

Code Evaluation for ESOC COVID-19 Misinformation Dataset Analysis

This document provides a structured self- evaluation of the work completed on the ESOC COVID-19 Misinformation Dataset, aligned with the provided **Project Code Evaluation Criteria**.

1. Use of Python and Analytical Libraries

- **Libraries Used:**
 - **Pandas:** For data manipulation, cleaning, and analysis.
 - **NumPy:** For efficient array operations where applicable.
 - **Matplotlib** and **Seaborn:** For creating visualizations.
 - **FuzzyWuzzy:** Applied for resolving text inconsistencies using fuzzy matching.
- **Implementation:**
 - Efficient data cleaning using Pandas functions like dropna, apply, and groupby.
 - Visualizations to explore and present patterns in the data, such as platform usage, narrative trends, and temporal patterns.

2. Defined Stages

Data Sourcing:

- **Source:** The ESOC COVID-19 Misinformation Dataset, a credible and publicly available dataset from Princeton University's Empirical Studies of Conflict (ESOC).
- **Supplementary Validation:** Contextual information from WHO and trusted health organisations.

Pre-Processing:

- Initial cleaning in **Excel**, followed by CSV conversion for analysis in Colab.
- Key pre-processing tasks:
 - Standardising text fields (e.g., narratives, countries).
 - Handling invalid dates in Publication_Date column.
 - Resolving text inconsistencies using FuzzyWuzzy.
 - Removing duplicates and addressing missing data.

Evaluation:

- Robust exploratory data analysis (EDA) conducted in Colab:
 - Identified trends in platforms, narratives, and motives.
 - Cross-sectional and temporal analysis for regional insights.

Visualization:

- Used **Matplotlib** and **Seaborn** to present findings:

- Bar charts for platform usage.
- Time series plots for misinformation trends.
- Heatmaps for regional differences.

3. Relevance of Analysis Topic

- **Objective:** Understand the spread and evolution of COVID-19 misinformation globally.
- **Core Questions:**
 - How do platforms and regions differ in misinformation dissemination?
 - What motives and narratives dominate, and how do they evolve over time?

Outcome:

- The analysis provided actionable insights into platform-specific trends, regional differences, and motives driving misinformation.

4. Data Sources

- **Primary Dataset:** ESOC COVID-19 Misinformation Dataset.
- **Supplementary Sources:** WHO reports and public health updates for validation.
- **Data Combination:**
 - Cross-referenced key findings with global health data to enhance insights.
 - Leveraged multiple datasets for contextual analysis.

5. Visualization

- **Tools:** Matplotlib, Seaborn.
- **Examples:**
 - Bar charts for platform usage frequency.
 - Treemaps for misinformation narratives by type.
 - Time series plots annotated with pandemic milestones.
- **Outcome:** Clear and concise visualizations effectively communicated findings.

6. Handling Data Anomalies

- **Missing Values:**
 - Imputed where logical (e.g., default values for categorical data).

- Marked others as "unknown" for transparency.
- **Incorrectly Formatted Data:**
 - Resolved invalid dates in the Publication_Date column using logical rules.
 - Applied FuzzyWuzzy to standardise inconsistent text fields.
- **Outliers:**
 - Identified and excluded extreme outliers in numeric data (e.g., dates like "32/13/2020").

7. Credibility of Analysis

- **Evidence-Based Findings:**
 - Insights tied directly to dataset trends and patterns.
 - Findings validated with supplementary health and demographic data.
- **Reproducibility:**
 - Code structured and documented for reproducibility in Colab.
 - Clear methodology enables future replication or extension.

8. Summary of Findings

- **Platforms:** Facebook and Twitter dominated misinformation dissemination globally.
- **Narratives:** False cures and government response narratives were most prevalent.
- **Temporal Trends:** Peaks in misinformation aligned with pandemic milestones (e.g., lockdowns, vaccine rollout).
- **Regional Differences:** US, China, and Turkey had distinct narrative patterns.
- **Motives:** Fear and politics were key drivers behind misinformation.

Next Steps for Improvement / Future Directions

1. **Templates:**
 - Develop reusable templates for similar misinformation analysis projects.
2. **Refinement:**
 - Deepen the analysis on demographic trends if demographic data becomes available.
3. **Collaboration:**
 - Share findings with public health organisations to inform misinformation counter-strategies.