MEDIA FRAMING
The bias in international news coverage

Arnaud Garin – Reza Hosseini – Virginie Piskin
Applied Data Analysis

Abstract

With the development of new technologies, people have access to tremendous amounts of sources to keep up with the events of the world. This increase of sources made it nearly impossible for people to get a sense of how biased the media are in transferring the facts and events. With an increasing number of sources, there is increasing risks of misinformation.

The goal of our project is to raise awareness of the potential bias of news sources in each country and develop a visual representation to demonstrate it. To that aim, we used the dataset provided by the GDELT project which contains the key information needed to investigate international news coverage. Analyzing the locations of the news sources and their targeted themes using techniques such as clustering or PCA, we were able to prove the existence of bias in media coverage. Our results are presented in various maps showcasing the different characteristics of media coverage in different countries.

Introduction

In less than two decades, the development of technology has completely changed our society in the way we communicate and share information. Obviously, the emergence of the Internet was the main factor in this revolution but the phenomena was also boosted by the increasing quality in hardware. Not only do we have a free access to the Internet but we also have a wireless connection almost everywhere on various devices with increasing computing power (computers, smart phones or tablets). The downside of this phenomena is the difficulty in handling so much data and being able to extract meaningful information while keeping track of the accuracy of the information.

We were inspired by this concern as individuals but also as computer scientists which is why we chose to work on media coverage. For this task, we used the database provided by the GDELT project which tracks all the articles written on the Internet about a specific event happening in the world. This database allowed us to carry multiple analysis to study media framing by answering two key questions, helping us grasp media bias. Nethertheless, we have also tried to come up with our own mesure of bias using information such as the average tone of an article.

Key Questions

1. *How differently are international news covered depending on the country or the media?*

We studied the distribution of the mentions around the world to determine which countries or areas are the most covered and those that are left out by news sources. For this task we counted the number of articles per geographic location and computed the corresponding overall proportion.

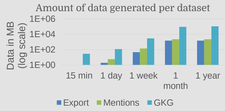
2. *Which category of news tends to be framed or less covered depending on the country or the media?*

Inspired by the bag of word approach we have created a bag of themes which essentially counts the occurrence of each themes per media source. From this matrix we were able to make a clustering of media sources by first applying a PCA to extract latent features and then use the K-means algorithm.

Material

The database provided by the GDELT Project (Global Database of Events, Language, and Tone) is an endless source, containing terabytes of data collected over three years (from 2015 to 2018) and still expanding. GDELT set up a system able to scrap the Internet every 15 min and keep record of the news shared during this period. The data is stored on the Big Query server provided by Google which allows anyone holding an account to make SQL requests. This is very convenient yet it can quickly become very expensive. Therefore, GDELT made it also possible to access the data freely by providing packages of compressed files to be downloaded from the Internet:

- **Export:** contains information to identify one specific event (ID, date, latitude and longitude coordinates, URL of one article about this event, average tone of the articles about this event and identification of the actors, countries or organization, taking part in the event.
- **Mentions:** contains information to identify the articles written about one specific event (event ID, date of the article, URL of the article, tone and score on the Goldstein scale).
- **Global Knowledge Graph:** GDELT developed algorithms able to parse the articles written about one specific event and extract very useful information (locations, people, countries or organizations mentioned in the article, themes of the subjects treated in the article, videos or images embedded in the article).



Duration	Export [0.3,0.5] MB	Mention [0.6,1.5] MB	GKG [15,30] MB	All (rounded)
Quarter	0.5 MB	1.5 MB	30 MB	32 MB
Hour	2 MB	6 MB	120 MB	128 MB
Day	48 MB	144 MB	2.9 GB	3 GB
Month	1.5 GB	2.25 GB	90 GB	93.75 GB
Year	18 GB	27 GB	1080 GB	1.13 TB
Total (3 years)	54 GB	81 GB	3 TB	3.3 TB

Methodology

1. Finding the country of mentions source:

In order to find the source country we used two tricks:

a. Looking into domains' extensions
b. Using IP-Whois to get the country of the source registered domain

2. Investigating the themes for topic modeling:

We used topic modeling techniques to investigate the overall context of news.
The steps are:

a. Creating the bag of words from all news themes
b. Using PCA/SVD decomposition to get the topics
c. Investigating the 4 top frequent topics by visualizing their top themes

3. Investigating the distribution and intensity of events' coverage

In this part we focused on news source from one country (U.S.A) and investigated their coverage of events from other countries. In order to do that, we did the following steps:

a. Removing noise from Lat/Long of event's action locations
b. Getting countries of news sources
c. Plotting on geo-map

4. Investigating the bias:

We defined a new feature named "bias" using 'Average Tone' feature from each event and 'Doc Tone' from each mention. We then, plotted the the weighted average bias using choropleth map.

