Socio Political Analyses in Regions Of World
Anindita Saha, Arul Bharathi, Namita Shah

MOTIVATION
- Increased number of conflicts, fatalities, mass migration of people due to socio-political events in many countries
- Loss of Educational and Medical Coverage to major population in some countries.

OBJECTIVES
1. How does socio-political conditions of a country impact its human development(health +education + livelihood) ?
2. How can NGOs and resource planners predict the Human Development Index of a country, given its socio-political features?
3. How does the news media cover and impact prominent socio-political events in a country?

PIPELINE

DATA COLLECTION & INTEGRATION	DATA EXPLORATION & INSIGHT ANALYSIS	PREDICTIVE MODELLING	IMPACT ANALYSIS & ANOMALY DETECTION	EVALUATION & VISUALIZATIONS

DATA COLLECTION & INTEGRATION: Google BigQuery, V-Dem INSTITUTE, ACLED Bringing clarity to crisis, The New York Times

DATA EXPLORATION & INSIGHT ANALYSIS: tableau, Pandas

PREDICTIVE MODELLING: jupyter, Natural Language Analysis with Python NLTK, learn

IMPACT ANALYSIS & ANOMALY DETECTION: Spark MLlib, Natural Language Analysis with Python NLTK, K + TensorFlow

EVALUATION & VISUALIZATIONS: D3, Google, Flask web development, one drop at a time

Storage: Google Big Query
Data Sources: V-Dem Institute, ACLED, Uppsala Conflict Data Program
APIs: New York Times, UCDP
Time Frame: 3 Weeks
Integration: 4+ sources
Preprocessing: Outlier & Missing Value Treatment

Exploratory Data Analysis: Leveraged the Data Visualization Tool - **Tableau** as a medium to explore the variables and interactions between them
Insights: Discovered several findings and formulated Insights using socio-political data metrics.

Regression Models: Performed Regression to predict the Human Development Index of a country given it's socio-political status
Classification Models: Performed Supervised multi-class Classification to identify the health access in a country, given its socio-political atmosphere

Misconception Analysis: Trained a Deep Recurrent Neural Network (LSTM) to classify 'Fake' News articles.
Topic Modeling: Latent Dirichlet Allocation(LDA) is used to find topics which were published in news articles and correlate the impact with current affairs.

Evaluation: The above built models are tested with standard evaluation metrics that could give more efficacy
Visualization and Web App: An integrated dashboard that consists of actionable findings and insights is built using Flask Web Application,D3 and Google Charts.

