DS4A FINAL PROJECT | WRITTEN REPORT
THE COLOMBIAN ARMED CONFLICT
In the eyes of the world

[redacted]

1. Introduction and Business Context

Numerous attempts of ending Colombia's armed conflict have existed since its start, in the mid-1960s. Due to the complexity of its social, economic and political dynamics, it has been the target of countless research projects derived from academic and diplomatic efforts to understand its causes and effects on Colombia as a society and as a nation. On November 24, 2016, after 4 years of negotiations, the Colombian government signed a Peace Agreement with the FARC guerrilla[1], one of the main actors in this armed conflict. However, there is still a high level of uncertainty towards the outcome of this treaty as there has been a lot of criticism and controversy around it, as well as an atmosphere of emotional distress and polarization that has taken over Colombian society. This situation is what brought Colombia's armed conflict to the global spotlight again.

With the rise of new technologies that allow collecting massive amounts of data and tools to understand the emotions involved in it, analyzing complex social and political phenomena through data has become an increasingly interesting field of study. Google Jigsaw's GDELT Project[2] uses these technologies, along with several others, to monitor news in any format (broadcast, print and web) all around the world.

With this dataset and complementary data on various economic indicators such as the USD/COP exchange rate and the price of an oil barrel, we aim to analyze Colombia's Armed Conflict in terms of its global perception and its effects on economic indicators that are commonly related to the perception of a country's stability. In other words, our project derives from the question:

WHAT IS THE GLOBAL PERCEPTION OF COLOMBIA'S ARMED CONFLICT ? HOW DOES IT RELATE TO THE PERCEPTION OF ITS SOCIAL AND ECONOMIC STABILITY?

[1] http://www.altocomisionadoparalapaz.gov.co/procesos-y-conversaciones/acuerdo-general/Paginas/inicio.aspx
[2] https://www.gdeltproject.org/

This question should be answered gradually according to the next project versions:

Component #1: Analyze how the perception of Colombia's Armed Conflict (and its main actors) has changed over time and across different countries.

Component #2: Analyze the relationship between the percent variance of the USD/COP exchange rate and other economic indicators (price of an oil barrel, for example) and the social indicators included in the GDELT dataset, such as the number of news articles published and the tone these were written with.

Component #3: Conduct an anomaly analysis on the number of news articles and their tone to identify interesting changes in the trends and relate them to specific political and social situations.

2. Executive summary of results

The main question that drives this project was: What do countries around the world think of Colombia's Armed Conflict? This conflict has ignited harsh debates in Colombian public media. But how does an outsider feel about the events it comprises?

A secondary question was: while economic indicators can easily measure social and economic stability, so how does the armed conflict relate to them?

To answer these questions, we used the GDELT Project databases. This project has data about news articles from all over the world, including the source, tone and actors involved in the events. It is updated every 15 minutes and it has data since 1979.

We found that in the world, countries have a different perception of the issue of armed conflict in Colombia. Through word cloud analysis we identify which are the relevant issues for each country.

We also noticed that the overall tones of the articles seem to be more negative when the events jeopardize the peace agreement or exacerbate the war.

We built an app that could be useful to any citizen that has an interest in exploring how global news communicate the armed conflict in Colombia, but this tool could have the same impact in any country regarding any social conflict.

Also, this could allow decision-makers to know the local narratives around social problems and address them with concrete solutions.

And our most important finding: however volatile economic indicators can be, the topics, tone and number of articles around global news can definitely guide an analysis of these indicators. Of course, further work needs to be done to reach a more reliable result.

We found through multivariate linear regression, a statistically significant negative relationship between the tone and the price of the dollar.

We predicted the dollar value next day using a Neural Network with a sprt(MSE) less than $140

We found by anomaly detection (or outlier detection) that the number of articles and tone are indicators of relevant events. Anomaly detection (or outlier detection) is the identification of rare items, events or observations which raise suspicions by differing significantly from the majority of the data.

3. Application

3.1 Application overview

The application has three components:

1) Perception across countries: a temporal space analysis of how the events associated with the armed conflict in Colombia have been perceived by the rest of the world
2) Indicators and Insights: it let us explore the relation with some economic indicators
3) Interesting anomalies, it let us identified some events with tone and numbers of articles outliers

Using GDELT the world's largest news database, the application allows us to identify insights to explore and validate periods and situations, related to the armed conflict in Colombia from the point of view of news from each country in the world, to recognize behaviors, concepts or strategies that contribute to your solution. That is, it allows identifying the issues that are considered relevant in the face of the armed conflict in Colombia, from the perspective of other countries.

