Global Data Analysis for Social Unrest Prediction

Raksha Harish
rha82@sfu.ca

Tavleen Sahota
tss13@sfu.ca

Harman Jit Singh
hjs10@sfu.ca

Summary

Prediction of social unrest in this project is done in 2 stages. In the first stage, different datasets are integrated for a specific timeline with the main GDELT dataset, and correlation analysis is carried out to select the most relevant features. Once this is done, we consider all these features to predict the Goldstein Scale value, which ranges between -10 and 10. We have implemented and compared 4 models to predict Goldstein values - Linear Regression with Lasso, Random Forest Regressor, H2O ai + Gradient Boosting Machine, and H2O ai + Isolation Forest Model. Higher Goldstein values represent a more intense unrest event.
In the second stage of the project, Markov Chain Simulation is implemented to simulate the sequence of 5 different types of events (i.e, low intensity to violent scale events) for a given country/timeline. Then, Hidden Markov Model is implemented to predict the event sequence transition probability for a random test event sequence. Finally, for a given country predicted to be in the state of unrest (based on Goldstein Scale value), trending social unrest twitter hashtags are analysed to identify the type of unrest event. In the final dashboard, these models are deployed to interact with the user of the application.

1. Motivation and Background

Social unrest can be characterized as dissatisfaction expressed by people about any issue, that might lead to conflict, protest, or even violent acts and civil disobedience. Till now, sociologists and policy makers around the world have been trying to collect country-wise data related to different incidents to study riot patterns, armed conflicts and protests. But there has been no automation or software built so far to predict social conflict before the actual unrest occurs. Prediction of future social unrest is of utmost importance because the

governments across the world can take effective steps and actions to at least control violence, if not prevent it.
Hence, we decided to take on this project to get a data science perspective for global social unrest and build a model that could help to understand different factors that cause unrest, and how these factors can be used to predict future social unrest.

2. Problem Statement

In this project, we integrated different datasets in order to analyse different factors like world food prices, Gross Domestic Product (GDP) rates, crime and poverty rates, armed conflicts and historical event patterns, and dynamic Twitter streams. While doing so, we were able to select relevant features by analysing their relationships. Using temporal burst patterns in historical events data helped in the prediction of impactful social unrest events across the world by capturing the social, political and economic contexts of different countries over different timelines. The main aim of this project is:

- To help world governments and social scientists make better policies by considering all the factors for establishing peaceful governance
- To help in predicting future social unrest, thereby giving enough time for the governments to implement actions to either handle the unrest events or prevent it altogether.

3. Data Science Pipeline

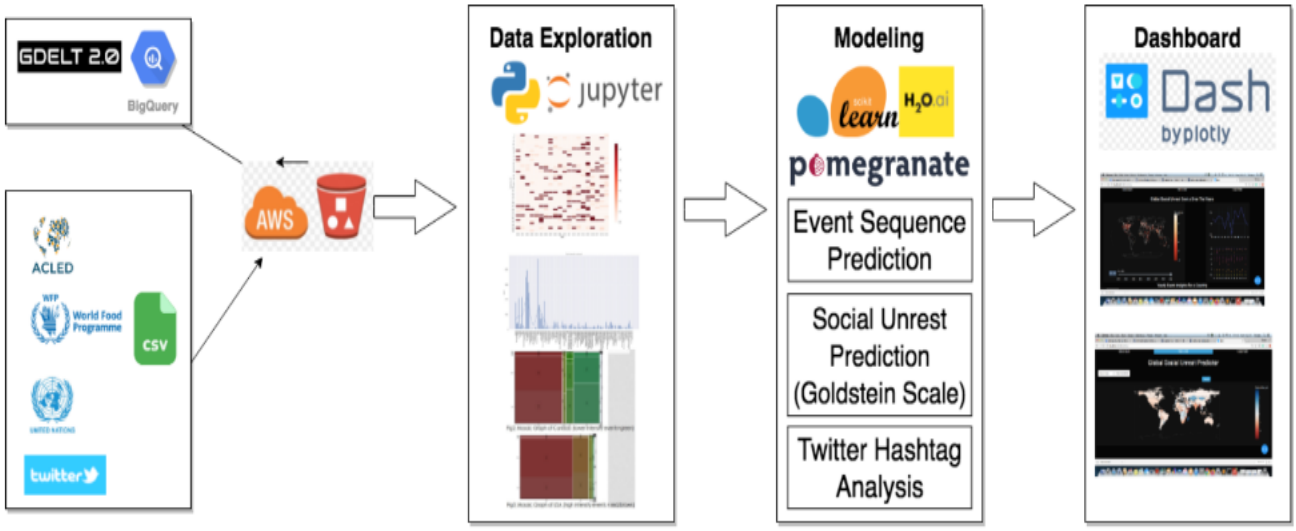

Figure-1 : Project Workflow

3.1. Data collection

The main dataset used for this project was the Global Database for Events, Language, and Tone (GDELT) dataset. This dataset is about 2TB, and is a streaming dataset which is updated by the GDELT project organization every 15 minutes. We integrated this dataset with ACLED (Armed Conflict Location and Event Data) dataset, and the UN Global Crime Rates dataset. We also used the World Food Program (WFP) data and World GDP data to study the changing food prices and GDP over the years. Finally, we used the Twitter API to get the live trending hashtags related to a specific country to analyse the type of event observed in the country.

3.2. Extraction, Transformation, and Loading (ETL)

In the ETL step, the datasets are integrated for a time period of 20 years between the year 2000 and 2020. All the duplicate records and the rows with null values are dropped. The combined data consists of around 70 features, out of which only 18 key features are selected based on a threshold defined for Pearson's correlation for the features. Almost 75% of irrelevant features are dropped. During ETL, we found that the increasing global food prices and the global GDP do not affect social unrest as much as world crime rates and armed conflict. Hence, the features from World Food Program and World GDP data were dropped.

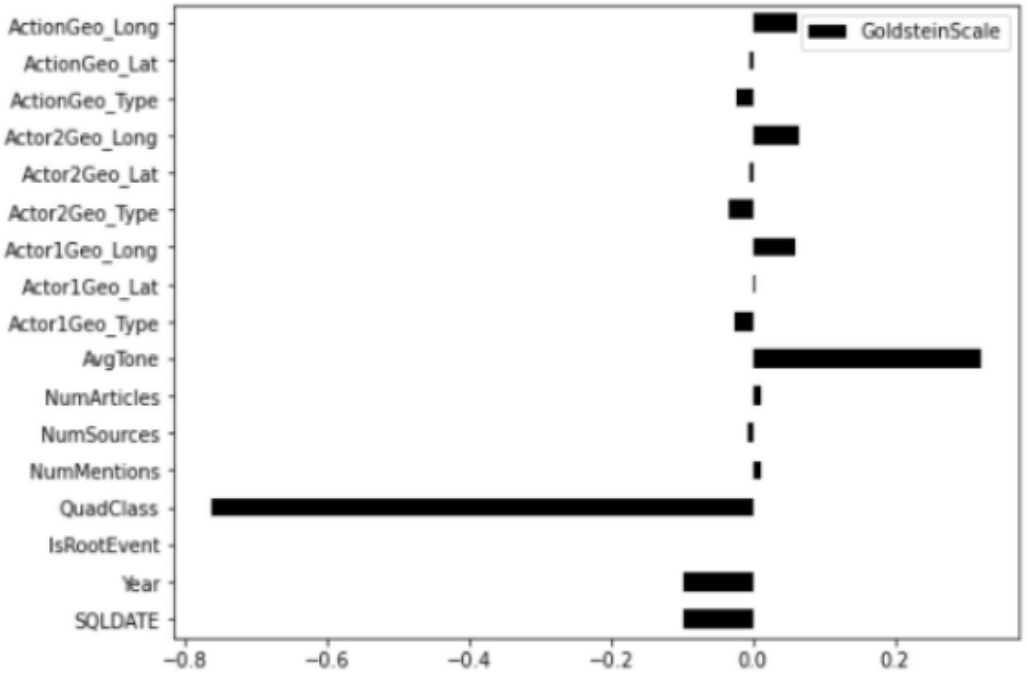

Figure-2 : Pearson's Correlation Plot for the most relevant features to predict Goldstein Scale

