0.15	0.14	9372
weighted avg	0.79	0.87	0.82	9372

Precision: 0.7941558350583235
Recall: 0.8725992317541613

Using n grams

The application of Naive Bayes to our sentiment analysis allowed us to classify public sentiment with a high degree of reliability, given the algorithm's suitability for text classification tasks. The Naive Bayes model, which inherently assumes feature independence, processed the frequency of word pairings and triplets within our corpus, capturing complex sentiment expressions beyond what individual words could express. The results obtained from this model pointed to specific linguistic patterns that signal varying emotional responses to policy measures during different lockdown periods. For example, the predominance of negative sentiment in the 'Pre-Lockdown' phase, as determined by the model, may signal initial uncertainty and fear among the populace. In contrast, a shift to more positive sentiment indicators in later stages like 'Vaccine Rollout' demonstrates increasing public confidence in response efforts.

Support Vector Models (SVM):

For our SVM model we applied Singular Value Decomposition (SVD) to reduce the dimensionality of our TF-IDF vectors. With SVD set to retain 100 components, we used the transformed data for the task. In the SVM pipeline, after scaling the data and applying SVD, we trained the model using a linear kernel. With nearly flawless precision, recall, and F1 score, our SVM model excels in the "Anxiety" category. It does not do well in the other categories, though, which could be because the "Anxiety" category has a considerably greater sample size than the other categories, which causes the model to prefer the majority of the categories in its predictions. This is something that needs work to avoid overfitting. The words or phrases that have the biggest influence on the model's predictions can already be found using our feature importance analysis. This aids in determining which characteristics have a strong correlation with particular sentiment groups to help us divide the group of the reviews and posts, and use the words to identify people's emotional variations in the different periods.

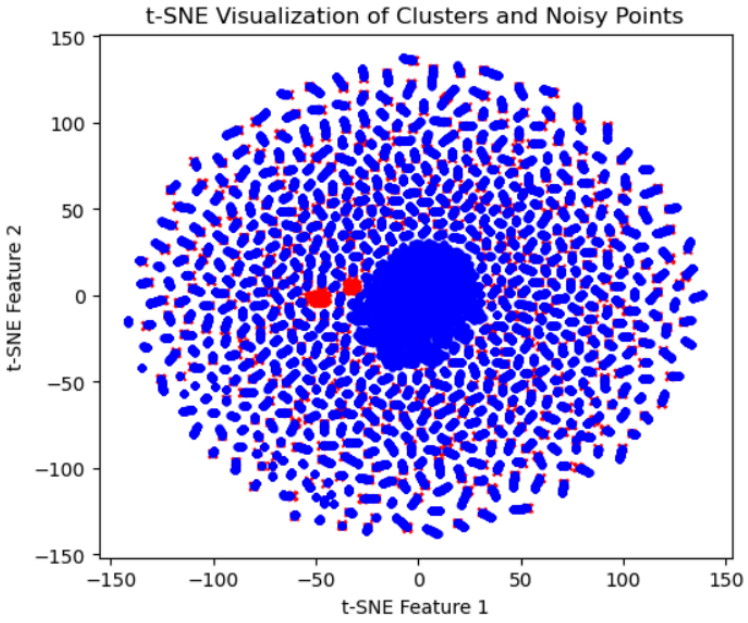
```
Accuracy: 0.9529449423815621
              precision    recall  f1-score   support

     Anxiety       1.00      0.99      0.99      8176
```

```
Top 10 features: ['fear doctor', 'care treatment', 'late outbreak', 'low rate', 'post use', 'dip best', 'announce new', 'pandemic
safe', 'beautiful also', 'try outside']
```

Clustering:

In our analysis, we utilized the DBSCAN algorithm for its proficiency in distinguishing outliers or 'noise points' from regular clusters of data points. The output of DBSCAN with eps=0.5, min_samples=5, includes cluster assignments for each data point. Points labeled as -1 are considered as noise points that do not belong to any cluster. Noise points might be associated with certain features that are less important or have higher variance. One such noise point highlighted a customer's negative experience with a moving company, citing unprofessional