3.2 User interaction

Through the selection of the period to analyze, and a measure of the tone of each news item (positive, negative or neutral), we can see how the armed conflict in Colombia has been seen in the world or in the country that is selected.

According to the selection of the mentioned variables, the application displays a world map with the prevailing tone for each country.

Additionally, the application presents the following results for the period, tone of the news and selected countries:

- Average tone over time
- Number of articles over time
- Word cloud
- Relevant actors or organizations

A user can easily choose the period of interest, within which the event under study was presented, specify if they want to analyze the news that were markedly positive, negative or a combination of them, choosing the tone (a number between -10 : negative to +10: positive) and if you want the country under analysis.

It also interactively displays the average tone behavior over time, for the selected input variables. This chart gives insights about dates when significant events occurred for the rest of the world.

The interest in the situation of the “Armed Conflict in Colombia” for a particular country or for the whole world is measured by the number of articles published.

The word cloud graphic tells us about the relevance of the issues associated with armed conflict from the eyes of the world.

Finally, we can see the actors or organizations that were significant in the news spread throughout the world.

3.3 Application models

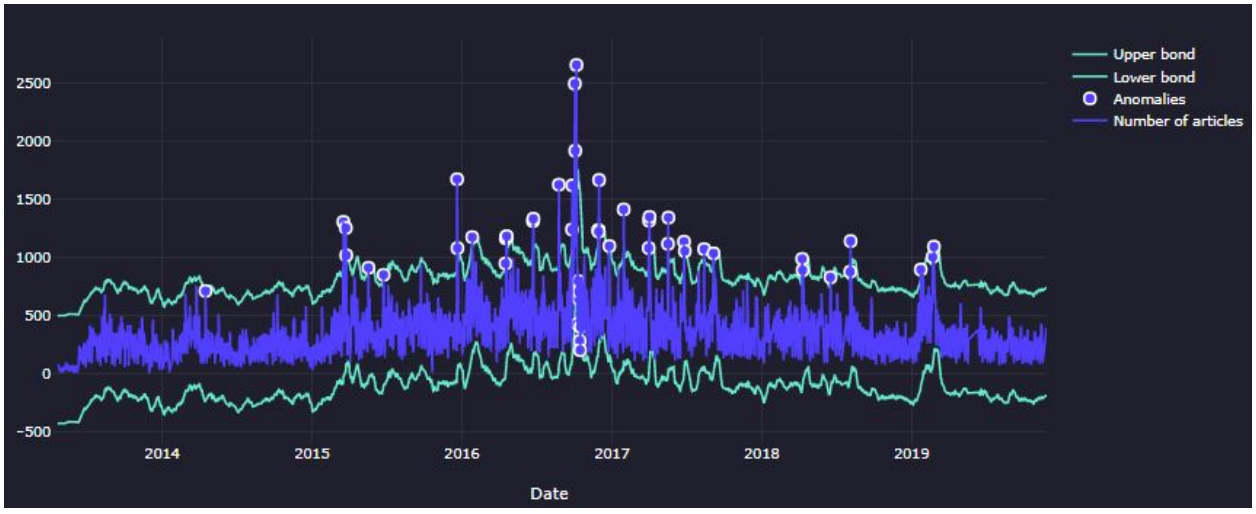
Anomalies Model

Firstly, we created two time series, one for the number of articles and the other one for the average tone, both in the time interval between april of 2013 and present. Once we have this we calculated the rolling mean of the time series of the tone and number of articles. This with the purpose of smoothing the original series and identify trends. In order to do this we used a window of seven days. Then we calculated the mean absolute error between the original series and the rolling mean, as well as the standard deviation of the difference between both of them. With this information we calculate the lower and upper limits of the interval as follows: .

$$lower\_bond = rolling\_mean - (mae + scale * deviation)$$
$$upper\_bond = rolling\_mean + (mae + scale * deviation)$$

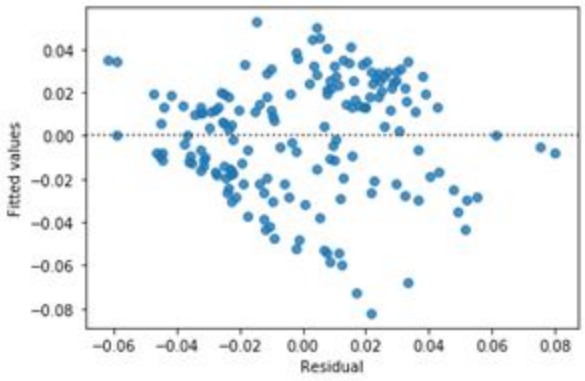
Where scale was set to 1.96.

