World Trade Index and Policy Effect Analysis using GDELT

Deepika Sanjay Khatri
Computer Science Dept., Courant
NYU
dsk420@nyu.edu

Shreya Santosh Bamne
Center for Urban Science and
Progress, NYU
ssb602@nyu.edu

***Abstract*—**

News showcase various events happening around us and their effect on our environment. One of the huge collections of news events across the world can be found in Global Dataset on Events, Language and Tone (GDELT) along with information about location, actors and tone of the event. Every news event irrespective of its location of origin has an impact on different parts of the world. We plan to show these impacts through our analysis of GDELT data through a heat map for different events.

Keywords—GDELT, HeatMap, Big Data Analytics

I. Introduction

The amount of news articles generated every moment globally is enormous and GDELT, created by Kalev Leetaru, collates all this data from conventional as well as non-conventional news sources and distills it into three major datasets, namely "Events", "Mentions" and "Global Knowledge Graph". Our aim is to be able to track the effect of a policy or a major movement across the world in terms of their subjective mass appeal as well as their economic impact in terms of the Goldstein scale. We will visualize this analyzed information using three popular events in terms of a choropleth map and a bubble intensity chart. The end game is to be able to visually see the progress of an event and/or policy.

II. Motivation

GDELT includes a GoldStein Scale in its dataset that explains about the cooperation and/or conflict of a news event. This informs only about the event for that country/region.. Visualizing the reaction of the world in a choropleth map, helps in knowing and comparing the magnitude of impact across the globe. It makes it easier to understand the broader picture for them who are not aware of the formal methods of measuring impacts of news events.

III. Related Work

There are many studies that have used GDLET in their research including news events. The news events comprise of various types of incidents, politics is undoubtedly a major category. One of the study used GDELT to forecast Domestic Political Crisis(DPC).[1] The paper proposes a graph based methodology to forecast DPCs based on GDELT dataset (from January 2003 to December 2013). The problem is treated as a classification problem. It uses frequent subgraphs for detecting and forecasting events from news events using classification. The approach was effective with the only caveat of computational complexity.

Another paper talks about the procedure to join the three datasets of GDELT[3]. While joining the Events and the Mentions dataset is easy as they both have a common identifier, the real issue is joining them with Global Knowledge Graph. We cannot proceed with our project, if we don't know how to join the three datasets together. This paper aims to provide a solution to construct a unified database easily and quickly by using the MentionIdentifier from Mentions and the DocumentIdentifier from Global Knowledge Graph. The paper talks about two phases.

The first phase is easy and simply and it asks us to join the Events and Mention Table using the identifier GlobalEventID that is present in both the datasets. Let's call this joined database Event_Mentions. After performing an inner join, it highlights two problems for us.
The Identifier that we want to use to join the Event_Mentions and Global Knowledge Graph datasets on is an URL. Since URLs are lengthy, the join will be time consuming process and could even lead to memory leakage problems.
The Global Knowledge Graph is considerably larger in length per record compared to the Mentions_Event.

The proposed solution is to preprocess the identifiers, MentionIdentifier from Event_Mentions and the DocumentIdentifier from Global Knowledge Graph, by using MD5 hash function. The result of the hash is relatively small compared to the size of the URL. Once you have obtained the dataset, there are several visualizations that could be obtained. The paper talks about heatmaps of coverage of events through news articles and bubble maps to show the detailed location on the map.

IV. Application Design

Figure 1 shows an overview of the flow of application. We have stored our files on Hadoop Distributed File System (HDFS). The files are then read into Spark environment where they are cleaned to get rid of irrelevant columns. The cleaned data is then profiled to make sure that the columns have the right data types. The profiles data is stored to a Spark dataframe. The further analysis is performed on this dataframe. The final output is then written to a csv file which is written to HDFS. The results are then visualized using Tableau.