We also found that social unrest is more common in certain geolocations (Latitude, Longitude features), and certain Governments (Actor based features). The history of social conflict can be inferred from the number of incidents, the date of occurrence and the number of times the event is mentioned in different platforms (NumSources, NumMentions). The most crucial features are "AvgTone" and the "ActionGeo_Type". There are 5 types of events represented by "ActionGeo_Type" in GDELT, with values 1 to 5, wherein 1 is the least impactful event and 5 is the most violent event. This feature helps in keeping track of the sequence of events observed for a specific country for a given timeline, and what sequence of these events could lead to social unrest. "AvgTome" is the average intensity of events for a given year.

Most of the key features are numerical, with some being categorical. All these features are used to predict the values of the Goldstein Scale, which is the Target feature. According to the GDELT organization, Goldstein Scale has a numeric value between -10 and +10, and is dependent mainly on the "AvgTone" feature. By getting the Goldstein scale prediction, we can understand the intensity of the event represented by "ActionGeo_Type" for different countries. A value of -10 represents low event intensity, for example, a regular event or a public gathering (ActionGeo_Type=1). Higher Goldstein values represent higher intensity of events, like riots or civil disobedience (ActionGeo_Type more than 3).

3.3. Exploratory Data Analysis (EDA)

Due to the huge volume of GDELT dataset (2TB) and the complexity of integration and storage, Google BigQuery free version was used to perform EDA for only 4 specific locations - Canada, USA, China and Africa. The entire EDA was performed using "seaborn" and "matplotlib" packages in Jupyter Notebook. For the final Dashboard, all the countries and their respective features between the year 2000 and 2020 were considered to predict social unrest.

During the EDA, we created a correlation heatmap, and also studied the occurrence of unrest events and their intensity with changing crime rates, GDP, armed conflict and increasing world food prices.

