**Navigating Covid-19 Research Data: An In-Depth Exploration Of Covid-19**

Shouzhifan Zhu

**Background and Motivation**

The COVID-19 pandemic, declared by the World Health Organization (WHO), has reshaped global health and profoundly impacted social dynamics, especially through shifts in public sentiment. As the virus spread, social media platforms, particularly Twitter, emerged as vital channels for public discourse. Early in the pandemic, emotions such as fear dominated public conversation, driven by uncertainty, panic, and concerns about the capacity of healthcare systems to cope with the crisis. However, as the pandemic continued, these emotions evolved, and anger began to take center stage, reflecting growing frustration with lockdowns, shortages, and government responses (Lu et al., 2024).

Understanding these emotional shifts is essential for effective public health communication. By analyzing sentiment on social media, real-time insights into public perceptions can be gained, allowing policymakers to adapt their messages accordingly. For instance, fear and anger were often linked to issues like medical supply shortages, xenophobia, and strict governmental measures. At the same time, emotions of sadness and joy emerged, associated with personal loss and collective gratitude for healthcare workers. These emotional shifts provide valuable insights into how populations react to crises, which can help guide communication strategies throughout the duration of global health emergencies.

This project aims to emphasize the importance of monitoring and analyzing these sentiment changes over time, as understanding the psychological impacts of pandemics can lead to more nuanced, effective messaging. The insights gained from sentiment analysis not only inform crisis management but also contribute to the mental well-being of populations during turbulent times (Lu et al., 2024). Therefore, the exploration of emotional narratives in COVID-19-related discussions on social media is crucial for improving future public health responses.

**Research Question**

The research question for this project is to explore:

(1). How do temporal factors (e.g., day of the week, month, weekend vs. weekday) influence the volume of COVID-19-related tweets?

(2). What are the geographical patterns of tweet volume across different countries and cities, and how do these patterns impact the prediction of tweet activity?

**Application Scenarios**

The findings from this project on the temporal and geographical variability in COVID-19-related tweets, as analyzed through SHAP, namely Shapley Additive Explanations, have significant real-world applications across public health communication, sentiment analysis, and predictive modeling. By understanding the factors that influence tweet volumes, this project can help healthcare providers develop more effective communication strategies and tailor public health interventions to specific regions and periods. For example, if SHAP analysis shows that tweet volumes peak on weekends, health organizations could schedule important announcements or updates during these high-engagement periods to ensure maximum reach. Additionally, by identifying geographical trends, policymakers can allocate resources and focus their efforts on regions with higher levels of social media engagement, ensuring that interventions are both timely and relevant.

This project can also be used to inform crisis management strategies by offering real-time insights into public sentiment. By examining the SHAP values that highlight key features like the day of the week, month, or geographical location, health officials can gauge how different regions are reacting to the pandemic and adjust their messaging accordingly. For instance, if SHAP analysis reveals heightened frustration or fear in certain regions, public health authorities can adjust their strategies to address these concerns, providing targeted messaging or increasing support where it is most needed. This could help mitigate public unrest and improve trust in public health measures, particularly during critical phases of the pandemic.

Furthermore, this project’s insights into tweet volume patterns can assist in predictive modeling and geospatial research. By using the temporal and geographical features highlighted in the SHAP analysis, researchers can predict future trends in tweet activity, which could correlate with real-world events such as outbreaks or changes in government policies. This predictive capability can support decision-makers in preparing for potential surges in public engagement and ensuring that resources are distributed efficiently. For example, countries with consistently high tweet volumes may require more urgent attention and allocation of resources, while regions with lower engagement might need targeted efforts to increase awareness or participation in health initiatives.

Ultimately, this project can bridge the gap between social media analytics and real-world applications, enabling a more data-driven approach to public health interventions. By understanding the factors that influence social media discussions around COVID-19, health organizations can improve their ability to communicate effectively with the public, address emerging concerns, and adapt their strategies as the pandemic evolves. This approach can also be applied to other global health crises, offering a valuable framework for real-time monitoring and response.