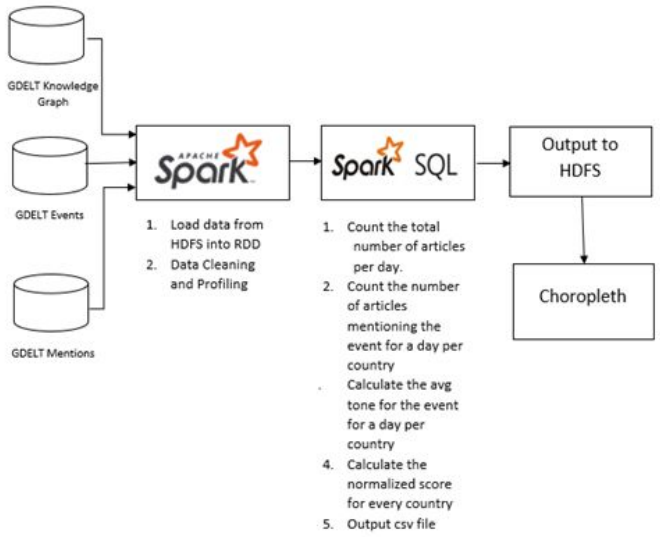

Fig 1: Design Flow Diagram of Analysis

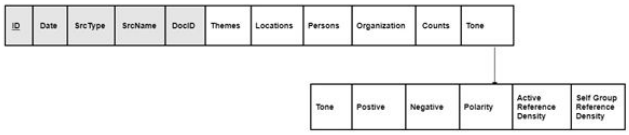

Fig 2: Schema of the GKG Dataset

Event ID	Date	Actor1 Code	Actor1 Name	Actor1 CountryCode	Actor2 Code	Actor2 Name	Actor2 CountryCode	IsRootEvent	Goldstein Scale	Avg Tone	Cameo Code	Num Mentions	Source URL

Fig 3: Schema of the Events Dataset

Event ID	Event Date	Mention Date	Mention Type	Actor1 Offset	Actor2 Offset	Action Offset

Fig 4: Schema of the Mentions Dataset

V. Datasets

We are using three datasets provided by GDELT:

1. GDELT Global Knowledge Graph (GKG): As the name suggests, this dataset comprises graph of global events. It is updated every 15 minutes and is available in 65 different languages. It also includes information about average tone that specifies the impact of an event. For our analysis, we plan to use the GKG 2.0 version with events in English.
2. GDELT Events Data: This dataset comprises of spatio temporal information of news events. Like GKG we are using the GDELT Events 2.0 version.
3. GDELT Mentions Data: This is an additional dataset added to the GDELT database since 2015. It has information regarding mentions of a event in the Events data.

VI. Remediation

With respect to our visualizations and analysis results for the particular week, we can compare the event intensity and the trend of the world trade index. A pattern emerges where we can see that when the event is most talked about, the world trade index takes a hit. Depending on the tone of the event, the world trade index correspondly responds. There is a lag in the response and it can be seen playing out in the world stock market. One application would be to track these indices and when the intensity of the event increases, a global investor should exit out of the equities market and move towards safer fixed interest stocks like bonds.

VII. Experiments

We plan to look at the dispersion of a news event across the globe and also analyzing the average tone at the country level. The analysis has been done using the Global Knowledge Graph, Events and Mentions datasets. For our analysis, we have considered one specific news event. We decided to look at the reaction to the tax tariffs imposed by the US on Aluminium and Steels goods. In order to find the event in each of the dataset, we perform following steps:

1. GKG: We look at the themes column. It generally has many tags in it. Examples of tag include X_POLITICAL_PARTY;TAX_POLITICAL_PARTY_DEMOCRATS;USPEC_POLITICS_GENERAL1;LEADER. For our analysis we will filter on themes starting with 'ECON_'. Further filters can be done on Actor as 'Donald Trump', 'US', 'China', etc

2. Events Dataset: We will filter using the cameo codes. GDELT has assigned different cameo codes for different types of events. Examples of cameo codes are: '01' is for Making a

Statement and has subcategories of the form '01x' depending on type of statement made.

3. Mentions: The event IDs of the events identified in events dataset will be used to filter the mentions file.

Before we filter for the events, we need to clean the data to remove irrelevant columns. This will make the data manageable and help in saving some space. The data cleaning for the datasets is done as follows:

For analysing the GKG dataset, we first computed the world trade index from January to June by calculating the average tone of the filtered articles per day and the number of filtered articles per day. A problem with this data is that the date format YYYYMMDDHHMMSS isn't compatible with the spark format and has to be striped of the HHMMSS to make it usable. To normalize this data, we divide it by the total number of articles per day. Here the articles don't originate in the country but rather we are accounting for those countries that are being mentioned in the article, thus either being affected or their policies affecting other countries. The World Trade Index is then mapped as a simple line graph. We also look at the country level by calculating the normalized average tone per day per country. This data is then mapped using a bubble size chart.

While the SQL dataframe does provide GROUPBY to sort the data, it is comparatively slow. Given more time, we would find a better parallelized code to perform this function. A further research area would be to find the causality of the events and predict the behaviour of stock market. The root event code in the events dataset can be utilized to expand on to that.

Events Dataset: 18 out of 58 columns were retained for the final analysis. The data type for numeric columns was changed from String to Integer/Double. The columns that were converted include – AvgTone, Goldstein scale, NumMentions and Is RootEvent This was done as we would plot the final choropleth map using these numeric columns. In addition to this, there is an additional filter on event codes based on cameo codes manual and also on country using the codes specified in the United Nations List. [6]

Mentions Dataset: 8 columns were retained for the final analysis. This dataset required some type conversions too like the events dataset.