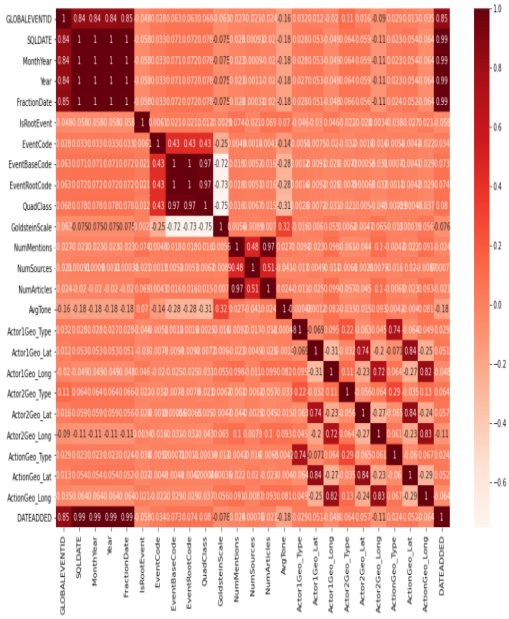

Figure-3 : Feature Correlation heatmap

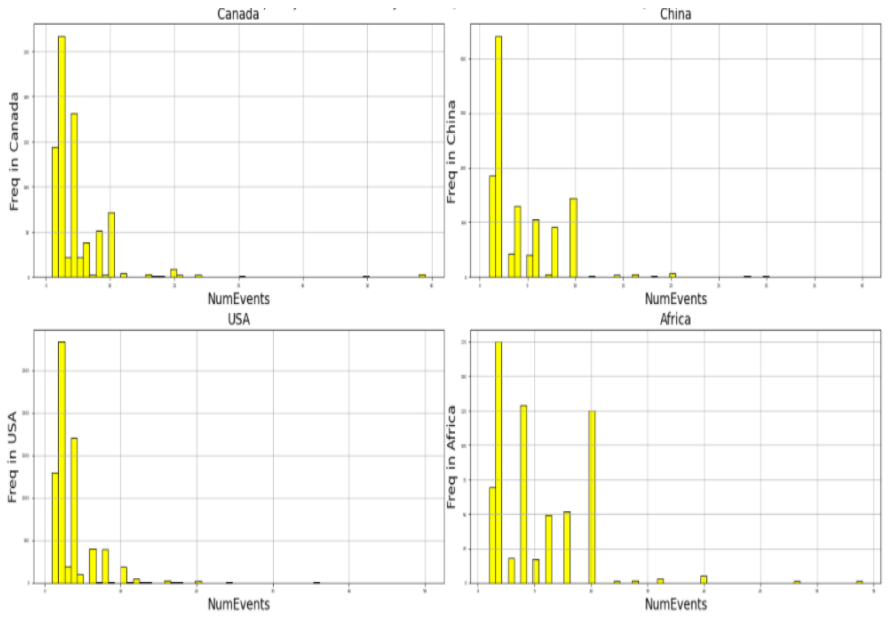

Figure-4 : Frequency of low intensity events in the 4 locations

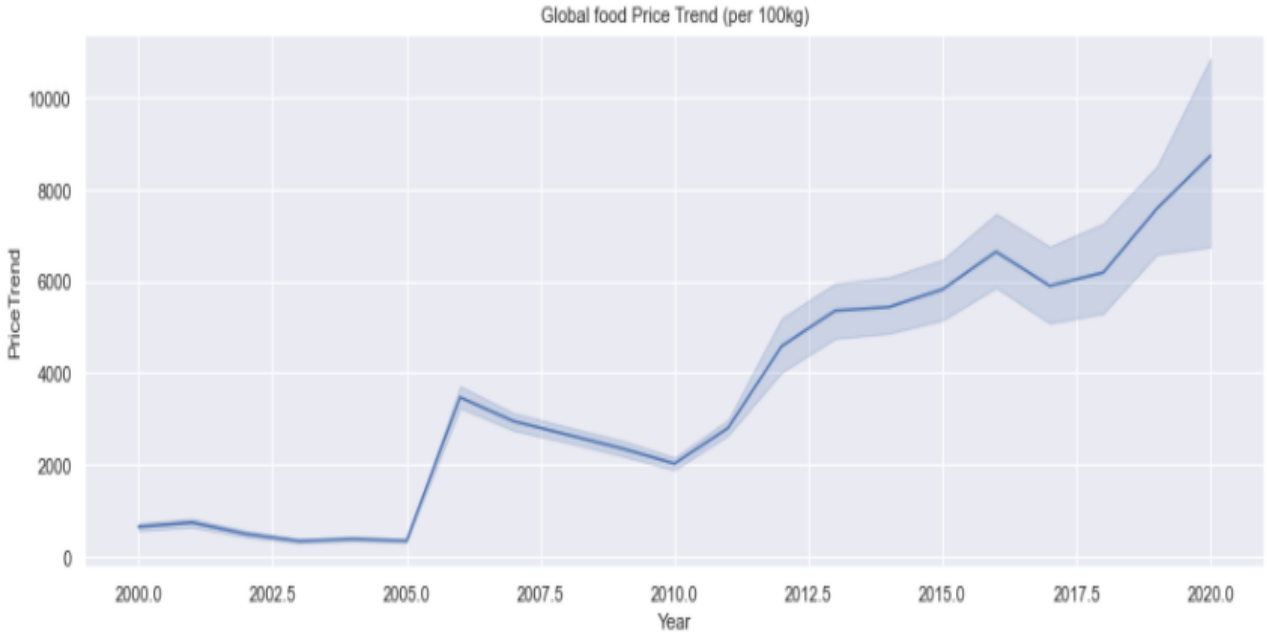

Figure-5 : Average food prices across the world between 2000 and 2020

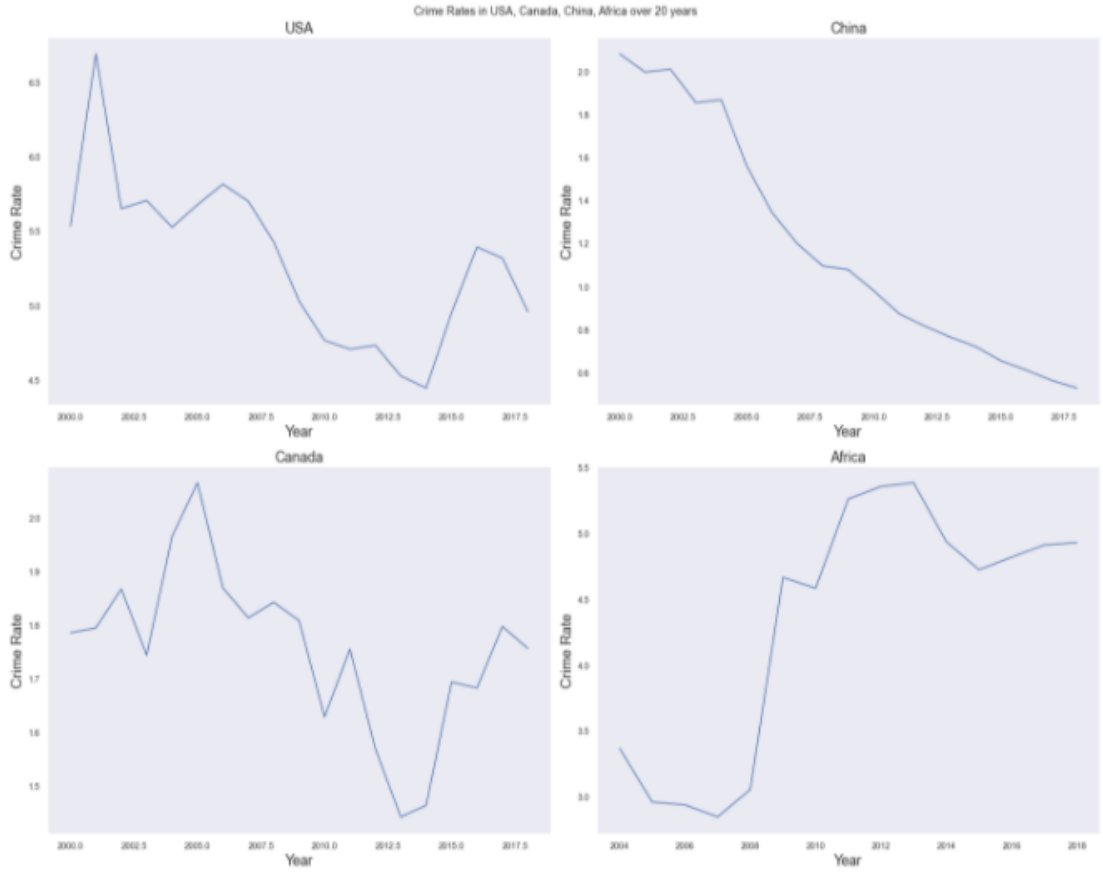

Figure-6 : Yearly crime rates observed in the 4 locations between 2000 and 2020

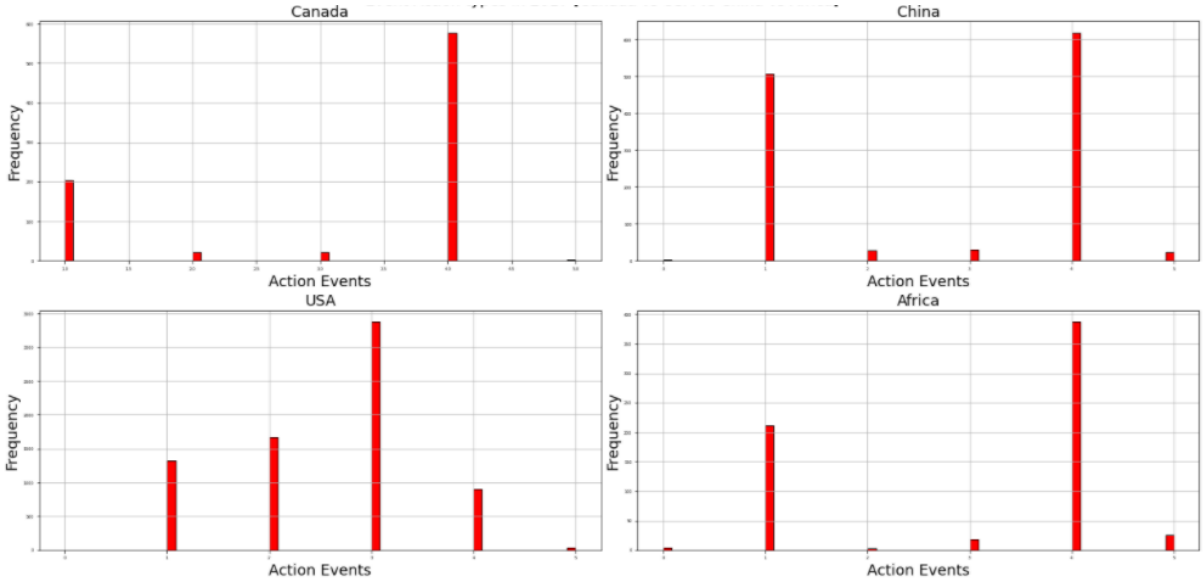

Figure-7 : Frequency of ActionGeo_Type (1 to 5 event types) in the 4 locations over 20 years

After EDA, we found that GDP and world food prices do not affect event patterns or lead to social unrest. Hence, further modeling was carried out on the dataset consisting of GDELT, armed conflict and crime rate features.