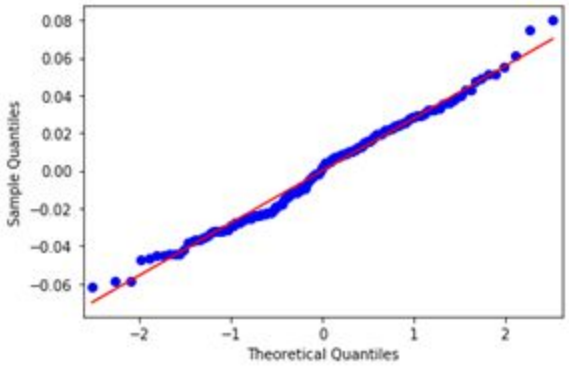
Having defined the confidence interval we could identify the moments when the average tone and number of articles leaves its normal behavior (anomalies). We found that there is a consistency between the identified moments with situations that were relevant to Colombian history. For instance in the following chart we could identify that during late 2016 and early 2017 Colombia was in the spotlight of the world and these dates correspond to Colombian peace agreement process and our pre and post conflict.



RL model to correlate the price of the dollar (TRM) with the all average of issues associated with the armed conflict in Colombia

We discovered that the daily average tone of several themes was correlated to the USD-COP exchange rate in the last year; therefore, we implemented a linear regression model.





This is the correlation matrix of some of the variables

Variable	agriculture_1	np.log(colcap)	oil_price	self_identified_humanitarian_crisis_2_lag_4
agriculture_1	100%	34%	6%	-41%
np.log(colcap)	34%	100%	25%	-23%
oil_price	6%	25%	100%	22%
self_identified_humanitarian_crisis_2_lag_4	-41%	-23%	22%	100%

The results of this model were not conclusive because around 50% of the variability of the USD-COP currency exchange can be explained with variables from the dataset. Since we had highly correlated variables and the necessary resources we then decided to implement a NN to predict the currency exchange rate.

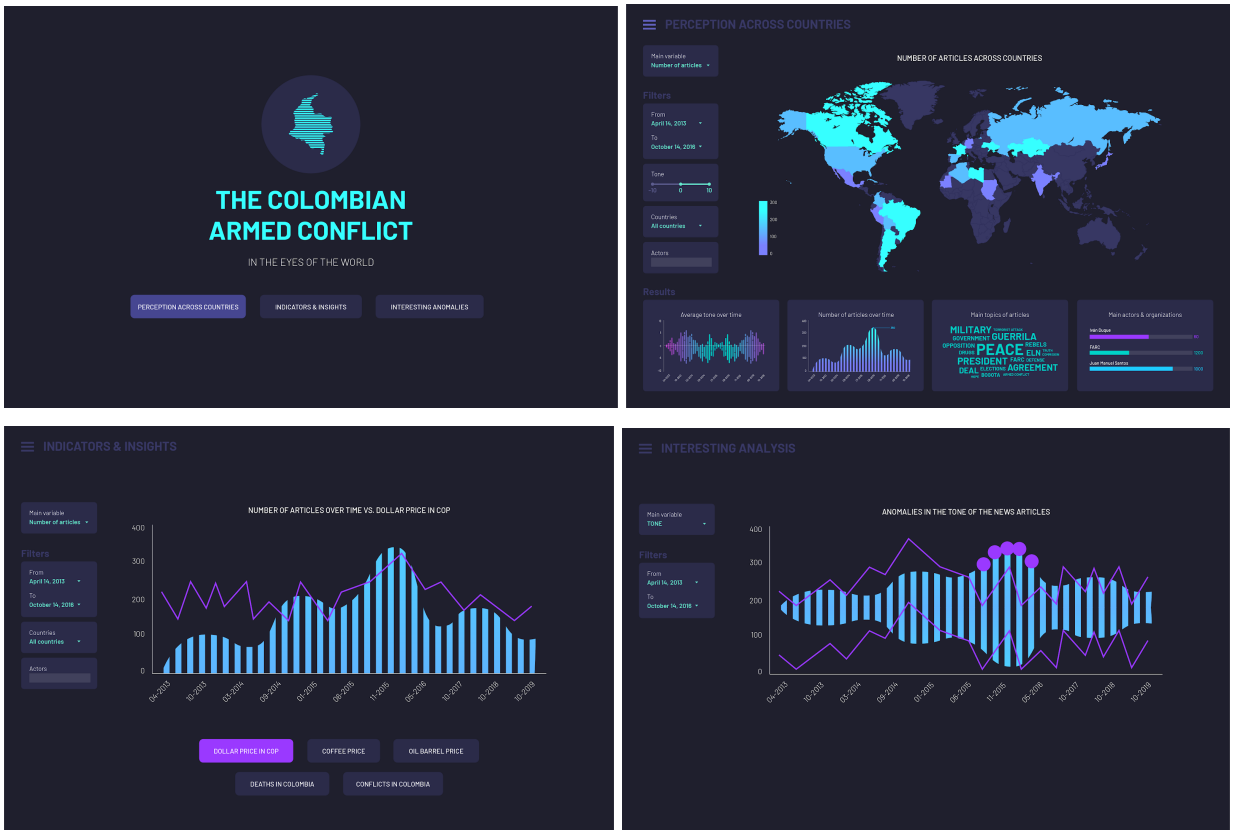
ML model for dollar price prediction from tone