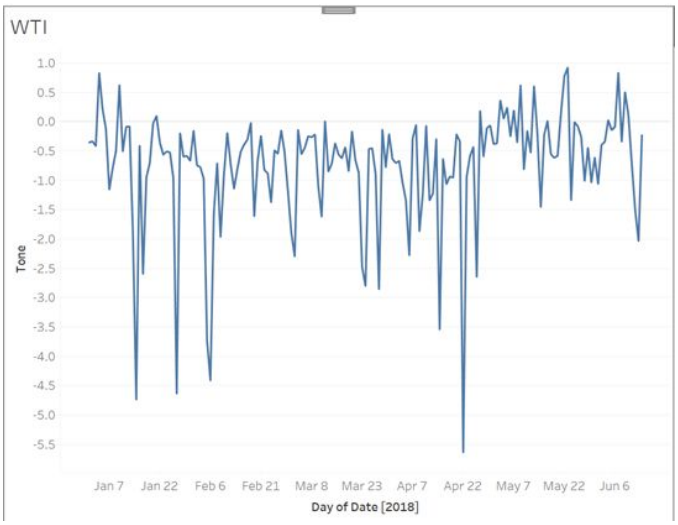

Fig 5: World Trade Index

Usually, the world trade index is more neutral and trending around the baseline of zero but with the tariffs announcements, it has huge spikes towards the negative tonality. As you can see, every dip in the world trade index is justified by an event happening just the week or a few days before it. For example:

- Mar 23 dip: Aluminum and Steel Tariffs go into effect.
- Apr 2: China Retaliates
- Apr 22 dip : On Apr 17, China imposed tariffs on US Sorghum
- May 7 – 22 neutral: China ends tariffs on US sorghum during negotiations
- June 1-6: US moves ahead with tariffs on Canada, EU and Mexico revoking their exemptions.

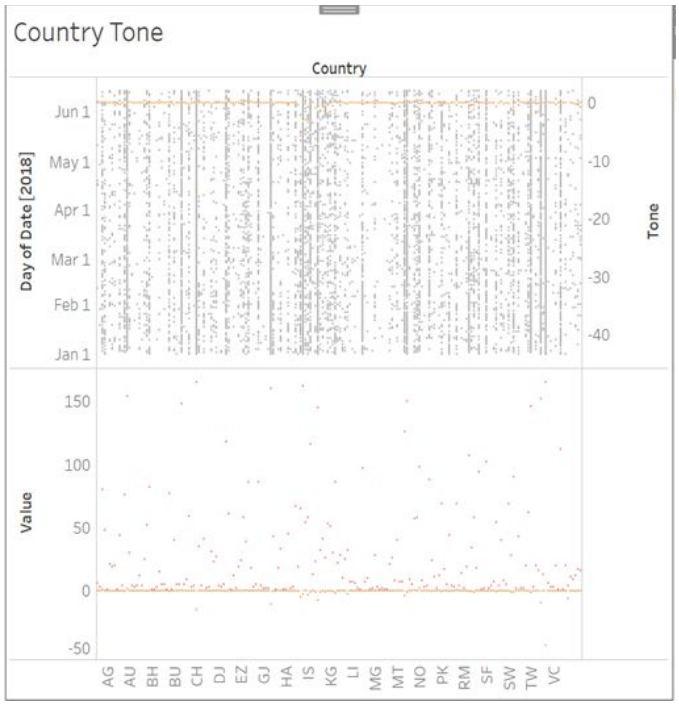

Fig 6: Country Tonality

From the following visualization, we can clearly see the top 5 countries with the most negative tone as:

- US
- China
- Germany (car exporter)
- UK (steel sector)
- Japan (Japanese automakers invested in US Auto-plants)
- West Bank (1/3rd of the economy dependent on exports)

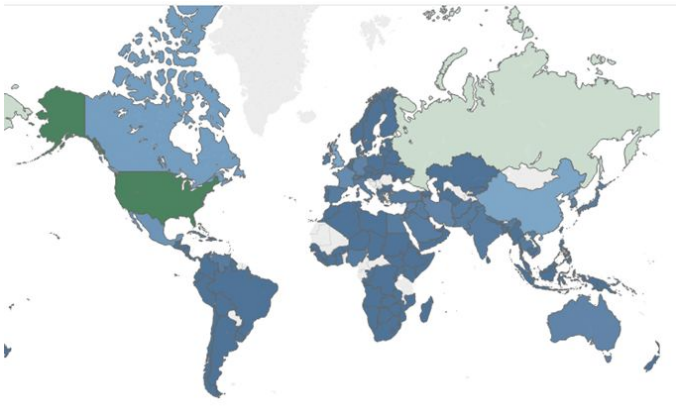

Fig 7: Choropleth of Number of Mentions of Tax Tariff Event based on first two weeks of March data