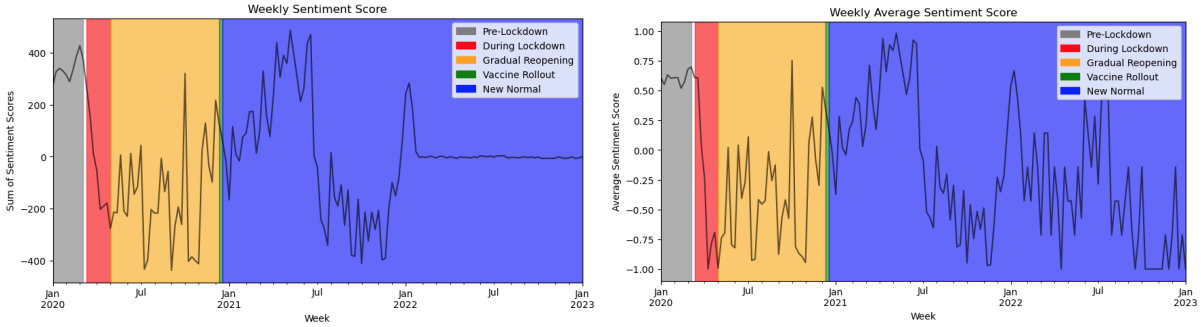
behavior, tardiness, and poor communication. This isolated entry, distinct in its negative sentiment and dissatisfaction, underscored the algorithm's ability to parse singular, impactful feedback from larger datasets. This insight demonstrates the value of unsupervised learning in extracting meaningful patterns and outliers from unstructured text data. The use of Truncated SVD, reducing the feature space to the 100 most significant components derived from the n-gram TF-IDF vectors and t-SNE further simplified our data, allowing DBSCAN to effectively identify clusters based on density and reveal outliers like the aforementioned customer review, which emphasized the diversity and individuality within our dataset. The evaluation of the clustering performance yielded a Silhouette Score of 0.749, indicating a strong cluster structure with well-separated sentiment groupings. Meanwhile, the Calinski-Harabasz Index, measuring cluster validity based on intra-cluster dispersion, was impressively high at 1964.36, further affirming the distinction among the clusters. A Davies-Bouldin Index of 0.895, while close to the ideal, suggests some inter-cluster similarity, which may warrant further investigation or parameter tuning.



INFO:root:Noisy Text 38377: pancake get everything else get love give place certainly come back pancake good conscience give bas...
INFO:root:Noisy Text 38416: place gem experienced bird owner nyc mom pop bird store loyal patron year close last year extensive ...
INFO:root:Noisy Text 38422: chef timothy witcher chopped food network champion unfortunately win streak end may seem harsh try g...
INFO:root:Noisy Text 38444: family order lunch special today mother recommend food past experience establishment get food miss c...
INFO:root:Noisy Text 38460: know say everyone bad day see good review guess yesterday bad day nail salon disappointed one techni...
INFO:root:Noisy Text 38502: go christmas eve first time go boyfriend several time server pleasant festive end seat towards back ...
INFO:root:Noisy Text 38528: hidden gem south jersey almost year travel near far get vietnamese specialty noodle boy son pretty p...
INFO:root:Noisy Text 38542: gripe restaurant server related go back well server review may get well good huge enclose circus ten...
INFO:root:Text in 10th Cluster: great place pizza many time always order full pie family pizza delicious service always spot girl professional phone person want delicious pizza place new confirm confirm covid case coronavirus va say coronavirus coronavirus disease coronavirus test virus coronavirus spread case covid coronavirus indiafightscorona het coronavirus wields coronavirus virus help unemployed record coronavirus health expert positive case response coronavirus death covid help u people die update crisis coronavirus coronaviruspandemic risk coronavirus sign coronavirus death population coronavirus death health amp case top health worker case per coronavirus lockdown kill pls help recovery via death back school bachchan test please help coronavirus trav□s case rise test positive case case amp death coronavirus argentina corona coronavirus coronavirus relief room doctor lockdown positivo coronavirus dy covid test positive case report positive covid open school coronavirus case india school open coronavirus could corona virus coronavirus test pandemic coronavirus trump coronavirus health department coronavirus high risk coronavirus infection positivos coronavirus get sick coronavirus guideline health care positive case increase despite coronavirus coronavirus concern refuse coronavirus excess death

Time Series Analysis:

Resampling the sentiment scores to a weekly frequency, we computed the sum of the sentiment scores for each week to visualize the trends in sentiment over time. As seen in the plot, during the lockdown of 2020, there is a noticeable dip in sentiment scores, indicating a period of potentially heightened negative sentiment. This trend is followed by a significant upturn in sentiment scores around 2021, suggesting a shift towards more positive or neutral sentiment.