**Methodology**

**Dataset**

The primary dataset consists of COVID-19-related tweets collected over a period of time using over 800 predefined keywords and hashtags associated with COVID-19. The dataset includes both geotagged tweets (378K tweets) and tweets with location data but without explicit geotags (5.4 million tweets). This comprehensive dataset offers insights into the global conversation around the pandemic, providing a rich source for analyzing the temporal and geographical distribution of tweet volumes. Particularly, we are using a subset of the GeoCov19 dataset from February 1 to February 10, 2020, due to the massive file size.

**Visualization Techniques**

The methodology used in this project integrates advanced visualization techniques to analyze and present the temporal and geographical distribution of COVID-19-related tweet volumes, with a particular focus on SHAP analysis of these features. Visualization plays a crucial role in making complex relationships in the dataset interpretable and accessible, especially in understanding how different temporal and geographical factors contribute to the prediction of tweet volumes.

Marks and channels were carefully selected to ensure clarity and enhance interpretability. For example, in the SHAP summary plot, dots are used as marks to represent individual features, with the x-axis encoding the SHAP value for each feature, and the y-axis listing features such as “Is Weekend” and “Month.” This design leverages position along the axes as a channel for quantitative comparison, enabling users to quickly assess the importance of various features. Additionally, color intensity is used as an auxiliary channel to highlight the significance of certain features, making it easier for users to identify key data points and understand which features have the greatest impact on the model’s predictions.

Geospatial data was incorporated into the visualizations using a GeoJSON map, which shows tweet volume by country and region. In this map, tweet volume is represented through color gradients, with darker colors indicating higher tweet volumes. This approach follows best practices for spatial data representation, where geographic regions are color-coded to reveal disparities in tweet activity across the globe. The map includes interactivity, allowing users to click on individual regions for more detailed insights into tweet volumes and their corresponding SHAP values. This interactivity empowers users to explore the data in a more granular way, providing deeper insights into regional patterns of COVID-19-related tweet activity.

Data abstraction was a key element in the visualization design to ensure that each tool supports specific analytical needs. For example, the SHAP beeswarm plot was designed to highlight the relative importance of temporal and geographical features in tweet volume predictions. The features are ranked along the y-axis based on their SHAP values, and the color coding highlights the importance of each feature in the overall prediction. This approach allows for a high-level overview of feature contributions while also providing the capability for detailed examination of individual data points.

The visualizations combine multiple data sources, such as geotagged tweets and temporal metadata, to provide a comprehensive view of tweet activity. By integrating these diverse datasets, the visualizations capture complex relationships between time, location, and tweet volume, revealing more nuanced patterns and trends. This combination of datasets helps to paint a fuller picture of the factors influencing the volume of COVID-19-related tweets, providing more robust insights.

Color theory and design considerations were also integral to the creation of the visualizations. For example, in the SHAP summary plot, color is used to differentiate between higher and lower feature values, ensuring that the visual representation remains clear and easy to interpret. In the GeoJSON map, color gradients represent varying tweet volumes across countries, allowing users to visually distinguish between regions with high and low tweet activity. The color design is implemented to ensure perceptual clarity and to avoid any confusion or misinterpretation, following best practices in color used for data visualization.

In conclusion, the visualization techniques used in this study effectively represent the complex temporal and geographical dynamics of COVID-19-related tweet activity. By selecting appropriate marks and channels, employing interactive elements, and using color strategically, the visualizations ensure that users can derive meaningful insights from the data, whether for high-level analysis or detailed exploration.

### **Advanced Tools**

**Interactive Visualization Tools**

In this project, we employ advanced interactive visualization tools to facilitate an in-depth understanding of the temporal and geographical patterns in COVID-19-related tweet volumes. These visualizations, which integrate SHAP analysis and machine learning models, offer both high-level insights and detailed, user-driven exploration of the data.