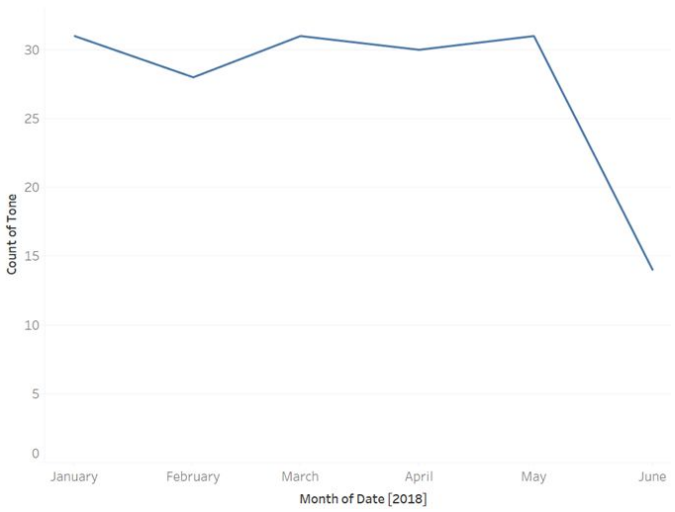

Fig 8: Average Tone of USA

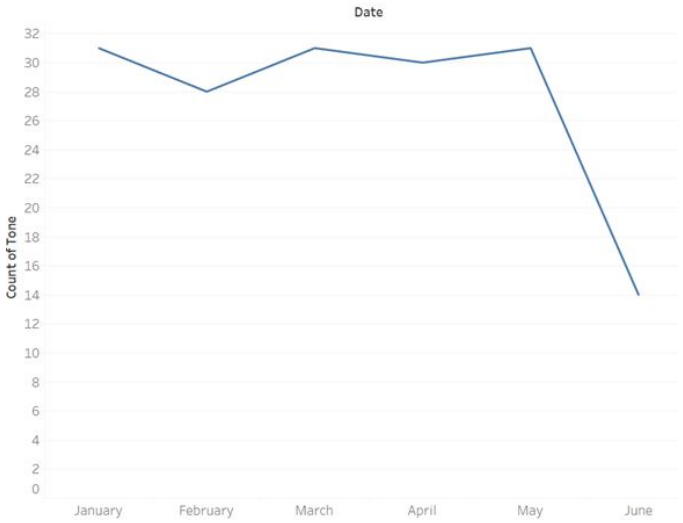

Fig 9: Average Tone of China

As we can see from the Tone Graphs of USA and China, their tone is negative around the same time when big changes in the tariff policies are announced by either governments.

VIII. Conclusion

In conclusion, our analysis of tax tariffs matches up with the global reaction unfolding across the globe. This method can be used to track the dispersion of any policy, once you have identified the right themes and cameo codes to track it in the GDELT database.

Acknowledgment

We would like to extend our thanks to Google for making this available for free to anyone who wants to play with it. The data is relatively clean and can be modelled to various degrees. We would also like to extend our gratitude towards Prof. Susan McIntosh who introduced us to this dataset and encouraged us to explore it.

References

1. Yaser Keneshloo, Jose Cadena, Gizem Korkma, Naren Ramakrishnan, Detecting and Forecasting Domestic Political Crises: A Graph-based Approach, WebSci '14 Proceedings of the 2014 ACM conference on Web science, June 2014.
2. . Hadoop: The Definitive Guide. O'Reilly Media Inc., Sebastopol, CA, May 2012.
3. Kedi Chen, Feng Qiao, Hui Wang, Correlation Analysis Using Global Dataset of Events, Location, and Tone, IEEE First International Conference on Data Science in Cyberspace, June 2016.
4. Kiran Sharma, Gunjan Sehgal, Bindu Gupta, Geetika Sharma, Arnab Chatterjee, Anirban Chakraborti, Gautam Shroff, A complex network analysis of ethnic conflicts and human rights violations, Scientific Reports, August 2017.
5. Haewoon Kwak, Jisun An, Two Tales of the World: Comparison of Widely Used News Datasets GDELT and Event Registry, Proceedings of the Tenth International AAAI Conference on Web and Social Media, 2016
6. https://unstats.un.org/unsd/methodology/m49/