Socio-Political Analysis In Regions of World

SPAROW

-By-

Anindita Saha

Arul Bharathi

Namita Shah

Motivation and Background:

The rise in number of conflicts, fatalities and mass migration of people across the world due to socio-political events, acted as the primary motivation for this project. During the time of a conflict, countries not only face fatalities but also lose educational and medical coverage to a percentage of its population which directly impacts the Human Development Index (HDI) of a country. In fact, the HDI was created to emphasize that people and their capabilities should be the ultimate criteria for assessing the development of a country, not economic growth alone. The HDI is a summary measure of average achievement in key dimensions of human development: a long and healthy life, being knowledgeable and having a decent standard of living[1]. We focused on HDI as our primary target to achieve a quantitative analysis of socio-political features in this domain. Though there were a lot of qualitative papers related to this study, we created an application that will show the quantitative aspect of this analysis.

Problem Statement:

SPAROW helps to answer the below questions:

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 the country, given its socio-political features ?
3. How does the news media cover and impact prominent socio-political events in a country?

Answering these questions is challenging because there is no consolidated dataset that give answers to all the above questions. Although a lot of qualitative studies have taken place in this domain, we could not find any quantitative analysis on it. Hence, we collected data from various sources and performed a quantitative analysis to answer the above questions.

Data Science Pipeline:

Data Collection & Integration

The data collected from different sources such as V-DEM (Varieties of Democracies), ACLED(Armed Conflict Location & Event Data Project) and UNDP(United Nations Development Programme) was integrated to create a master dataset. V-DEM[2] data corresponds to Socio-Political Quantifiers for each country for every year, ACLED[3] data corresponds to conflicts in countries with type and year of occurrence, and UNDP[4] data consists of HDI values. The second master file that was used for media impact analysis was created by using the UCDP (Uppsala Conflict Data Program) - a dataset containing data points of conflicts that have been derived from news headlines about the conflicts. The news corpus of the news headline was scrapped from New York Times API and merged along with the data points, which were being used for both Topic Modelling and Misconception Analysis. Google Bigquery was used to store the structured data.

We faced challenges in collection as well as integration of the dataset. First of all, it was difficult to gather the datasets from a single source. We had to search a lot of sources and talked to a number of people to finally shortlist the data sources. Though a lot of data are available in the internet, but all the data are segregated. Some datasets had a timeline from 1985 to 2000 whereas some had a timeline from 1990 to 2015.

Multivariate Exploratory Data Analysis:

In this part, we performed multivariate analysis on the first master file that was created. We aimed to get a comprehensive knowledge of what kind of socio-political indicators of a country influence the human development of a country.

Predictive Modelling:

We performed regression and classification on the integrated data files to find HDI value of a country, given its socio-political atmosphere. Also, we predicted the health access quality in each country based on its social and political atmosphere especially during conflicts.

Impact Analysis and Anomaly Detection:

We performed topic modelling to get the most frequent topics published by the news media of the conflicted countries and extracted the media coverage analysis from it. Also, we used deep neural network to calculate the probability of genuineness of each news article published related to the conflicts in each country, to extract the extent of media misconception.

Data Visualization:

We built an interactive web application that serves as an integrated dashboard to be used by political science scholars, enthusiasts and NGOs to visualize their learnings and get more insights and information.

Methodology and Evaluation:

Figure 1 shows a flow chart that explains our methodology used in deriving concrete answers to all three 3 specific questions posed in the objectives.

Multivariate Exploratory Data Analysis:

We used Tableau to get an initial insight about the variables and their influence on HDI values. We used it as an exploratory tool rather than as a visualization tool. There were some interesting answers that we derived using the process listed below:

- Civil Society and Social Regime of a country are prominent aspects that influences the HDI Value. Liberal democratic countries tend to have both high Civil Society Index and Human Development Index [see Appendix C].
- Almost all the countries that have HDI value below than 0.5 are autocratic in nature.
- Democratic countries seem to have higher values of Civil Society Index. varying values. In terms of democratic countries, electoral democratic countries have their HDI values spread out from as low as 0.35 to 0.90, whereas most of the liberal democratic countries have higher HDI.
- Law Equality and Civil Liberty Index are prime influencers of HDI value. Countries that are most free in civil liberties seem to have higher HDI value. We also found out that the law

equality of a country linearly increases with the increasing levels of civil liberties [see Appendix D].

