CMPT733 PROJECT MILESTONE
TOPIC : "GLOBAL DATA ANALYSIS FOR SOCIAL UNREST PREDICTION"

TEAM "EXA-MINE":
RAKSHA HARISH, TAVLEEN SAHOTA, HARMAN JIT SINGH

PROBLEM STATEMENT

- How can we predict social unrest events (protests, riots, etc.) across the globe based on factors like historical event patterns, poverty, hunger, crime rate, natural calamities, etc.?
- What insights can be derived from the study of these factors and are there important relationships among these factors?
- What other approaches can be implemented to accurately predict social unrest?

- **<u>Application domain:</u> World governments, policy making, Social science, world hunger/disaster management.**

APPROACHES

Our project will be exploring 3 approaches to predict social unrest:

- **1. Temporal Burst Patterns in Historical Events data**
- **2. Additional Data Sources that capture social, political and economic contexts –** poverty, hunger(food prices), crime rates, etc.
- **3. Predictors from Twitter-feed**

APPROACHES

Our project will be exploring 3 approaches to predict social unrest:

- **1. Temporal Burst Patterns in Historical Events data:**
  Utilizes temporal burst patterns in Event streams to uncover the underlying event development mechanics and predict social unrest.

- **2. Additional Data Sources that capture social, political and economic contexts:**
  Study global factors like poverty, hunger, crime rate, natural calamities and social biases in order to determine the factors that may lead to social unrest in the future.

- **3. Use Predictors from Twitter-feed**
  Twitter data is more vulnerable to propaganda. Since Twitter data is annotated based on Crowd sourcing it isn't sufficiently reliable to predict the sequence of event. Also, using live-twitter streams may be computationally challenging. Because of these and other issues, we concluded to exclude Twitter live-stream for this iteration of the application.

CHALLENGES

1. **Domain Knowledge:** There is limited work done in this area with very little reference material. Therefore, extensive literature review was required.
2. **Handling Data:** Historical datasets are extremely big and handling them requires expertise (familiarity with Google BigQuery etc.).
3. **Availability of Datasets:** Many datasets are available only for a certain time period making it difficult to aggregate datasets without losing key information.

DATASETS

- The GDELT Project (Global Data on Events, Location, and Tone) data – (over 2.5 TB), using Google BigQuery.
- World food program (WFP) data
- International Relations and Human Rights data – Harvard University
- World Poverty data – The World Bank
- Hazards and Disaster Risk (Socioeconomic Data) – NASA, USA
- Crime rates data – UN Office on Drugs and Crime
- Net National Household Income per annum – The World Bank
- ACLED Data – The Armed Conflict Location and Event Data Project

ACTION PLAN

1. EDA: Perform comprehensive analysis including descriptive statistics, temporal analysis, comparative analysis, etc. Use these findings to determine next analysis/course-of-action.
2. Visualize data patterns to observe previous trends that led to social unrest.
3. Integration of a variety of data sources based on relevance (for specific time period).
4. Feature Selection and Engineering from the available datasets and (possibly) real-time twitter feed, to predict social unrest.
5. Analysis of the twitter feed, and finding similarities in the available cleaned datasets.
6. Build competing models to predict how and when these global factors may lead to social unrest and fine-tune the best one.
7. Build an interactive UI allowing users to access real-time social unrest predictions and other insights.

METHODOLOGY

DATA SCIENCE PIPELINE

DEVELOPMENT

- SCRUM: We used an adaptive scrum methodology with pair-programming. We held bi-weekly scrum calls to quickly resolve issues and make decisions. We documented meeting minutes.

TECHNOLOGIES

- **EDA + Visualization:** Python with relevant packages + Jupyter Notebook/JupyterLab (in AWS)
- **Data Storage:** AWS Cloud S3 buckets
- **Interactive Web-app:** Dash
- **SCM:** GitHub