3.4. Data Modeling

The data modeling was carried out in 2 stages. In the first stage, 4 models were built and compared to predict the Goldstein Scale values for the dataset of 20 years between 2000 and 2020. In the second stage, Markov chain simulation was used to simulate the sequence of "ActionGeo_Type" events, and Hidden Markov Model (HMM) was built to predict the event sequence transition probability for a random test event sequence. These 2 stages aided in the prediction of social unrest in different countries across the world, as it combined the prediction of intensity of unrest events and also the sequence of events that lead to social conflict.
Different python libraries like "pomegranate", "mchmm", "sklearn", "h20 ai framework" and "graph tool" were used to build and analyse the models for the project.

3.5. Data Product - Dashboard Visualization

The final data product of this project is a front-end Dashboard which is developed using Dash by Plotly. The final dashboard consists of 3 tabs. In the first table, we can get the visualizations of historical data for events observed across the world, and observe their patterns. The second tab of the dashboard is used to predict social unrest by allowing users to choose a specific time period and get the Goldstein Scale values, and also to predict the sequence of unrest events in a graph format. In the third, and the final tab of the dashboard, the users can choose a specific country, predicted to be in the state of unrest and analyse the latest trending Twitter hashtags related to social unrest for the selected country.

4. Methodology

Our aim was to model the historical event patterns observed across the globe to predict the Goldstein Scale and also the probability of occurrence of a simulated sequence of events, thereby predicting social unrest based on the key features of the dataset. We have implemented regression based models and statistical Markov models to predict social unrest.

We found that classical regression based algorithms work better for this data when boosted with the "h2o" python package, which is basically the AutoML H2O ai framework. H20 Driverless AI, a product by H2O.ai, is capable of performing feature engineering as well as

finding ML models best fit for your data and are capable of deployment. This was used to compare different model performances, and to have an idea about the impact of different input features on the model performance.

Since unrest in society is volatile in real time, we have assumed that the predictions are fairly accurate for the chosen timeline, and the integrated dataset. Real time prediction can be performed on getting access to live global data, but this was out of the scope of our project because the entire project was built using the freely available softwares/data.

For our bespoke dataset between the years 2000 and 2020, we have used the random Train-Test split for modeling and predicting the Goldstein Scale values (-10 to +10 : no impact to violent event scale, by specifying 70% of the data as the train set, and the remaining 30% of the original dataset as the test set.

For Markov modeling, the inbuilt "mchmm" (markov chain hidden markov model) and "pomegranate" python packages were used, wherein the entire dataset was taken as input for the HMM to predict the transition probabilities for the simulated event sequence.

4.1. Data Modeling

4.1.1 Goldstein Scale Prediction

The "Goldstein Scale" column is taken as the target feature, which is predicted using the most relevant (i.e., correlate) input features. A prediction value between -10 and 0 indicates a low intensity social event, whereas a prediction between 5 and +10 indicates a violent impact unrest event. The following 4 models were implemented for the prediction:

4.1.1.1. Model-1: Linear Regression with Lasso Regularization -

Linear regression is the most common approach of regression analysis. In this model, the relationships are modeled using linear predictor functions whose unknown model parameters or coefficients are estimated from data. It has considerably lower time complexity as compared to other machine learning models due to its simple nature but suffers from problems like overfitting and underfitting.
To reduce overfitting in our models we used lasso (least absolute shrinkage and selection operator) regularization in which cost function is altered by adding a penalty equal to magnitudes of coefficients.

4.1.1.2. Model-2: Random Forest Regressor -

Random Forest is a popular machine learning model which uses ensemble learning methods for classification or regression. A random forest is a meta-estimator (i.e it

combines the results of multiple predictions) aggregates many decision trees. It is able to handle complex data but requires high computation and training times. In our experiments, it outperformed the linear regression model by a good margin. This performed better than Linear regression when we increased the number of estimators for the model.

4.1.1.3. Model-3: Gradient Boosting Machine (GBM) with H2O ai framework -

Gradient Boosting Machine is a forward learning ensemble method. The guiding heuristic is that good predictive results can be obtained through increasing refined approximations. H2O's GBM sequentially builds regression trees on all features of the dataset in a fully distributed way to reduce time complexity. While random forests build an ensemble of deep independent trees, GBMs build an ensemble of shallow and weak successive trees with each tree learning and improving on the previous. When combined these weak trees make a high accuracy prediction. This model gave us the best result for Goldstein Scale prediction.

4.1.1.4. Model-4: Isolation Forest Model (IFM) with H2O ai framework -

Isolation forest is similar in principle to Random Forest and is built on the basis of decision trees. Isolation Forest, however, identifies anomalies or outliers rather than profiling normal data points. Isolation Forest isolates observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of that selected feature. This random partitioning produces noticeably shorter paths for anomalies. Due to the huge variations in our dataset, outlier detection was slightly difficult. Hence, this model got a lower accuracy of prediction.

4.1.2 Event Sequence Prediction

For this prediction, we have considered the "ActionGeo_Type " column values over the 20 years time period. This column consists of categorical values '1' to '5', wherein '1' represents 'no impact event' and '5' represents a violent/unrest event. For the sake of encoding these categorical values for Markov chain simulations, we have replaced "1,2,3,4,5" by "A,B,C,D,E" for the "ActioGeo_Type" feature column.

In the first step, we use the inbuilt python “mchmm” package to simulate a sequence for the existing values for “ActionGeo_Type”. This is the initial Markov chain simulation with the pre-defined event state transition probabilities. For this simulation, we build a Hidden Markov Model (HMM) from scratch and then train the model based on the initial simulation to get the likelihood of occurrence of the simulated Markov Chain. In order to test this HMM, we input a random sequence of events A-E (usually with min length=3, max length=10). For this input test sequence, we use the pre-trained HMM to predict the probability of occurrence of the test sequence in the future based on the prior event sequence transition probabilities derived from the initial Markov Chain simulation for the original dataset.

Finally, we use the “graph tool” package with “pomegranate” package to draw the graph for the Markov Chain simulations (both original and test simulations), wherein, the nodes are represented by A-E (i.e., the type of action events) and the edges of the graph represent the transition probabilities from one event state to another. This HMM prediction for the simulation is evaluated in our project by Viterbi algorithm and the Forward-Backward algorithm.

4.2. Twitter Hashtag Analysis

Twitter, a prominent microblogging and social networking service, is one of the best indicators of trending topics in a country. It is used by Governments to unveil new policies and other updates, protesters use it to plan, organize and announce protests or demonstrations. Each news or update is often accompanied by a hashtag which users can use to check other similar posts. In order to gauge the most recent and current events in a country we can use the “top trending hashtags” feature in Twitter. As these top trending hashtags are often indicative of major happenings in a country, we performed analysis on these to further our analysis of the potential reasons for unrest in a country.

We used Twitter API to get top trending hashtags for a country. For each hashtag, tweets are downloaded for analysis (we downloaded only 50 tweets per hashtag due to quota constraints). Firstly, we cleaned the downloaded tweets by removing their respective URLs. Next we removed all stop words that are not of much relevance, by using the Natural Language Toolkit package “nltk” in python.

Often tweets containing multiple hashtags are those that are linked to one another. Hence we saved the list and counts of all identified related hashtags, and defined a list of unrest related keywords containing words like protest, war, uprising, revolt, insurgence etc. We further analysed the tweets by counting the number of unrest

related words found in the tweets and then saved their counts. Similar analysis was done for all the trending hashtags.

At the end of the analysis, we had a list of top trending hashtags, counts of unrest related keywords (by analysing the tweets for that hashtag) and a list of related hashtags. A network graph with the country in the middle and trending hashtags in a circular layout was created for top trending hashtags in a country, with node and edge sizes proportional to the tweet volume associated with that country.