Correlation Analysis:

Our statistical assessment explored potential relationships between public sentiment and COVID-19 impact, as quantified by new cases and deaths. The Pearson correlation between sentiment scores and new cases was statistically insignificant ($r = 0.0008$, p-value = 0.861), indicating no linear relationship. Conversely, the correlation with new deaths was negative, albeit weak ($r = -0.087$), but highly significant ($p < 0.001$), suggesting a slight inverse association where an increase in new deaths weakly corresponds with more negative sentiment scores. Additionally, we performed a lagged correlation analysis to observe the relationships over 14 successive time intervals. The findings exhibited a consistent trend: as the time lag increased, the sentiment scores' correlation with new cases and deaths fluctuated albeit remained negative and weak (ranging from about -0.030 to -0.052 for cases and about -0.026 to -0.045 for deaths). This pattern suggests sentiments become slightly more negative as new cases and deaths rise, though the impact diminishes slightly over time. These correlations, though relatively stable, reflect minimal changes in sentiment with respect to the evolving pandemic situation, suggesting that public sentiment is affected by the pandemic's progression but only to a limited extent.

Topic Modeling:

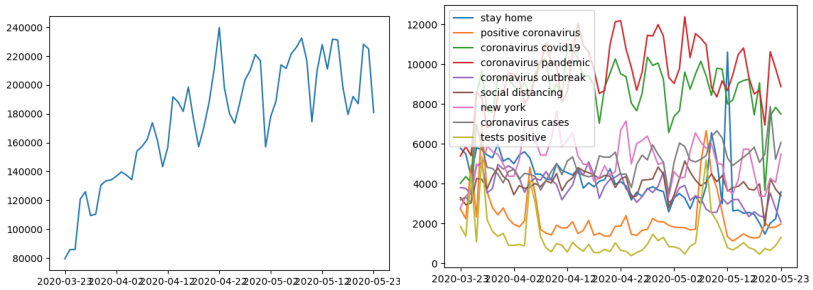
Topic modeling, when combined with sentiment analysis, offers a strong way to understand and interpret large-scale text conversations. Our exploratory topic modeling analysis using Latent Dirichlet Allocation (LDA) with a CountVectorizer highlighted key thematic clusters within the corpus of COVID-19 related text data. Topic 0, encapsulates terms central to the public discourse on the health crisis such as 'coronavirus', 'COVID-19', 'cases', and 'deaths', suggesting a focus on

the spread and impact of the virus alongside a notable mention of political leadership ('trump'). Topic , echoes these concerns with a slight variation possibly depicting public health responses including 'tests' and 'lockdown'. Contrastingly, Topic 2 diverges to the restaurant terms (which is valid as we are using yelp dataset) centering on culinary experiences during the pandemic. Topic 3, captures aspects of day-to-day life with references to 'school', 'work', and 'new jersey', indicating discussions around routine adaptations due to the pandemic. Lastly, Topic 4 seems to revisit and reinforce the concerns found in Topics 0 and 1, with an additional focus on 'patients', which may relate to healthcare discussions.

```
Topic 0:
coronavirus 19 covid cases covid19 new death positive trump deaths virus health crisis confirmed help

Topic 1:
coronavirus cases covid19 health deaths 19 covid positive virus new death died tests die lockdown

Topic 2:
food good place ordered chicken like great order got time really delicious restaurant sauce cheese

Topic 3:
new said school like told time going people called jersey need car schools know work

Topic 4:
coronavirus cases 19 covid covid19 health deaths positive virus help death died new patients update
```

Visualization of github twitter words

This graph represents the period from March 2020 to May 2020, illustrates the trend in talks or searches for various COVID-19 virus related topics. Each line represents the trend of this word during the pandemic period. We can see there is obviously fluctuation for different words, these reflect the impact of epidemic-related information on public attention. For instance, the term "stay home" exhibits multiple spikes in the graph, which could be due to the local government's implementation of the home stay order or the different timing of relevant news articles.”Positive coronavirus’ and ‘coronavirus cases’ show similar trend on the graph ,this may indicate that these two topics are correlated in public topic. “New York” is also a word that is mentioned many times in the period. This reflects people in New York’s attention and how the public reflects on New York when it was an outbreak-hit area. For the ‘covid 19’ , this may be coordinated with the public's response and media coverage as the number of incidents rises. This analysis can help governments and health departments to know how people reflect their emotions by words and help us to do future analyses.