A key interactive tool is the “GeoJSON” map, which visualizes tweet volumes across different countries and regions. This map provides a global perspective on COVID-19 tweet activity, with each country color-coded based on tweet volume. Darker shades indicate higher tweet volumes, allowing users to quickly identify regions with significant social media engagement. By clicking on a specific country, users can access detailed information on the tweet volume and SHAP values for that region, offering insights into both geographical trends and the underlying factors driving tweet activity.

This map allows for a dynamic exploration of tweet distributions, and the ability to click on countries provides users with deeper insights into how COVID-19 discussions vary across the globe. By combining geographical and temporal data in this interactive map, users can observe the correlation between tweet volumes and the pandemic’s impact in different regions.

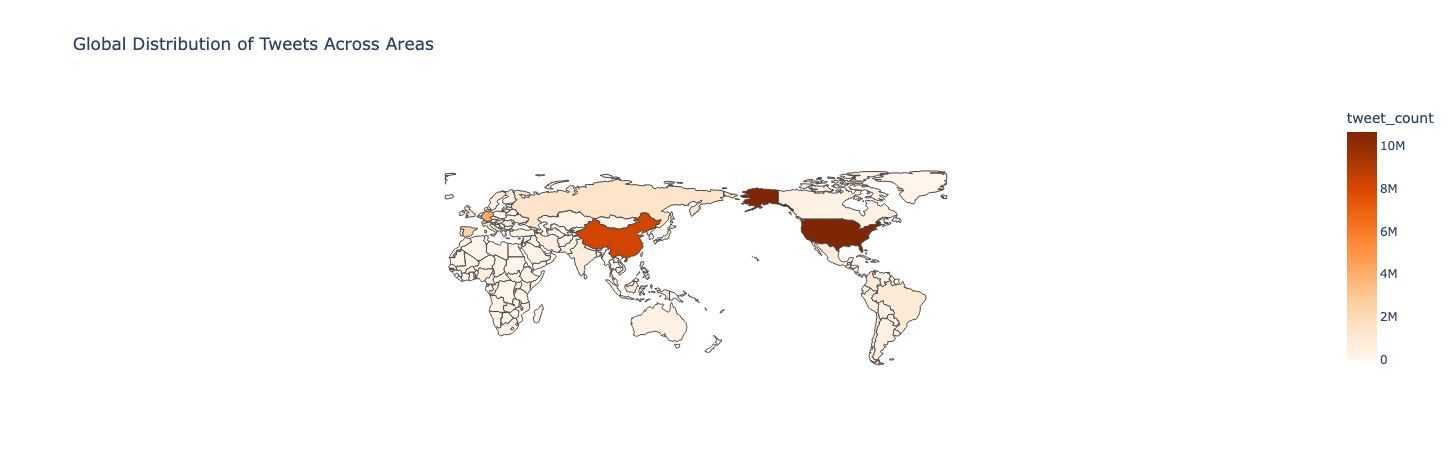


Figure 1: Global Distribution of COVID-19-related Tweets

**Machine Learning**

To predict tweet volumes based on temporal and geographical features, we used multiple machine learning models, including Logistic Regression and XGBoost. These models are crucial for identifying and quantifying the factors that influence tweet activity, providing both predictive power and interpretability.

Logistic Regression is employed to model the linear relationship between features, such as whether a tweet was posted on a weekend (“Is Weekend”) and tweet volume. This model allows us to understand how specific features, like weekends, directly impact the volume of tweets. Despite its simplicity, Logistic Regression provides interpretable results, with coefficients indicating the strength of each feature’s influence on the prediction.

In contrast, XGBoost is used to capture more complex, non-linear relationships between features through an iterative process of decision tree construction. XGBoost is particularly well-suited for handling high-dimensional data, identifying subtle interactions between features, and mitigating overfitting through regularization techniques. This makes it an ideal tool for modeling the relationships between various features and tweet volume.