Multiple other circular network graphs were drawn with hashtags (having high count of unrest related keywords) at the centre and related hashtags nodes on the circumference of the circle. Node and edge size was set according to the number times the hashtags appeared together in tweets.

5. Evaluation/Results

Part-1 : Results for Goldstein Scale Prediction

We experimented with multiple models for prediction. We found out that GBM along with H2O framework performed best with accuracy of 86.31%.

Models	RMSE	MAE	Accuracy
Linear Regression + Lasso	0.5873	0.2651	59.01%
Random Forest Regressor	0.0145	0.2392	78.17%
GBM + H2O framework	0.4567	0.7864	**86.31%**
IFM + H2O framework	1.0043	1.7295	64.03%

Part-2 : Results for Event Sequence Prediction

- Markov chain simulation for existing event sequence
  Initial Simulation sequence : ACBDDAAABB
  HMM prediction for the initial simulated sequence : 1111111111

- Markov chain simulation for random input test sequence
  Test Sequence Input : AEEAEEAEAC
  HMM prediction for the random input sequence : 1001001110

- Validation using Forward-Backward Algorithm for initial Sequence "ACBDDAAABB"
  Sequence: 'ACBDDAAABB' -- Log Probability: -17.735770945821383
  Sequence: 'CCE' -- Log Probability: -5.33298645408333

Sequence: 'EEAC' -- Log Probability: -7.394981082159444
Sequence: 'AAACBB' -- Log Probability: -10.876577233453677

- Validation using Viterbi Algorithm for new test Sequence “ACDEEC”

Sequence: A, Log Probability: -2.3025850929940455
ACDEEC
A
Sequence: DA, Log Probability: -4.017383521085972
-ACDEEC
DA
Sequence: AC, Log Probability: -4.017383521085972
-ACDEEC
AC
Sequence: AE, Log Probability: -4.017383521085972
-ACDEEC
AE
Sequence: AECC, Log Probability: -7.446980377269825
---ACDEEC
AECC
Sequence: ACEDB, Log Probability: -9.16177880536175
----ACDEEC
ACEDB
Sequence: ABBB, Log Probability: -7.446980377269825
---ACDEEC
ABBB

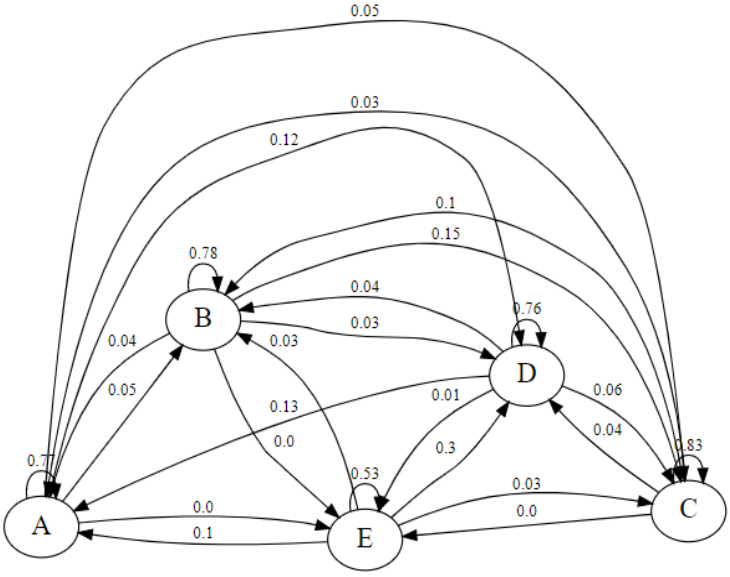

Figure-8 : Initial simulation graph for “ACBDDAAABB”

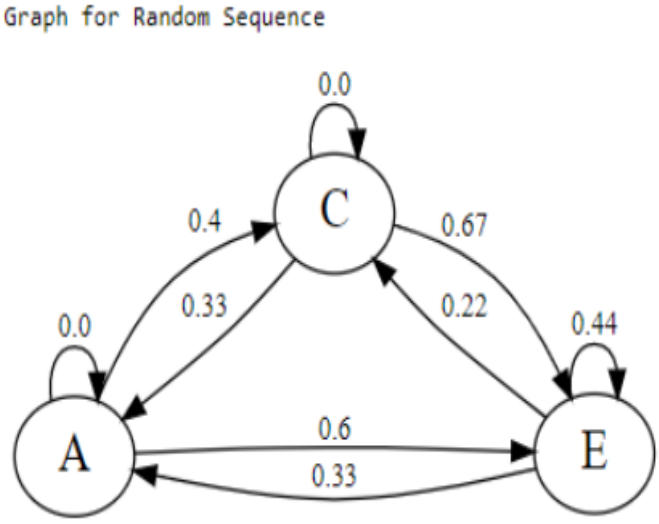

Figure-9: Graph for test sequence “AEEAEEAEAC”

6. Data Product

Our final product is an interactive dashboard built using Dash by plotly. It has 3 major functionalities:

6.1. **Explore Insights:** This functionality allows the end-user to explore historical data and understand patterns associated with social unrest. It enables temporal analysis such as visualization of global social unrest events (Goldstein Scale >= 5) over the years, visualization of global food, crime and economic trends and frequency of various types of EventRootTypes each month to detect possible seasonal trends.