- Socio-Economic Position of individuals places an important role in deciding the human development index of a country. Countries where there is unequal power distribution in political positions seem to have less HDI values than the countries where the power distribution is equal irrespective of socio-political position. This insight seems to be an interesting because the dominance of the wealthy people seem to be a striking factor in deciding the development of the country [see Appendix B].
- Equal Access to resources to individuals is one of the prime indicators of good HDI value in a country.
- Education Equality and Women Empowerment Index [see Appendix A] seems to have great influence in deciding the human development index of a country.

To sum up, a country which gives equal basic education, equal access to resources, equal power distribution in political positions, equal rights to women in society and politics, equality in law and order, and has liberal civil society seems to have higher development index and less conflicts.

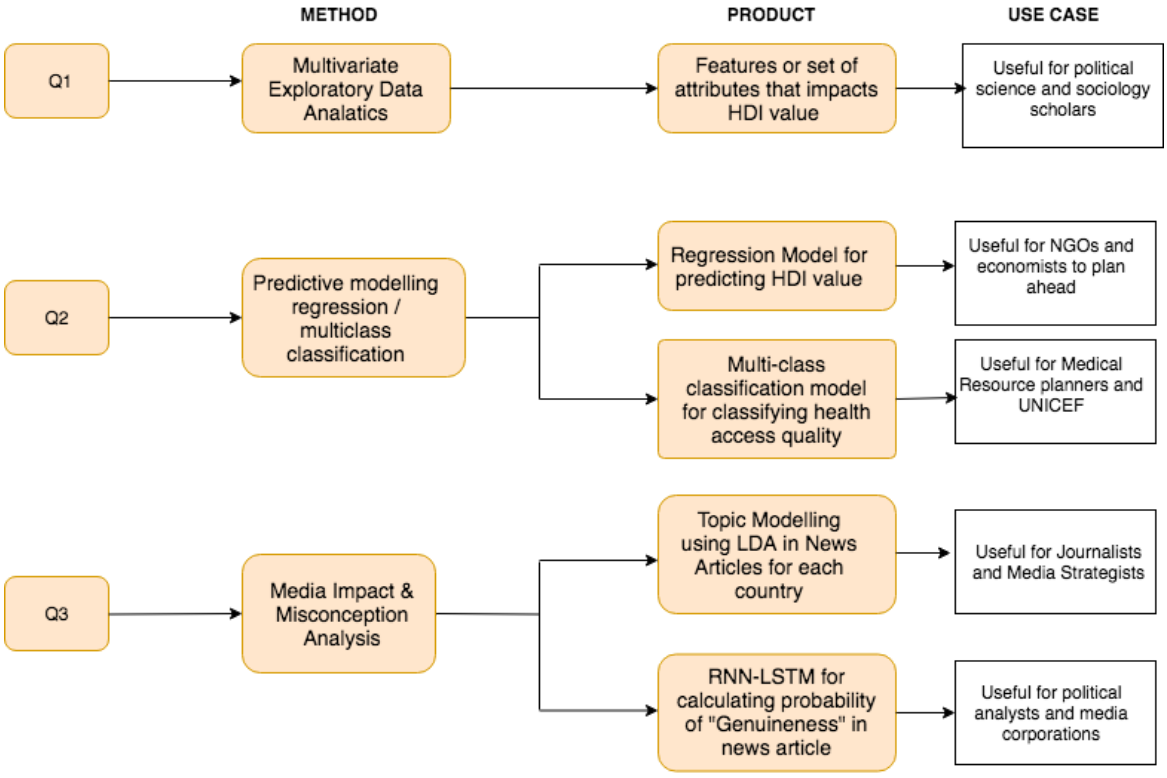


Figure 1

Predictive Modelling:

In predictive modelling, we predicted the **HDI index** and **Health Equality Access** given the socio-political features of a country. The dataset “Conflict_polity_master” was used for this purpose. A total of 33 categorical and 23 continuous features were taken from the dataset for predicting the HDI index. The missing values (which were only 2 in the whole dataset) were removed and the categorical values were converted to dummy variables which were then joined to the main dataset.

2/3rd of the dataset was used as training set and the rest was kept for testing. We applied **linear regression** model and achieved a RMSE score of **0.952962** on the test data. The most significant features that directly impacted the HDI value are “Civil Liberties Index”, “Participatory Democracy Index”, “Electoral Democracy Index”, and “Women Civil Liberties Index”.

Health Equality Access is a categorical variable which depicts the extent of high quality basic healthcare guaranteed to all, which is sufficient to enable adult citizens to exercise their basic political rights. Its values are explained below:

0: Extreme. Because of poor-quality healthcare, at least 75 percent (%) of citizens’ ability to exercise their political rights as adult citizens is undermined.

1: Unequal. Because of poor-quality healthcare, at least 25 percent (%) of citizens’ ability to exercise their political rights as adult citizens is undermined.

