NAVIGATING COVLD-19 RESEARCH DATA: AN IN-DEPTH EXPLORATION OF COVID-19

INFOSCI 301 Final Project
Author: Shouzhifan Zhu
Instructor: Prof. Luyao Zhang

INTRODUCTION

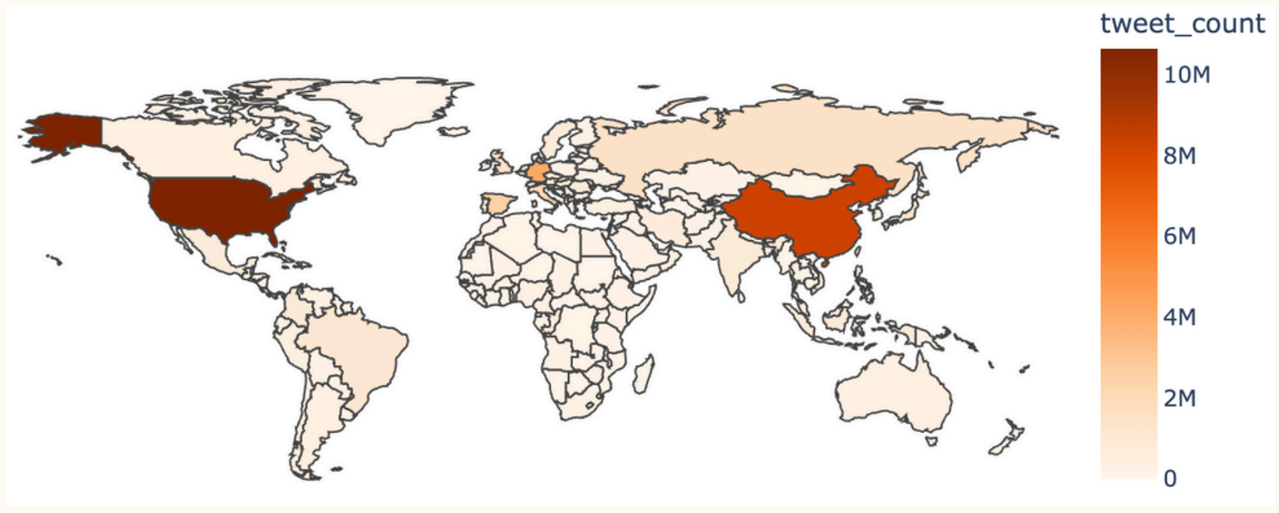
This project explores the distribution of COVID-19-related tweets from February 1 to February 10, 2020. We aim to uncover temporal trends (daily, monthly) and geographical patterns (by country, city), while also assessing the impact of different features on tweet volume predictions using SHAP (Shapley Additive exPlanations).

By highlighting these trends, we hope to offer insights into global engagement and discussion patterns around the COVID-19 pandemic during its early stages.

RESEARCH QUESTION

- Investigate Temporal Trends
  - Examine how tweet frequency varies by day of the week and month.
  - Explore whether there is a noticeable increase on weekends or specific months.
- Analyze tweet volume across different countries and cities.
  - Identify regions with the highest activity and compare trends across locations.
- Assess the Impact of Features (via SHAP)

RESULT & VISUALIZATION



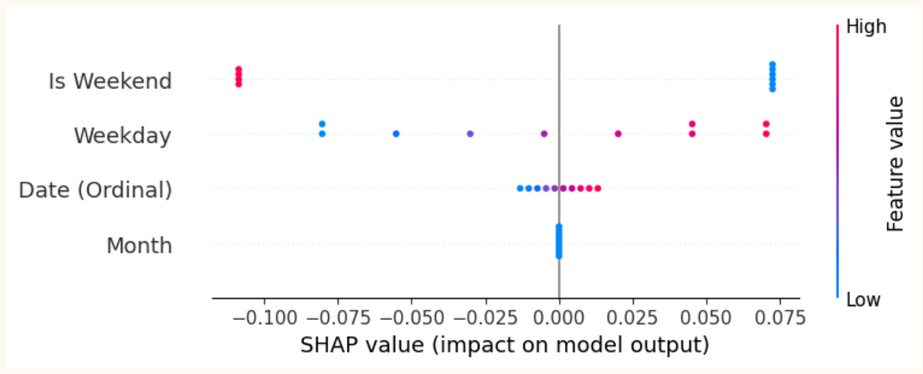
A map left visualizing COVID-19-related tweet volumes. A color gradient represents tweet counts per country:

- Darkest shades: Indicates highest tweet counts (e.g., USA, China).
- Medium shades: Moderately high tweet activity (e.g., India, Brazil, Russia).
- Lightest shades: Very low tweet activity (e.g., certain areas in Africa, and Central Asia).