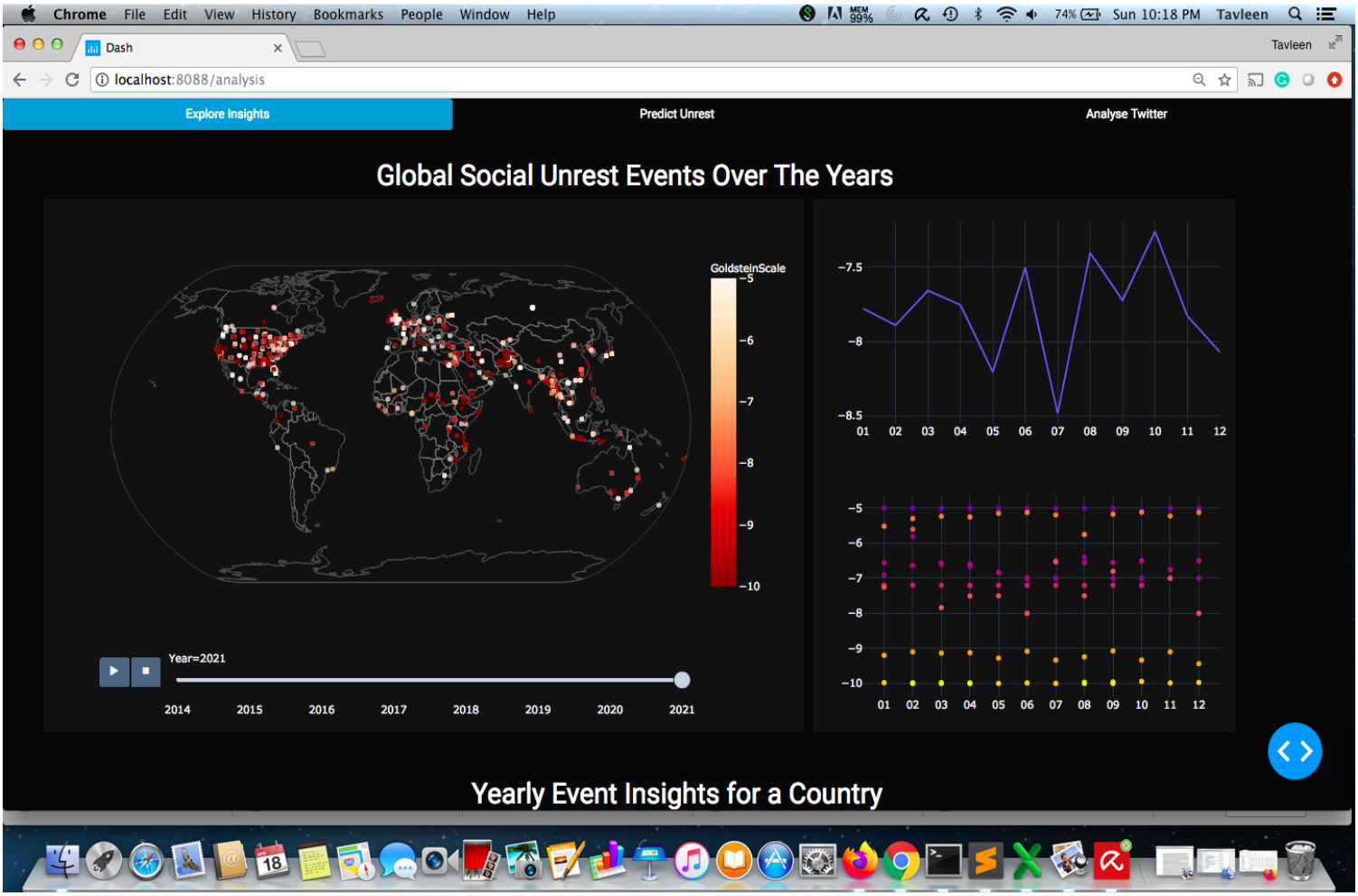

Figure-10: Dashboard Tab1 “Explore Insights” : Part-A

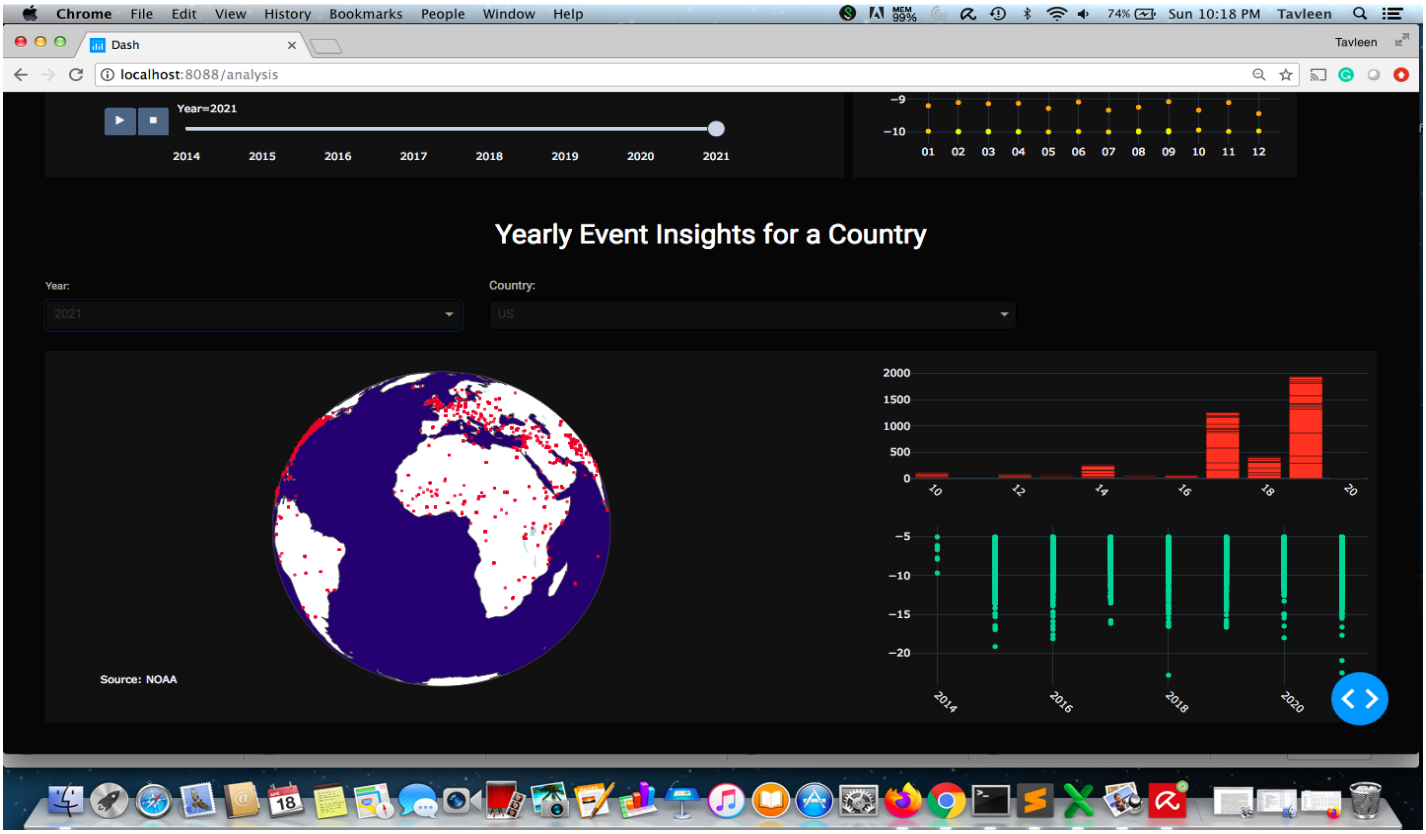

Figure-11: Dashboard Tab1 “Explore Insights” : Part-B

6.2. **Predict Unrest:** This functionality allows the end-users to make 2 kinds of predictions: ‘Global Social Unrest Prediction’ and ‘Event Sequence Prediction’. First, the user can visualize the predicted global social unrest across the specified time-period and predict the probability of occurrence of a sequence of events. Hovering over the countries shows the predicted Goldstein Scale for that country. We use the Linear Regression with Lasso Regularization model for prediction.

Then, the user can enter an event sequence (ActionGeo_Type feature) and predict it’s likelihood. A sequence consisting of letters A-E of length 3 to 7 must be inputed. The letters represent the intensity of an event with A being the least violent and E the most. A sequence of 0s and 1s is displayed as prediction.1 signifies that the event is likely to occur after the preceding event and 0 signifies the opposite. A graph is also displayed that visualizes the transition probabilities between these events using Markov Chain.