2: Somewhat equal. Because of poor-quality healthcare, ten to 25 percent (%) of citizens’ ability to exercise their political rights as adult citizens is undermined.

3: Relatively equal. Basic health care is overall equal in quality but because of poor-quality healthcare, five to ten percent (%) of citizens’ ability to exercise their political rights as adult citizens is undermined.

To predict Health Equality Access, we used Multiclass classification. We used **Random Forest Classifier**, **Linear SVC** and **Logistic Regression** to apply cross validation and found out that Logistic Regression was giving us the best result with a mean accuracy of **0.845854**. The confusion matrix was plotted using matplotlib and seaborn.

Media Impact Analysis:

In this section, we calculated the impact of media coverage and misconception of news articles. We used topic modelling to find the variety of topics that were covered in conflicted country’s news and calculated what percentage of them talks about socio-political situation of the country.

Topic Modeling:

Dataset: It was collected from “The New York Times” articles. We used NYTimes Api to get the news headlines and web urls. We took top 50 news details of 12 countries of each month from year 2005 to 2015. Then, corpus of those news was scrapped from web url using lxml.

Preprocessing: Used NLTK a Natural Language Processing python library for preprocessing the news article. First we tokenized the sentence then removed all the stop words. Used lemmatizer to lemmatize each words. Since, we wanted Noun words for better topics we also used POS tagger.

Feature Extraction and Model Construction: Latent Dirichlet allocation (LDA)[5] is a probabilistic model that discovers the topics out of the documents. The words from preprocessing step were converted into vector of token words using CountVectorizer and fed to Pyspark’s ML LDA model to perform topic modeling on the extracted features. It produced a topic distribution matrix as output. LDA was performed on each country individually to find out what are the main topics that are published in the new york times article. And we found out that in countries like Sudan even though there was lots of conflicts and most of the topics were related to sports. And in country like Iraq, there were lots of conflicts and most of the news were related to politics and wars. Used word cloud to visualize all the top topics.

Evaluation: Used log likelihood and Perplexity to evaluate our model. For all the countries model performed better when used 10 topics. The perplexity was low for k=10 and log likelihood was high.

Misconception Analysis using Recurrent Neural Network

Dataset: The second master file explained in the data integration part was used as the test data set for the misconception analysis. The training data set on which the model was trained is a crowdsourced dataset containing labelled data of Fake and Real news. We did this analysis using references from existing kernels on this data set.

Preprocessing: The news corpus was preprocessed with multiple mechanisms. We replaced multiple period with a single period, removed white spaces, removed punctuations and stop words in order to make the corpus more meaningful.

Mechanism: We used **Glove vectors** for text analysis and used Recurrent Neural Network with LSTM. Custom embedding was used for the analysis of the corpus. The model consists of 62 LSTM layers one 1 output layer with sigmoid activation function. The main function used for calculating the genuineness is "Binary Cross Entropy" and the optimizer used is **rmsprop** with **Accuracy** as the evaluation metric. We reached validation accuracy of **86%**. We calculated the ratio of fabricated and genuine news in a conflicted country and also calculated the ratio be used on the dashboard.

Data Product:

Our data product is a web application that acts as an integrated dashboard that displays various metrics and insights about conflicted countries. There are specific options to select each country and each country page will be loaded dynamically in the application generating plots and insights. The web application is built using Flask API with bootstrap. We retrieved data from google BigQuery and processed the data in Pandas and processed that as Jinja template and used the template in creating dynamic UI charts using Google Charts.

Lessons Learnt:

- 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.
- Old Twitter data was not available for free and many important tweets were not in English, hence, could not perform sentiment analysis.

Conclusion

Our analysis on the socio-political features clearly showed us the importance of women empowerment, power distribution in politics, basic education equality, international influence, media coverage, etc. in the development of a country. With our data product, it is now clearly visible that what factors in which way directly impacts the HDI. This gives us visibility of the features which are really important in overall development of human being in a country. The news media coverage helped us identify how news media fabrication can impact a country. Our analysis showed that most of the conflicted countries had a significant level of media misconception and fabrication in them. The predictive modelling part of our project gave us a clear understanding of most important features that are directly impacting HDI. It also showed us that given certain socio-political features of a country, other factors such as 'health equality access', can be predicted easily.