Policy Related

The actionable insights derived from this sentiment analysis could inform strategic policy decisions within the mental health landscape. The outcomes from the decision tree model, which has identified features such as 'new cases' and 'let alone' as being most significant, can inform policy by flagging the terms most pertinent to consumer sentiment in current discussions. For instance, a high importance attributed to 'new cases' suggests the need for policies addressing consumer concerns about public health dynamics. Concurrently, the top SVM features such as 'fear doctor' and 'care treatment' suggest consumer focus on healthcare and personal well-being which could influence the design of people support and service policies. The presence of 'late outbreak' and 'pandemic safe' among the key features indicates the priority people place on real-time updates and safety measures during health emergencies, signaling a need for transparent communication policies during crises. Additionally, terms like 'beautiful also' and 'try outside' highlight elements that may form part of a positive experience, suggesting a possible expansion towards outdoor experiences. Additionally, the emergence of 'noisy' texts identified by DBSCAN in our dataset could serve as a key policy point where regulations or services may need adjustment to address uncommon but impactful issues. Similarly, the emergence of health-related topics, including terminology around 'coronavirus', 'COVID-19', 'cases', 'deaths', and 'health crisis', underscores the prevalent public concern regarding the pandemic's progression and implications. These areas could be focal points for policy intervention, requiring the provision of clear communication, enhanced health services, and support mechanisms. By tying the output of the Naive Bayes model to these distinct time frames, we can assess the effectiveness of policy communication and the public's receptiveness to policy decisions at critical junctures. Such data-driven insights enable policymakers to understand the impact of their actions on public sentiment and adjust their strategies accordingly to build trust and achieve better outcomes in managing public crises.

Discussion

In reviewing our analytical part, we encountered challenges in sourcing Twitter data due to restrictions on our developer account. This limitation constrained our ability to conduct a more detailed sentiment analysis. Additionally, our primary text data source, Yelp, primarily focuses on restaurant reviews. This posed a challenge when adapting the data for sentiment analysis related to mental health. Our models demonstrated that using bigrams and trigrams for vectorization outperformed the TF-IDF Word2Vec methodology. This is likely because n-grams can capture contextual cues and word patterns more effectively, thus improving sentiment analysis results. This finding led us to prioritize the use of n-grams in our analysis. A challenge we faced during the project was the computational power required for lemmatizing and vectorizing the data and integrating BERT embeddings that leverage more sophisticated natural language. We could only apply SMOTE to the Naive Bayes model, as it did not work effectively with other models, which were trained using TF-IDF gram vectors from the lemmatized text.

Furthermore, our custom sentiment dictionary, which was crucial for labeling mental health sentiments, presents an opportunity for improvement through collaboration with domain experts. This collaboration could enhance our classification performance.

An acknowledged limitation of our study was the absence of demographic data for the texts. This raises the consideration that the prevalence of sentiment expressions in data may not be universally representative. It is possible that those experiencing anxiety may be more inclined to write reviews.

Conclusions

In conclusion, this study investigated the rates of anxiety and depression during and after the COVID-19 lockdown period utilizing ‘natural language processing’ techniques and ‘time series analysis.’ Advanced machine learning techniques were employed to perform sentiment analysis, revealing valuable insights into the psychological state of individuals as they expressed their experiences online. The application of Decision Trees and Support Vector Machines facilitated a deeper understanding of the prevailing emotional landscape, as reflected in textual data. The use of TruncatedSVD and StandardScaler preprocessing techniques demonstrated their effectiveness in enhancing model performance and interpretability despite the inherent high-dimensionality of text data. Additionally, clustering helped to identify distinct groups within the data, revealing patterns in sentiment that correspond to different mental health indicators. While the findings of this study are significant, highlighting specific text contexts associated with mental health indicators, we acknowledge areas for improvement in our methodology. The accuracy of our models could be further enhanced through more extensive hyperparameter tuning, which could potentially uncover more nuanced insights into the emotional states captured in the data. Moreover, ensuring a more balanced representation in our dataset could improve the generalizability and reliability of our findings.

Looking forward, there is substantial scope for future studies to build on our work. We recommend that subsequent research should not only tune the existing analytical models but also extend the application to a variation of demographics and broader datasets. This would open the research for a more comprehensive understanding of psychological impacts of major global events such as the COVID-19 pandemic. By continuing to explore the untapped potential of machine learning in data analytics, specifically sentiment analysis, researchers can further contribute valuable insights to fields such as public health, ultimately providing more effective mental health interventions and policies.