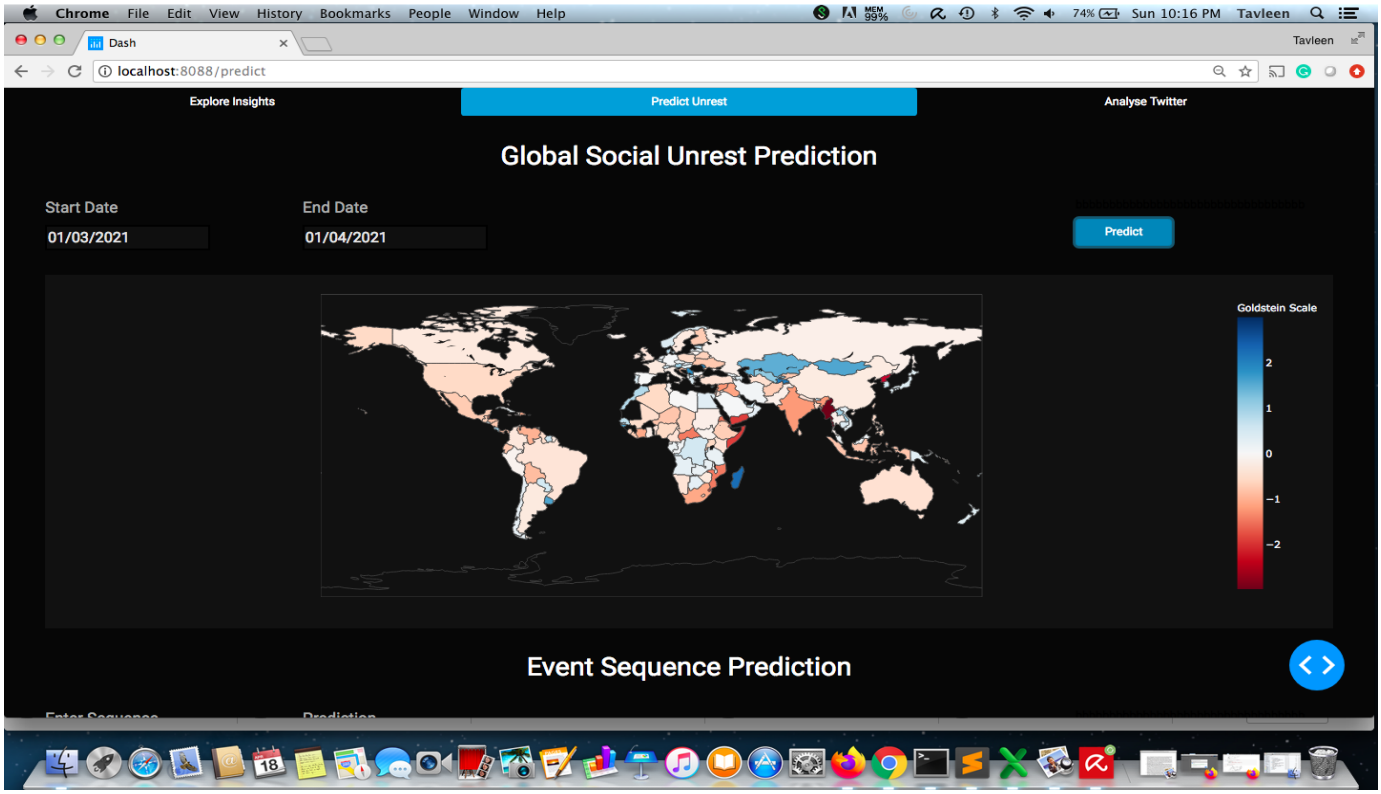

Figure-12: Dashboard Tab2 “Global Social Unrest Prediction” : Goldstein Scale Prediction

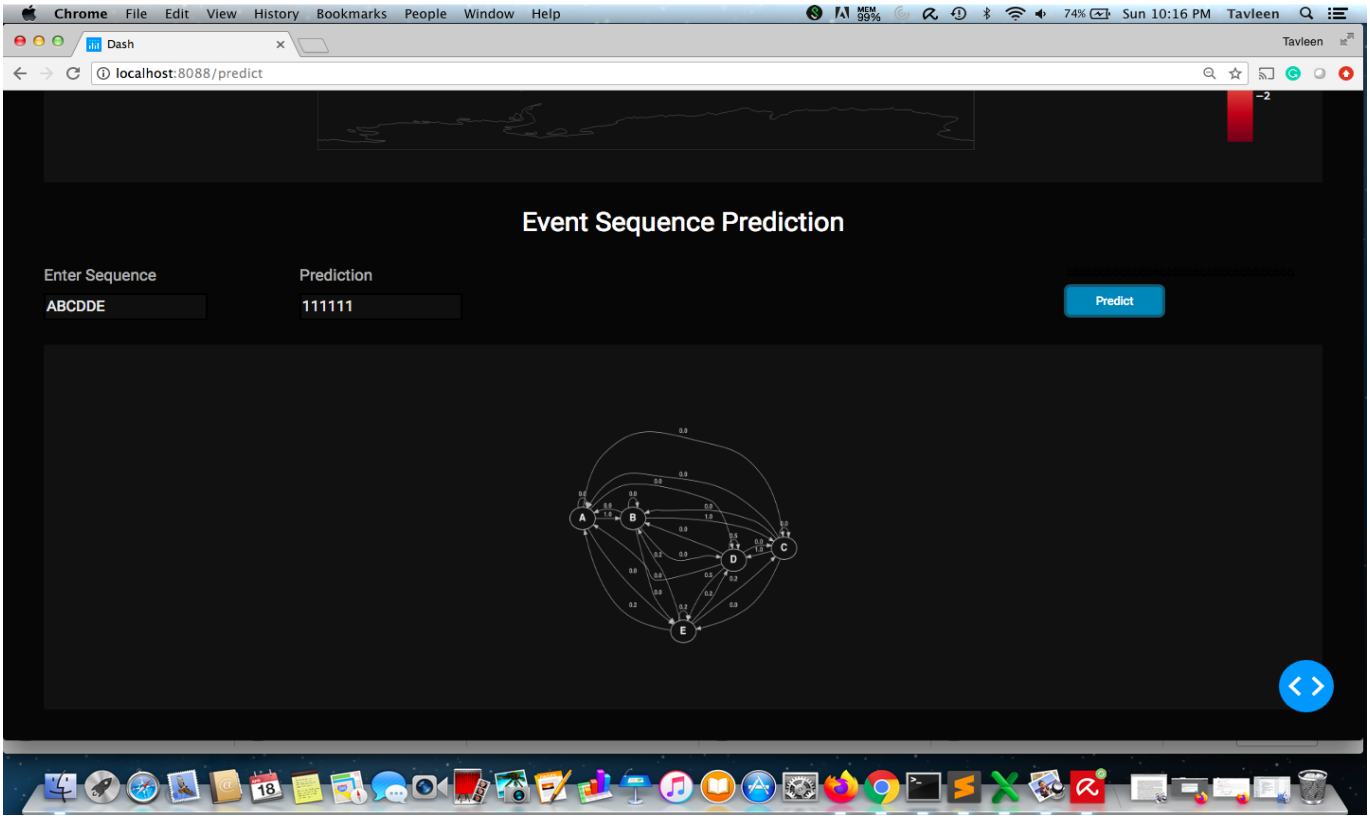

Figure-13: Dashboard Tab2 “Global Social Unrest Prediction” : Event Sequence Prediction

6.3. **Analyze Twitter**: End users can view top trending twitter hashtags countrywise on this page. Users can select a country from the drop down and the network graph of top trending hashtags is displayed in a circular layout.

The user can scroll down and select a hashtag from the drop down menu to see all related hashtags to that hashtag. Note that this drop down menu contains only identified unrest related hashtags. The user can select a country and then select unrest related hashtags for analysis.