The SHAP (Shapley Additive Explanations) analysis is integrated with these machine learning models to provide interpretability and transparency in the predictions. SHAP values quantify how much each feature contributes to the model’s output, making it easier to understand which factors are driving the model’s predictions. For example, the SHAP Summary Plot for Temporal Features (Figure 2) reveals the impact of time-based features, such as weekends, weekdays, and months, on tweet volume predictions. This plot shows that “Is Weekend” has the strongest influence on tweet activity, with tweets posted on weekends contributing significantly to higher tweet volumes.

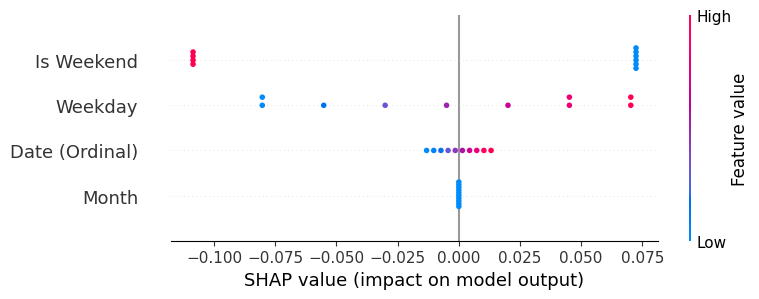


Figure 2: SHAP Summary Plot for Temporal Features

Similarly, the SHAP Summary Plot for Geographical Features (Figure 3) visualizes the influence of geographical factors, such as country and city, on tweet volumes. The plot shows that countries with higher social media engagement, like the United States (USA) and China (CHN), have a stronger positive impact on the model’s predictions. This highlights the regional disparities in tweet activity, with some countries contributing significantly more to the overall tweet volume than others.

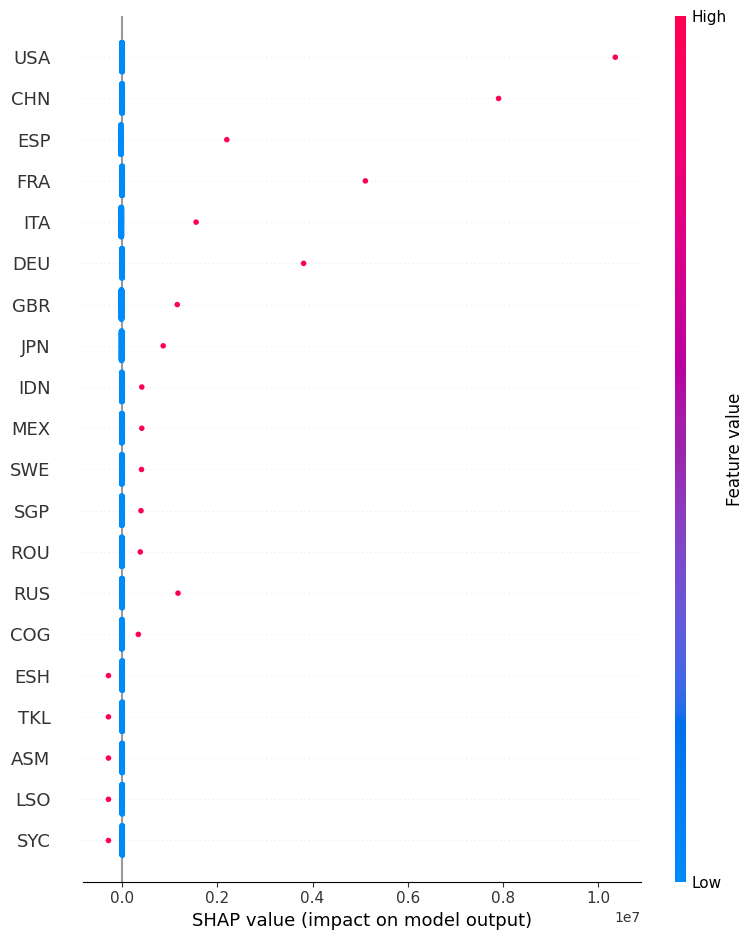


Figure 3: SHAP Summary Plot for Geographical Features

By integrating machine learning models with SHAP analysis, we are able to not only predict tweet volumes but also explain the role of temporal and geographical features in shaping those predictions. These advanced tools offer a clear, interpretable way to identify which factors are most influential, helping researchers and policymakers understand the dynamics of COVID-19-related social media discussions and how they vary over time and across regions.