References:

[1]. http://hdr.undp.org/en/content/human-development-index-hdi

[2]. https://www.v-dem.net/en/data/data-version-7-1/

[3]. https://www.acleddata.com/

[4]. http://hdr.undp.org/en/data

[5]. https://spark.apache.org/docs/2.2.0/ml-clustering.html#latent-dirichlet-allocation-lda

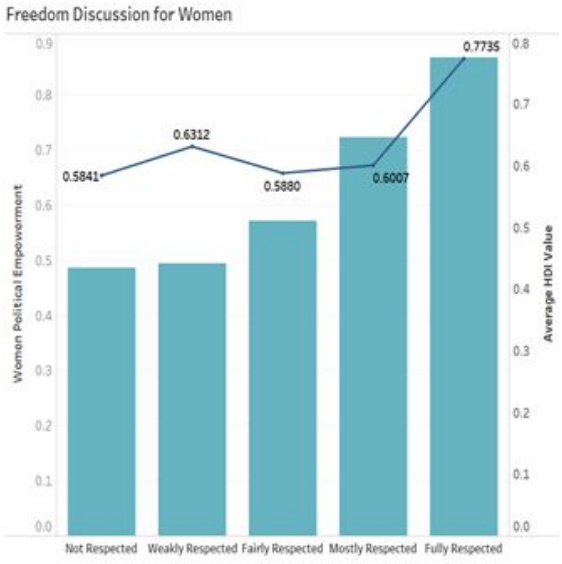
[6] https://www.datacamp.com/community/tutorials/scikit-learn-fake-news

[7] https://www.datasciencecentral.com/profiles/blogs/on-building-a-fake-news-classification-model

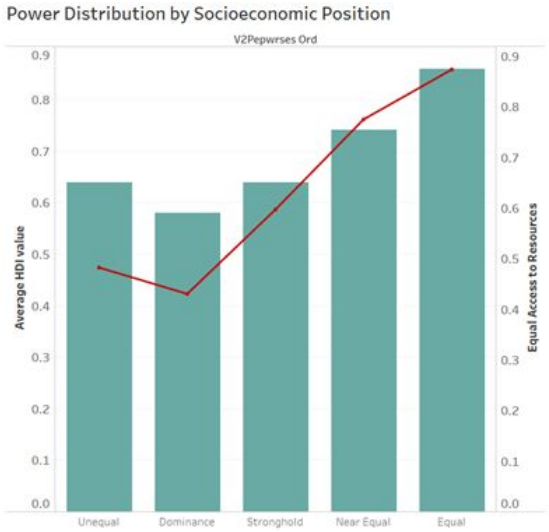
[8] https://towardsdatascience.com/fake-news-classifier-e061b339ad6c

Appendix:

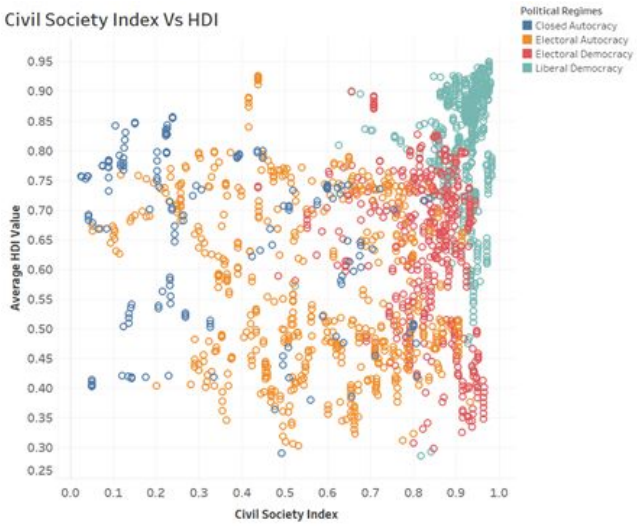
[A]



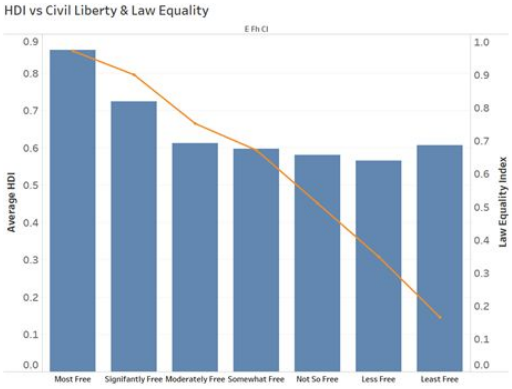
[B]



[C]


Civil Society Index Vs HDI
Political Regimes
Closed Autocracy
Electoral Autocracy
Electoral Democracy
Liberal Democracy
Average HDI Value
Civil Society Index

[D]


HDI vs Civil Liberty & Law Equality
Average HDI
Law Equality Index
Most Free
Signifantly Free
Moderately Free
Somewhat Free
Not So Free
Less Free
Least Free