EDA & INSIGHTS

- The average HDI value for countries with equal access to basic education is 41% high
- Nearly 90% of Liberal Democratic countries have high HDI and Civil Society Index
- Countries that respect personal autonomy and rights seems to have 47% higher HDI
- Most of conflicted countries seems to have more international influence in their policies
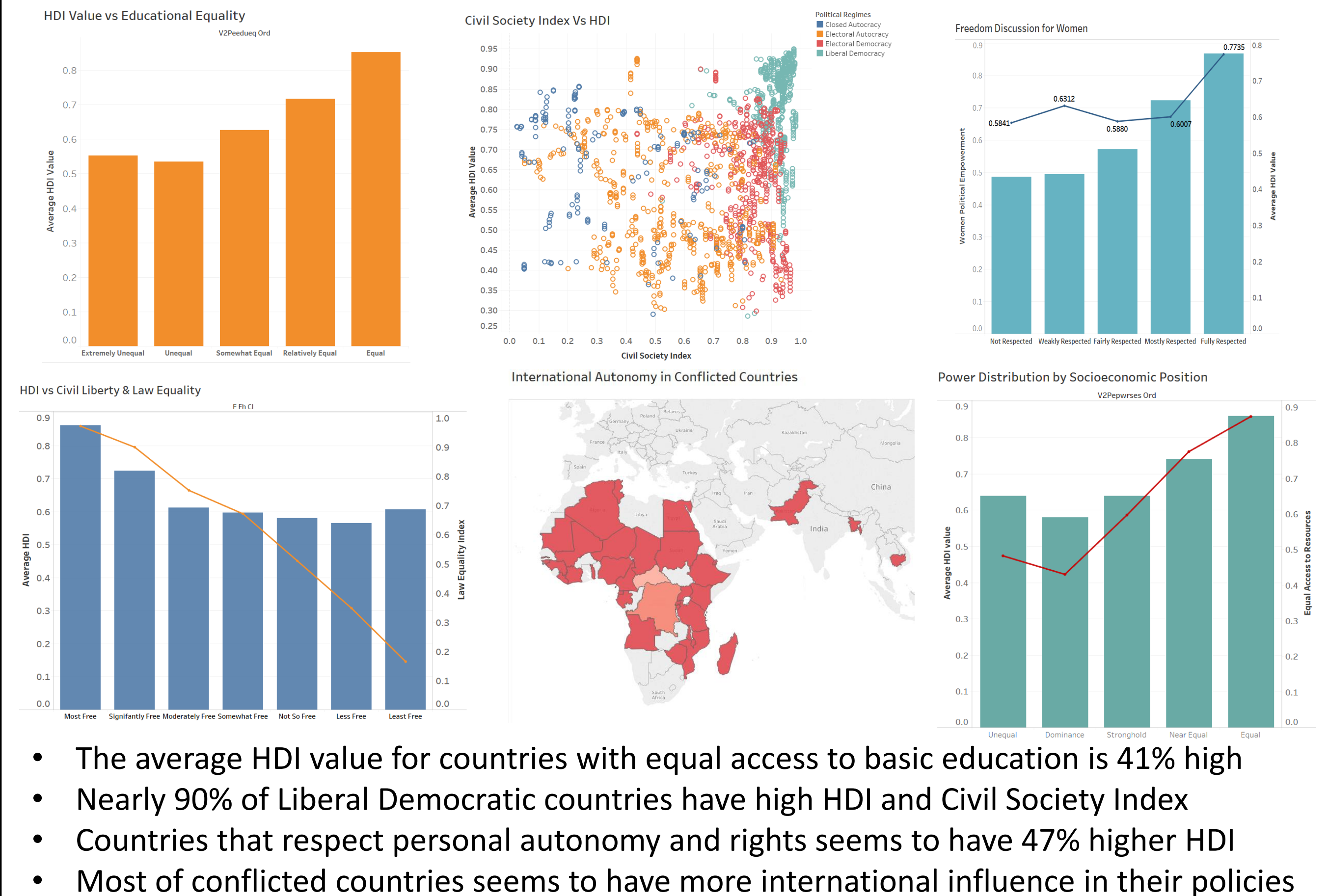
PROBLEM SOLVING
Problem Time Frame – 2005 to 2015
1. Combine polity and human development data to perform exploratory analysis and discover interesting findings and insights
2. Perform predictive modelling on socio-political features of a country to predict its HDI value and classify it's health coverage access
3. Analyze conflicts and fatalities in a country and correlate them with its social and political conditions that paved way for them.
4. Perform news media impact analysis by finding the most-published topics vs actual political status of a country.
5. Find the percentage of media misconception and fabrication in a conflicted country and highlight it with several country-related metrics.
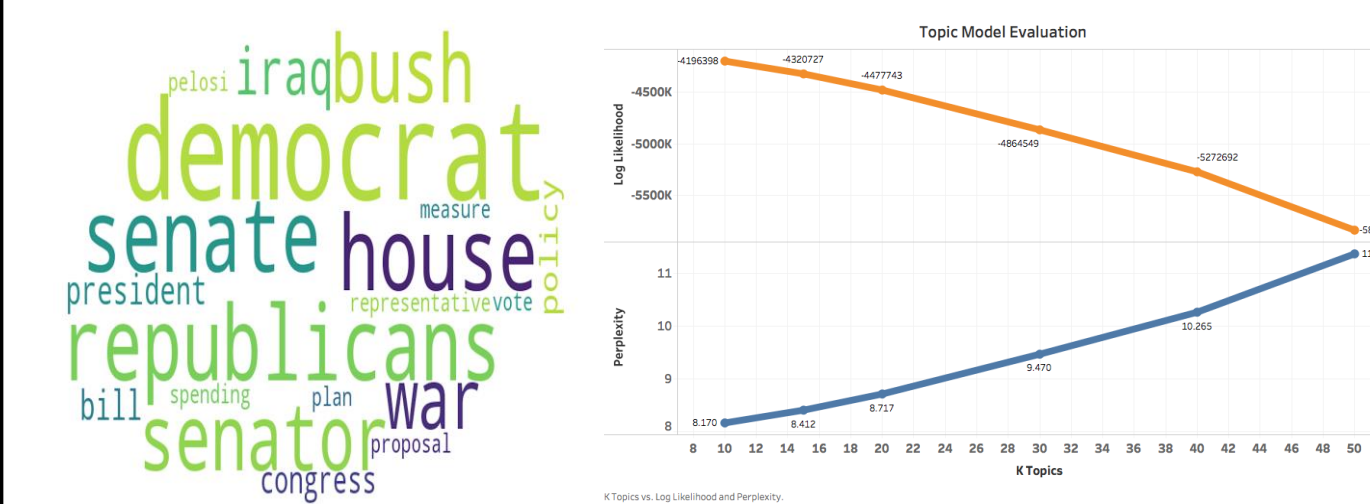
PREDICTIVE MODELLING
- Top features impacting HDI of a country are – **Civil Liberties Index**, **Participatory Democracy Index**, **Electoral Democracy Index, Women Civil Liberties Index**
- **Regression - RMSE** – 0.95 **(Target Variable –** HDI Index**)**
- **Multi-class Classification - Accuracy - 84.7%** **Target Variable** - Health Equality Access **Feature Variables –** Socio-Political Features