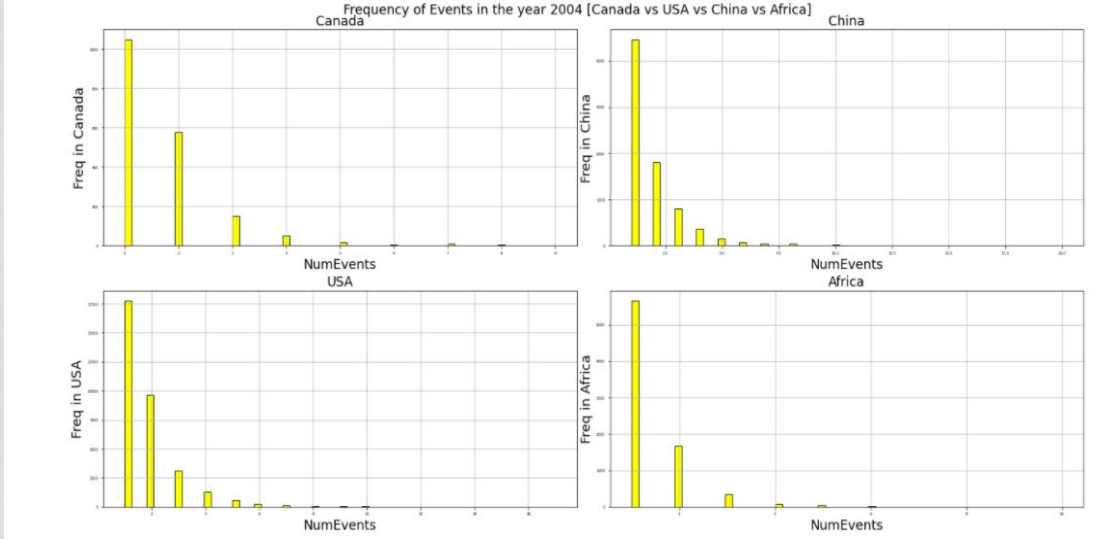
TENTATIVE TIMELINE

Task	Week	Status
Literature Review	Week 6-7	Completed
EDA + Twitter API	Week 7-9	Completed
Modeling	Week 10-11	In-progress
EDA - iteration 2	Week 11-12	In-progress
Improving Model	Week 11-13	
UI	Week 10-12	In-progress
Integration + Debugging	Week 12-13	
Documentation	Week 13	

Note: This is a tentative timeline to track progress, not to be mistaken for Waterfall methodology.

KEY INSIGHTS FROM EDA

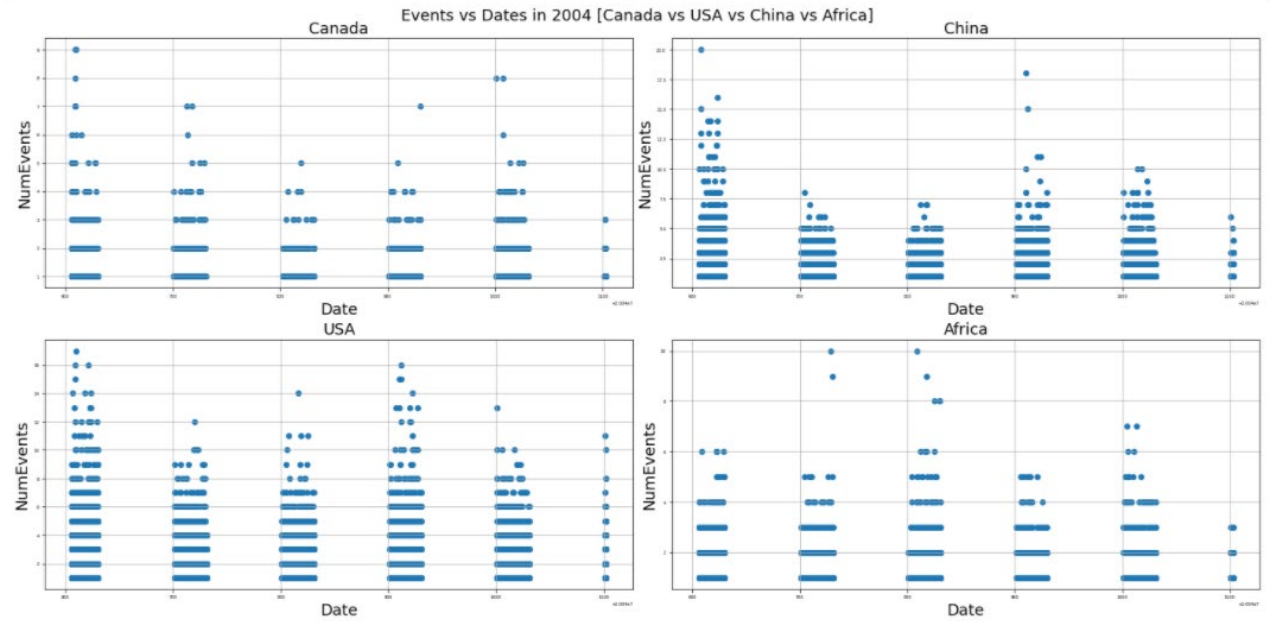
```
axs[1, 1].hist(n4['NumEvents'], bins=60, edgecolor="black", color="yellow")
axs[1,1].set_xlabel('NumEvents', fontsize=18)
axs[1,1].set_ylabel('Freq in Africa',  fontsize=18)
axs[1,1].set_title('Africa',  fontsize=18)
axs[1,1].grid()

fig.tight_layout()
```




```
[129]: fig, axs = plt.subplots(2, 2, figsize=(20,10))
```

Fig1: Frequency of Events Histogram

```
fig.tight_layout()
```




```
fig, axs = plt.subplots(2, 2, figsize=(20,10))
```