References

Barnhart, B. (2024, April 10). *The importance of social media sentiment analysis (and how to conduct it)*. Sprout Social. https://sproutsocial.com/insights/social-media-sentiment-analysis/

Boden, M. T., Zimmerman, L., Azevedo, K., Ruzek, J. I., Gala, S., Magid, H. S. A., Cohen, N., Walser, R. D., Mahtani, N., Hoggatt, K. J., & McLean, C. P. (2021a). Addressing the mental health impact of COVID-19 through population health. *Clinical Psychology Review*, *85*, 102006. https://doi.org/10.1016/j.cpr.2021.102006

COVID-19 Mental Health Information and Resources | National Institutes of Health. (n.d.). NIH COVID-19 Research. https://covid19.nih.gov/covid-19-topics/mental-health

COVID-19's toll on mental health. (n.d.). https://www.bc.edu/bc-web/bcnews/campus-community/faculty/anxiety-and-stress-spike-during-pandemic.html

Garcia, K., & Berton, L. (2021). Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. *Applied Soft Computing*, *101*, 107057. https://doi.org/10.1016/j.asoc.2020.107057

Goel, A., & Gupta, L. (2020). Social media in the times of COVID-19. *Journal of Clinical Rheumatology*, *26*(6), 220–223. https://doi.org/10.1097/rhu.0000000000001508

Hong, W., De Liu, R., Ding, Y., Fu, X., Zhen, R., & Sheng, X. (2021). Social media exposure and college students' mental health during the outbreak of COVID-19: the mediating role of rumination and the moderating role of mindfulness. *Cyberpsychology, Behavior and Social Networking*, *24*(4), 282–287. https://doi.org/10.1089/cyber.2020.0387

Kecojevic, A., Basch, C. H., Sullivan, M., & Davi, N. (2020). The impact of the COVID-19 epidemic on mental health of undergraduate students in New Jersey, cross-sectional study. *PloS One*, *15*(9), e0239696. https://doi.org/10.1371/journal.pone.0239696

Kumar, A., Khan, S. U., & Kalra, A. (2020). COVID-19 pandemic: a sentiment analysis. *European Heart Journal*, *41*(39), 3782–3783. https://doi.org/10.1093/eurheartj/ehaa597

Morin, C. M., Bjorvatn, B., Chung, F., Holzinger, B., Partinen, M., Penzel, T., Ivers, H., Wing, Y. K., Chan, N. Y., Merikanto, I., Mota‐Rolim, S., Macêdo, T., De Gennaro, L., Léger, D., Dauvilliers, Y., Plazzi, G., Nadorff, M. R., Bolstad, C. J., Siemiński, M., . . . Espie, C. A. (2021). Insomnia, anxiety, and depression during the COVID-19 pandemic: an international collaborative study. *Sleep Medicine*, *87*, 38–45. https://doi.org/10.1016/j.sleep.2021.07.035

NJHA's Center for Health Analytics, Research and Transformation (CHART). (2021b). The other epidemic: the mental health toll of COVID-19. CHART Bulletin Series, VOL.20. https://www.njha.com/media/638857/Mental-Health-Toll-COVID-19.pdf

Park, B., Jang, I. S., & Kwak, D. (2024). Sentiment analysis of the COVID-19 vaccine perception. *Health Informatics Journal*, *30*(1). https://doi.org/10.1177/14604582241236131

Tsao, S.-F., Chen, H., Tisseverasinghe, T., Yang, Y., Li, L., School of Public Health and Health Systems, Faculty of Science, Seneca Libraries, University of Waterloo, Seneca College, & Butt, Z. A. (2021). What social media told us in the time of COVID-19: a scoping review. *The Lancet Digital Health*, *3*, e175–e194. https://doi.org/10.1016/S2589-7500(20)30315-0

Turna, J., Zhang, J., Lamberti, N., Patterson, B., Simpson, W. S., Francisco, A. P., Bergmann, C. G., & Van Ameringen, M. (2021). Anxiety, depression and stress during the COVID-19 pandemic: Results from a cross-sectional survey. *Journal of Psychiatric Research*, *137*, 96–103. https://doi.org/10.1016/j.jpsychires.2021.02.059

World Health Organization: WHO. (2022, March 2). COVID-19 pandemic triggers 25% increase in prevalence of anxiety and depression worldwide. *World Health Organization*. https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide

Appendices

Contributions:

Anne Mercado: Data Collection, Data Methodology, Presentation, Introduction, Literature Review, Conclusion. (20.5%)

Fan Fan: Data Collection, Data Clean-up, Modeling, Data Analysis, Presentation, Abstract, Methods, Results. (20.5%)

Sridevi Turaga: Data Collection, Data Clean-up, Modeling, Data Analysis, Presentation, Outline Methods, Results, Conclusion. (25%)

Sonali Mhatre: Data Collection, Literature Review. (17%)

Sneha Trichy Shekar: Data Collection, Data Clean-up, Data Analysis, Methods. (17%)