There is a significant regional variation, influenced by population, internet penetration, and social media usage, reflecting global disparities in pandemic-related discussions.

RESULT & VISUALIZATION (CONTINUED)

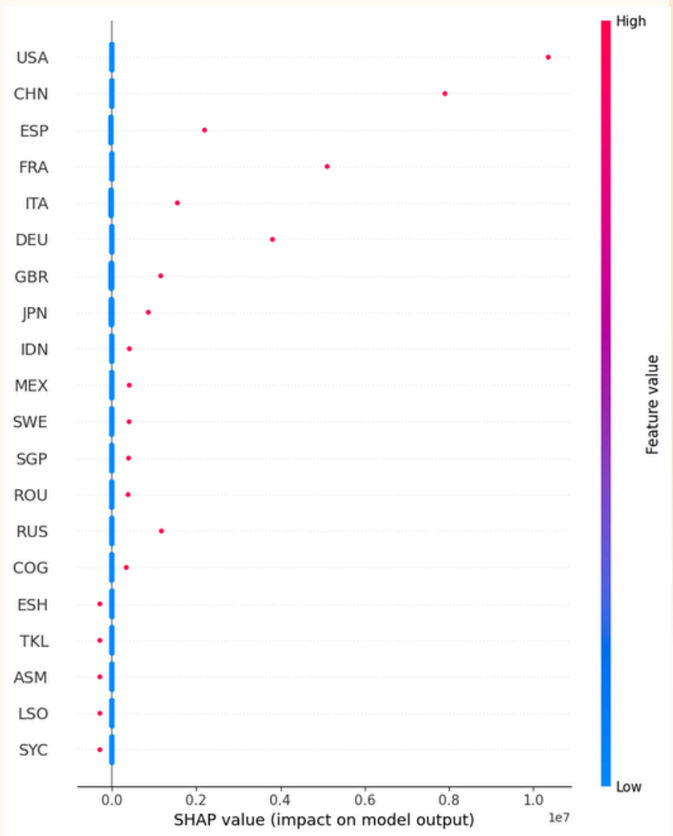
1. **Is Weekend**: Higher positive SHAP values on weekends indicate an increase in tweet volume.
2. **Weekday**: Values hover around zero, implying weekdays do not strongly influence tweet counts.
3. **Date (Ordinal)**: SHAP concentrated near zero, showing no strong day trend in tweet volume.
4. **Month**: Slight variations around zero; some months show mildly positive or negative contributions.



The SHAP plot right reveals how regions contribute to tweet volume:

- **Color Scale**: Red = High tweet volume, Blue = Low tweet volume.
- **USA**: Dominant influence (strongest red), indicating the highest impact on predictions.
- **China** (CHN): Also significant (red), reflecting large tweet counts.
- Spain (ESP), France (FRA), Italy (ITA): Moderate-high influence.
- Less Active Countries (e.g., Seychelles (SYC), Lesotho (LSO)): Minimal effect (blue).

Large, populous, and social-media-active nations (USA, China) drive the overall global tweet volume prediction.



CONCLUSION

- Weekend Effect: Tweet activity spikes on weekends, while weekdays and specific dates show minimal influence.
- Geographical Disparities: USA & China dominate tweet volume, while some regions show low engagement.
- SHAP Analysis Findings: Weekend factors and high-engagement countries are the strongest predictors of tweet volume.
- Global Social Media Trends: Population density, social media usage, and regional policies influence COVID-19 discussions.
- Impact: Results highlight the importance of real-time social media analysis for public health communication & policy strategies.

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.