Results - How differently are international news covered depending on the country or the media?

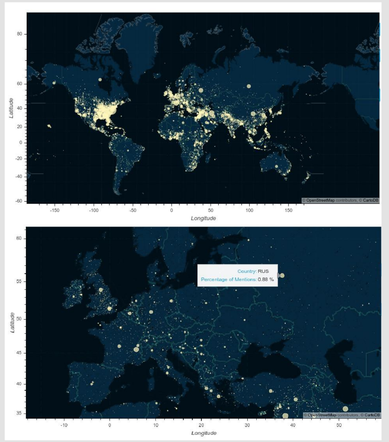


Figure 1a – 1b: Distribution of the world's news coverage from U.S. media sources

Explanation: This visualization is an interactive map on which we can zoom in and out to see the locations of different events. The scaling of the dots represents the proportion of articles about the events: the bigger the dot the more important is the coverage. By hoovering above the events we can have the proportion in percentage.

Interpretation: This representation shows that the distribution of the events is centered on the USA, western Europe and on the some of the middle east countries.

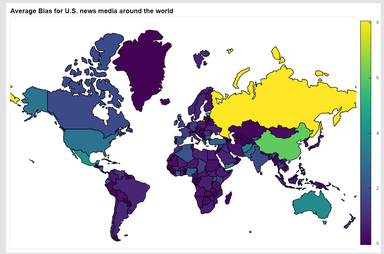


Figure 2: Average bias for US media sources around the world

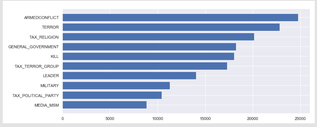
Explanation: This map is a visualization of the bias in US news coverage per country using a bias metric we computed ourselves.

Interpretation: It is pretty clear that there is a higher bias when countries such as Russia or China are concerned by the news, with whom the US has a political and/or economical conflict of interest.

Results (*continued.*)- Which category of news tends to be framed or less covered depending on the country or the media?

Figure 3: K-Means clustering of media sources according to their themes

Figure 4: Theme composition of one latent feature



Explanation: The first image shows the result of the PCA followed by the K-Means clustering of the media sources according to the theme occurrence. The second image presents the composition in terms of themes of one of the latent feature extracted from the PCA dimensionality reduction.

Interpretation: This visualization depicts the existence of two major groups of media who each supports distinct notions. Also, by investigating the most frequent themes in each topics, we found the top 4 topics as: "**Government and public services**", "**Environmental and educational** (future concerns)", "**Socio-politics**", and "**Military and terrorists related news**"

Conclusion

In this project, we used GDELT v2.0 dataset in addition to advanced data analysis tools in order to answer our research questions.

We first analyzed the U.S media news coverage to investigate any pattern or discrepancy of coverage for different countries . We saw a pattern that there are a lot of news from U.S., Europe, and some middle east countries.
Then, we introduced a new features named "average bias" to explore how positively or negatively, compared to the average, U.S. media sources talks about other countries. We clearly saw a high negative bias for countries that United States or their allies are in conflict with (namely Russia, China, Mexico, and Yemen).
We also explored news themes and tried to find the most spoke topics. By investigating the most frequent themes in each topics, we found the top 4 topics as: "Government and public services", "Environmental and educational (future concerns)", "Socio-politics", and "Military and terrorists related news".

As for future steps, one can extend the usage of the define "average bias" to investigate which countries might be in conflict with each other. Also, by combining our method for detecting the clusters of media news with "average bias", one can investigate which countries each media in one clusters are talking positively/negatively and try to find any pattern from their supports.