### **Results**

The visualizations and analyses conducted in this study offer critical insights into the distribution and dynamics of COVID-19-related tweet volumes. The SHAP summary plots (Figures 2 and 3) are particularly valuable in understanding the significance of temporal and geographical features in predicting tweet volumes. The SHAP summary plot for temporal features (Figure 2) highlights the considerable impact of the “Is Weekend” feature, where tweets posted on weekends contribute significantly to higher tweet volumes. This insight is in line with the hypothesis that social media engagement peaks during weekends. On the other hand, features like “Weekday” and “Date (Ordinal)” exhibit much smaller impacts, suggesting that tweet volume is less influenced by specific weekdays or dates throughout the year. This highlights the critical role of weekends in driving social media activity during the pandemic.

The SHAP summary plot for geographical features (Figure 3) presents similar insights, revealing the dominant influence of countries such as the United States (USA) and China (CHN) on the overall tweet volume. These countries, with their large social media user bases, have a significant positive impact on the model’s predictions. Other countries with substantial social media activity, like Spain (ESP) and France (FRA), also contribute notably to the tweet volume, while smaller nations or those with lower engagement, such as Seychelles (SYC) and Tokelau (TKL), show minimal impact. This geographic variation underscores the disparities in global tweet engagement, likely influenced by factors such as population size, social media penetration, and COVID-19 case prevalence.

The interactive map (Figure 1) further emphasizes these geographical disparities, offering a global view of tweet distribution. The map visualizes tweet volumes across different countries, with darker shades representing higher tweet activity. It reveals that countries like the USA and China not only dominate in terms of tweet volume but also that regions with high population density and widespread social media usage show consistently high levels of COVID-19-related discussion. In contrast, regions like parts of Africa and some small island nations display lower levels of tweet activity, potentially reflecting lower social media penetration or different levels of public concern about the pandemic.

These visualizations reveal critical patterns in the data, showing that both temporal factors (weekends) and geographical factors (high tweet activity in populous countries) strongly influence COVID-19-related tweet volumes. While the temporal features reveal predictable patterns, the geographical features emphasize the disparity in social media engagement across regions.

The results highlight the usefulness of SHAP analysis in explaining how different features contribute to tweet volume predictions. By focusing on temporal and geographical data, this study provides actionable insights into the dynamics of social media discussions during the pandemic, with implications for public health communication strategies. Specifically, the clear patterns observed in the SHAP plots indicate that targeting weekend periods for communication efforts might be more effective, while understanding geographical disparities can help focus efforts in regions with higher levels of social media engagement.

In conclusion, this study successfully combines machine learning with SHAP analysis to uncover important trends in COVID-19-related social media activity. The results underscore the importance of considering both temporal and geographical factors when analyzing the spread of information and public sentiment during a global health crisis. These findings can inform future strategies for managing public health communications and improving social media engagement in different regions.

This research contributes significantly to the academic community by combining machine learning models and SHAP analysis with interactive data visualizations to explore the dynamics of COVID-19-related tweet volumes. Traditional methods of analyzing social media data often rely on static visualizations, which may limit the depth of insights that can be drawn from the data. By integrating interactive tools, such as the SHAP summary plots and global distribution maps, our study advances the field by making complex datasets more accessible and interpretable. This approach allows researchers, policymakers, and the public to better understand how temporal and geographical factors influence the spread of information on social media during the pandemic.

**Intellectual Merit and Practical Impacts**

A unique aspect of our study is the use of SHAP analysis to explain the impact of temporal and geographical features on tweet volumes. While many studies focus solely on statistical models to predict tweet activity, incorporating SHAP analysis provides transparency and interpretability, enabling users to understand the contributions of specific features in the model’s predictions. By offering detailed visualizations that show the relative importance of features such as “Is Weekend” and “Month,” we provide a more nuanced understanding of how these factors drive social media engagement. This combination of machine learning and SHAP analysis sets our work apart, offering a methodology that can be applied to a wide range of social media studies, not just in public health but in other fields such as political science, marketing, and communications.