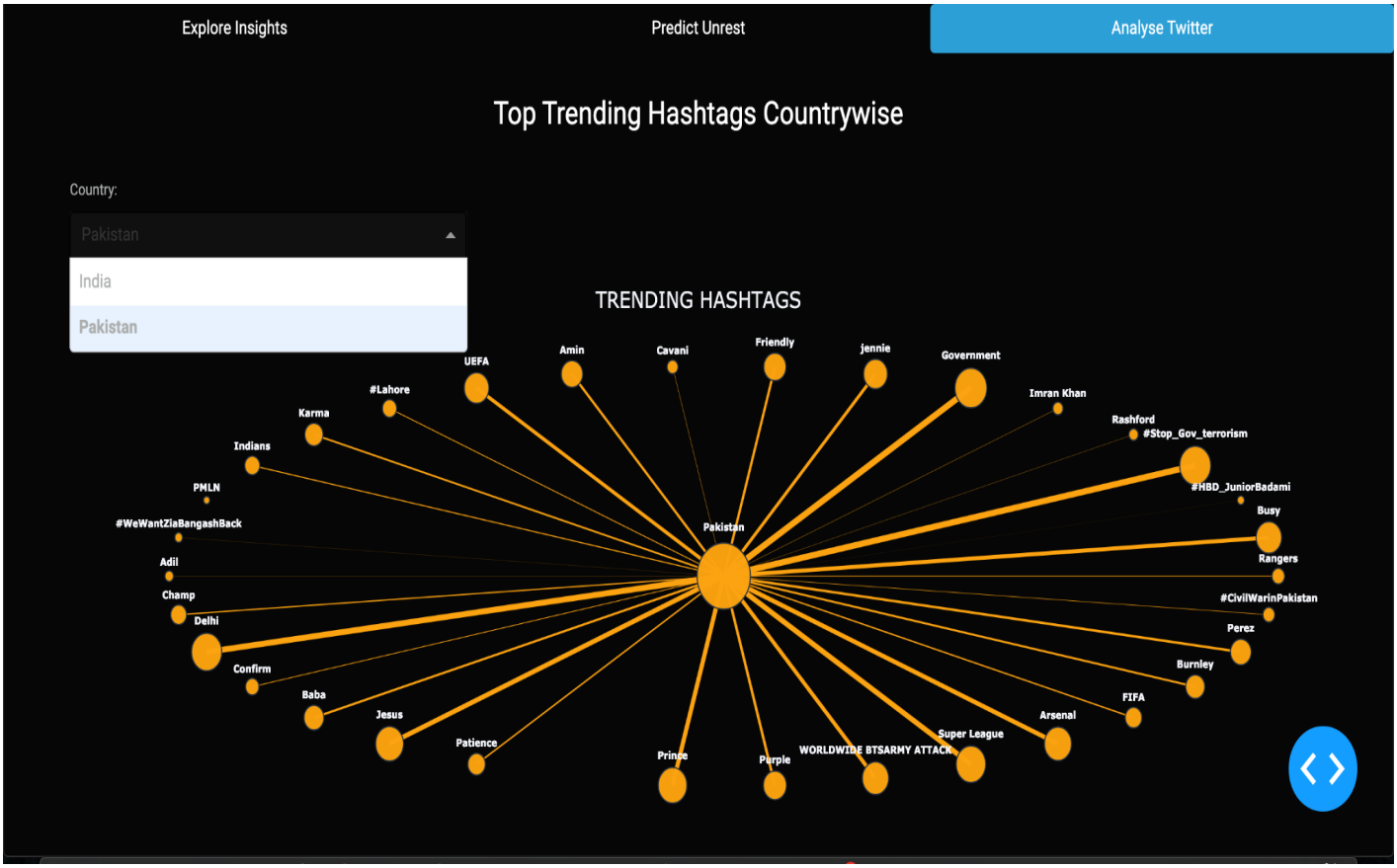

Figure-14 : Dashboard Tab3 “Twitter Analysis for Trending Hashtags in Unrest Country”

Figure-15 : Dashboard Tab3 Twitter Analysis for Hashtags related to the Trending Hashtags for a Country under Social Unrest

Lessons Learnt

- During the initial project development, we found that there was very limited work done in this field. Hence, we had to refer to different sociology and political articles to understand what factors lead to social unrest, and how we can use concepts of data science to understand their relationships.
- Huge volumes of unstructured data was available online, but access was limited due to data quota constraints, and additional costs for querying and storage. Hence, we were able to use only the free tier versions of Google BigQuery and Twitter API to build the project.
- EDA and integration of different datasets took more time because of inconsistent timelines of different events around the globe, and also due to missing data for certain time periods.
- Configuring Auto-ML tools such as H20.ai was beneficial for understanding the underlying assumptions in data during the analysis phase.

- Due to variations in the final cleaned data, we found it slightly difficult to generalize a model. Hence, we built four models and studied their performance. A more sophisticated model with better performance can be built on getting access to the data/API which isn't included in the free versions.
- Twitter API free tier can only be used to download past seven days of twitter data with limited quota. Initially we wanted to use twitter data for adding features such as daily tweet count, number of tweets with unrest related keywords, total tweet count with negative sentiment etc to complement GDELT data but due to these restrictions we changed the scope of analysis to show the possible reasons for unrest in a country.

Conclusion

In this project, we found that crime rates, armed conflict and other factors like history of events, and geolocation determine how and when social unrest events can be observed in a country. For the country which is predicted to be in the state of social or political conflict, trending Twitter hashtags can be narrowed down and analyzed to get an idea of the type of unrest event, i.e., if it is a peaceful protest or a violent event. A few projects have made use of the Golden Standard Report (GSR), a report published by scholars in the sociology field to validate certain findings. But this report uses different standards for different countries, and hence cannot be used as ground truth to evaluate a common model that can be applicable to all the countries in the world. Hence, we ended up using the Goldstein Scale of GDELT data as a common feature to predict the intensity of social unrest at any given place in the world and validate our prediction based on historic data.

Future Work

This project showed how likely it is to observe social unrest in different countries for a given time period, and also the probability of occurrence of a given sequence of event types. In the future work, we intend to predict the future sequence events by considering a dataset with a bigger timeline and more features. This would help in improving the performance of the model, wherein, we could get a more accurate forecast of events and this information can be used beforehand to prevent future social unrest. A statistical approach can be adapted to achieve better integration of different sources of data, and the scale of social unrest analysis can be extended to get better predictions in future studies.

References

1. Parang Saraf and Naren Ramakrishnan, “EMBERS AutoGSR: Automated Coding of Civil Unrest Events”, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* Pages 599-608, August 2016.
2. Pei Li, Xin Zhang, Zhaoyun Ding, Jiajun Cheng, and Hui Wang, “Predicting Social Unrest Events with Hidden Markov Models Using GDELT”, *Journal of Discrete Dynamics in Nature and Society,* Article ID 8180272, May 2017
3. Gizem Korkmaz, Jose Cadena, Chris J. Kuhlman, Achla Marathe, Anil Vullikanti, and Naren Ramakrishnan, “Multi-Source Models for Civil Unrest Forecasting”, *Article-6 Springer Journal*, July 2016
4. https://blog.gdeltproject.org/
5. https://acleddata.com/#/dashboard