Fig2: Distribution of events across the year

KEY INSIGHTS FROM EDA

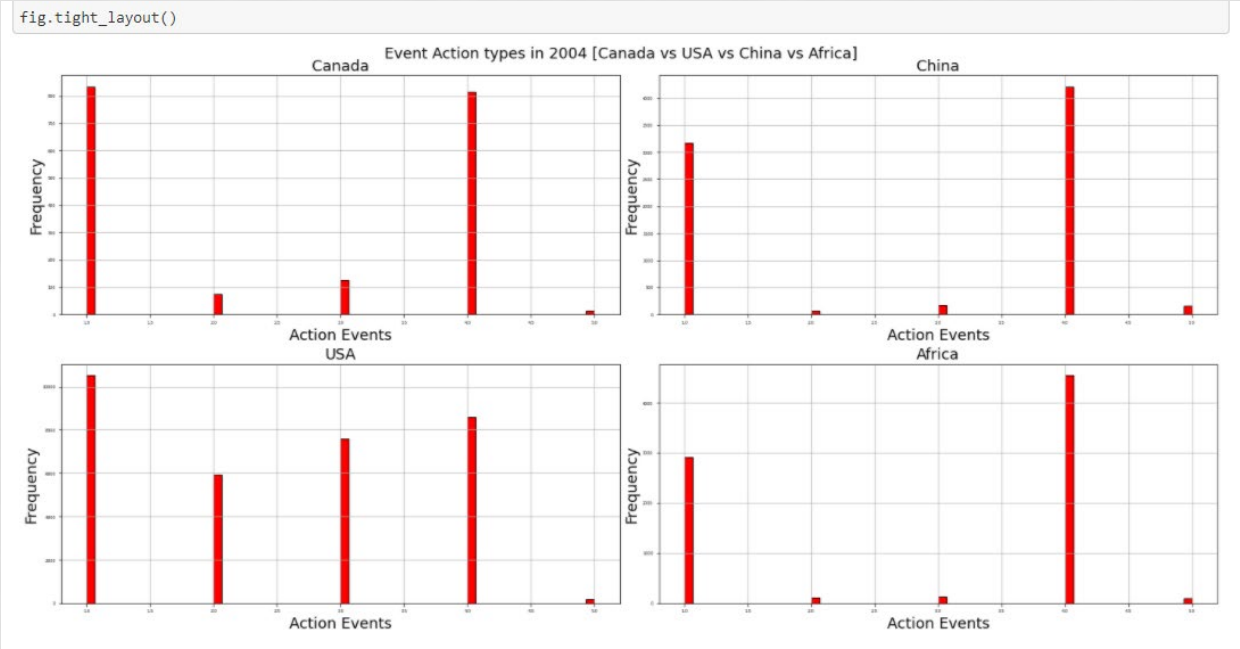


Fig1: Distribution of 5 types of events based on their intensity

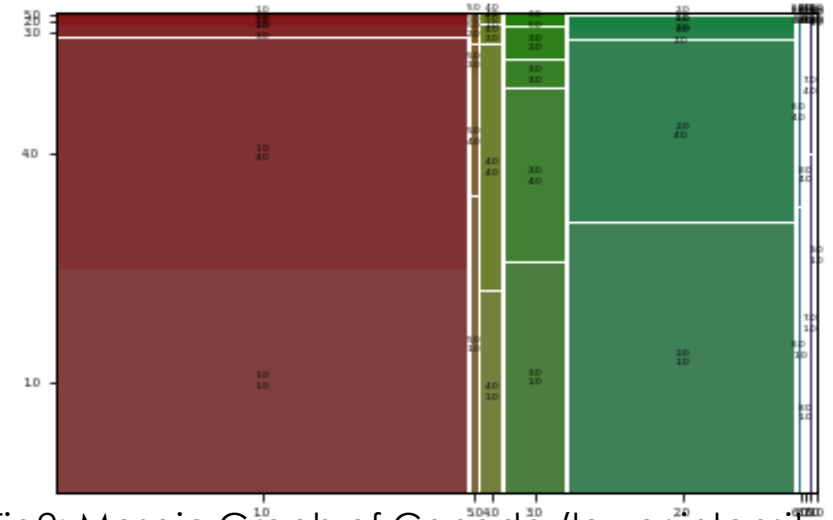


Fig2: Mosaic Graph of Canada (lower intensity events=green)

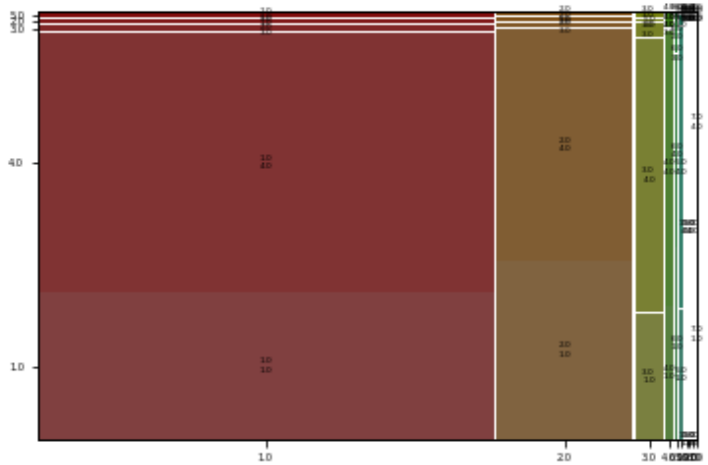


Fig3: Mosaic Graph of USA (high intensity events = red/brown)

THANK YOU!