This research aligns with recent advancements in computational tools and machine learning, similar to breakthroughs in AI applications for global health and epidemiology. By integrating temporal and geographical data into predictive models, our work addresses a significant gap in understanding the dynamics of public discourse across different regions and times. The ability to explain the impact of these factors on tweet volumes not only improves the accuracy of social media prediction models but also provides valuable insights into the regional disparities in social media engagement during health crises. Furthermore, the interactive tools we developed, such as the global tweet distribution map, allow users to explore tweet data in a way that is both informative and accessible to a broad audience. These innovations contribute to the democratization of social media data analysis, making it available to interdisciplinary scholars, healthcare professionals, and the general public.

The practical implications of this research are vast, especially in public health communication and crisis management. By identifying regions with high levels of tweet activity and understanding how temporal features, such as weekends or specific months, influence engagement, this study offers actionable insights for optimizing health communication strategies. For example, public health officials can target communication efforts during periods of high social media engagement, such as weekends, to maximize outreach. Additionally, understanding geographic disparities in tweet volumes helps policymakers tailor interventions to regions with more active public discourse, ensuring that efforts are more focused and effective.

On a broader societal scale, this study promotes inclusivity by highlighting the role of geographic and temporal factors in social media engagement. The interactive visualizations provide a clear picture of how the pandemic was discussed across different regions, encouraging policymakers and health organizations to consider regional differences when planning interventions. Moreover, by making complex social media data more accessible, this research ensures that insights are not confined to experts but can be shared with educators, clinicians, and even the general public, fostering a more inclusive approach to health communication and decision-making.

Ultimately, this project bridges the gap between social media data analysis and real-world public health applications. By providing a detailed, interactive exploration of the temporal and geographical factors that drive COVID-19-related tweet volumes, we contribute to a future where data-driven decision-making is more inclusive, accessible, and globally relevant. The findings of this study not only improve our understanding of social media’s role in pandemic discourse but also contribute to more effective, region-specific communication strategies that can help mitigate the impact of future health crises.

**Supplementary Materials**

GitHub repository: <https://github.com/ManchesterBlue/Info301_Final>.

**References**

Natural Earth. (2020). *Admin 0 - Countries (1:110m scale)*. Retrieved from <https://www.naturalearthdata.com/downloads/110m-cultural-vectors/110m-admin-0-countries/>.

Qatar Computing Research Institute (QCRI). (2020). *GeoCoV19: A Dataset of Hundreds of Millions of Multilingual COVID-19 Tweets with Location Information*. Retrieved from <https://crisisnlp.qcri.org/covid19>.

Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). (2020). *COVID-19 Dashboard*. Retrieved from <https://coronavirus.jhu.edu/>.

Abd-Alrazaq, A., Alhuwail, D., Househ, M., Hamdi, M., & Shah, Z. (2020). *Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study*. *JMIR Public Health and Surveillance, 6*(2), e19016. Retrieved from <https://pmc.ncbi.nlm.nih.gov/articles/PMC7247466>.

Dar, A. B., Lone, A. H., Zahoor, S., Khan, A. A., & Naaz, R. (2023). *Social Media Sentiment Analysis During the COVID-19 Pandemic: A Systematic Review*. *Frontiers in Public Health, 11*, 10133460. Retrieved from <https://pmc.ncbi.nlm.nih.gov/articles/PMC10133460>.

Sridhar, A., Balasubramanian, P., Khatun, F., & Subramani, R. (2023). *Artificial Intelligence in COVID-19 Twitter Data Analysis: Trends and Challenges*. *BMC Medical Informatics and Decision Making, 23*, 10487588. Retrieved from [https://pmc.ncbi.nlm.nih.gov/articles/PMC10487588](https://pmc.ncbi.nlm.nih.gov/articles/PMC10487588" \t "_new).