MISCONCEPTION ANALYSIS
- A Deep Recurrent Neural Network (LSTM) is trained (accuracy = **87%**) to analyze the news corpus related to particular conflicts in a country and calculate the percentage of genuineness in it, classifying it as FAKE or REAL.

IMPACT ANALYSIS
- **Topic Modelling** is performed on the news articles from year 2005 to 2015.
- **Preprocessing** : NLTK is used to tokenize the sentence, remove the stop words, lemmatize words. Then POS tagger is used to extract Nouns words.
- **Model :** CountVectorizer for feature extraction, LDA for clustering the topics.
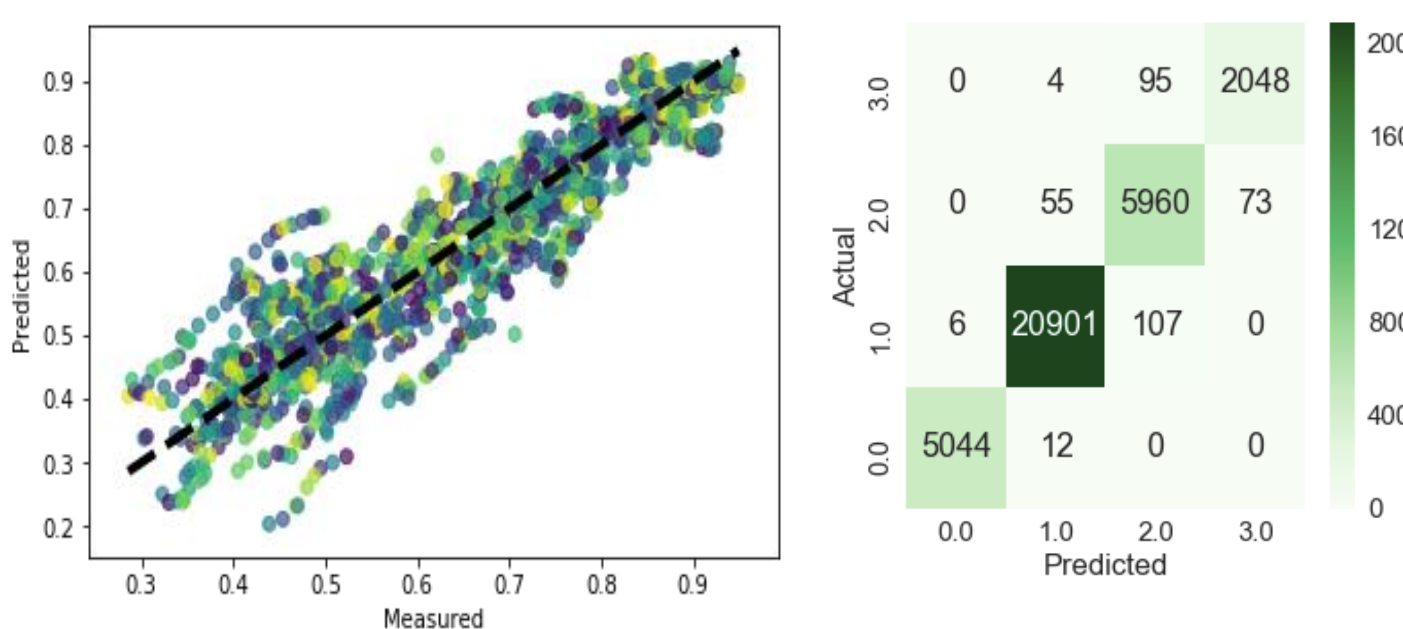- **Evaluation**: **Log Likelihood** -4937939.45 **Perplexity** 8.24

FUTURE WORK
- Including Social Media Impact Analysis in SPAROW, to reflect sentiments about a country and its events.
- Expanding the website to a more generalized format that could be used for users in politics and sociology.

LEARNINGS
- News API does not provide news article older than two months.
- It was difficult to filter the relevant news article specific to a country or an event in NY Times API.
- Preprocessing of the conflicted news headlines was extremely challenging as dataset was loaded with noise.