We found a statistically significant negative relationship between the average tone of the issues associated with the armed conflict in Colombia and the price of the dollar, that is, the greater the negative perception of the news are, the greater the value of the Representative Market Rate.

4. Technical Exposition

4.1 Interactive Front-end

The frontend was designed in Adobe Illustrator first, in order to ensure the functionalities and design decisions were truly centered in the user's experience, rather than the restrictions of the frameworks and libraries. This was the initial design:



The web application is hosted in an EC2 instance, in the same network as the RDS instance, which makes the response time much faster. With every new user input event on the front end, a query is sent to the Postgres database to retrieve the selected data. And we do not store any information about our users.

For the implementation of the dashboard we used Dash, HTML, CSS, and JavaScript. The visualizations were made using mostly the Matplotlib library and have the following characteristics:

- World map colored according to scale of values for the average tone in the selected period and tone.
- Line graph to show the dynamics of the average tone over time.
- Line graph to show the dynamics of the number of related articles in the period and all selected, for one country or for all countries.
- Word cloud to present insights of the relevant topics during the selected period, tone and country.
- Histogram of the importance of actors or organizations measured by the frequency in the appearance of news.

These are actual screenshots of our current dashboard:

4.2 AWS-hosted Database

GDELT Project Datasets

The GDELT Project reports the people, locations, organizations and events involved in our everyday news. It includes reports as old as January 1, 1979 and as new as 15 minutes ago, which is the amount of time that passes between every new update to the database. It provides two datasets: the GDELT Event Database and the GDELT Global Knowledge Graph (GKG). The first one includes physical events of social and political nature reported by the news media all over the world, specifying the people, locations and organizations involved. This dataset will provide us with information about the number of events that were reported in the news from 2007 to 2018, the tone in which those were reported and the categories to which these events belong to.

On the other hand, the GKG focuses on the context in which an event happens, providing "a list of every person, organization, company, location and several million themes and thousands of emotions from every news report"[3]. This dataset will help us understand how other countries perceive Colombia's Peace Agreement and the main actors that have participated in its creation and its current implementation.

[3] Ibid.

THE GDELT EVENT DATABASE DATA FORMAT CODEBOOK V2.0, presents an overview and descriptions of the fields in de GDELT Event. It is important to highly that "GDELT Event records are stored in an expanded version of the CAMEO format (see below), capturing two actors and the action performed by Actor1 upon Actor2. A wide array of variables break out the raw **CAMEO** actor codes into their respective fields to make it easier to interact with the data, the Action codes are broken out into their hierarchy, the Goldstein ranking score is provided, a unique array of georeferencing fields offer estimated landmark-centroid-level geographic positioning of both actors and the location of the action, and a new "Mentions" table records the network trajectory of the story of each event "in flight" through the global media system." GELTD data descriptions see TABLE 1: GDELT FIELD DESCRIPTIONS below.

GDELT data use Conflict and Mediation Event Observations (CAMEO) coding for recording events.
http://data.gdeltproject.org/documentation/GDELT-Data_Format_Codebook.pdf

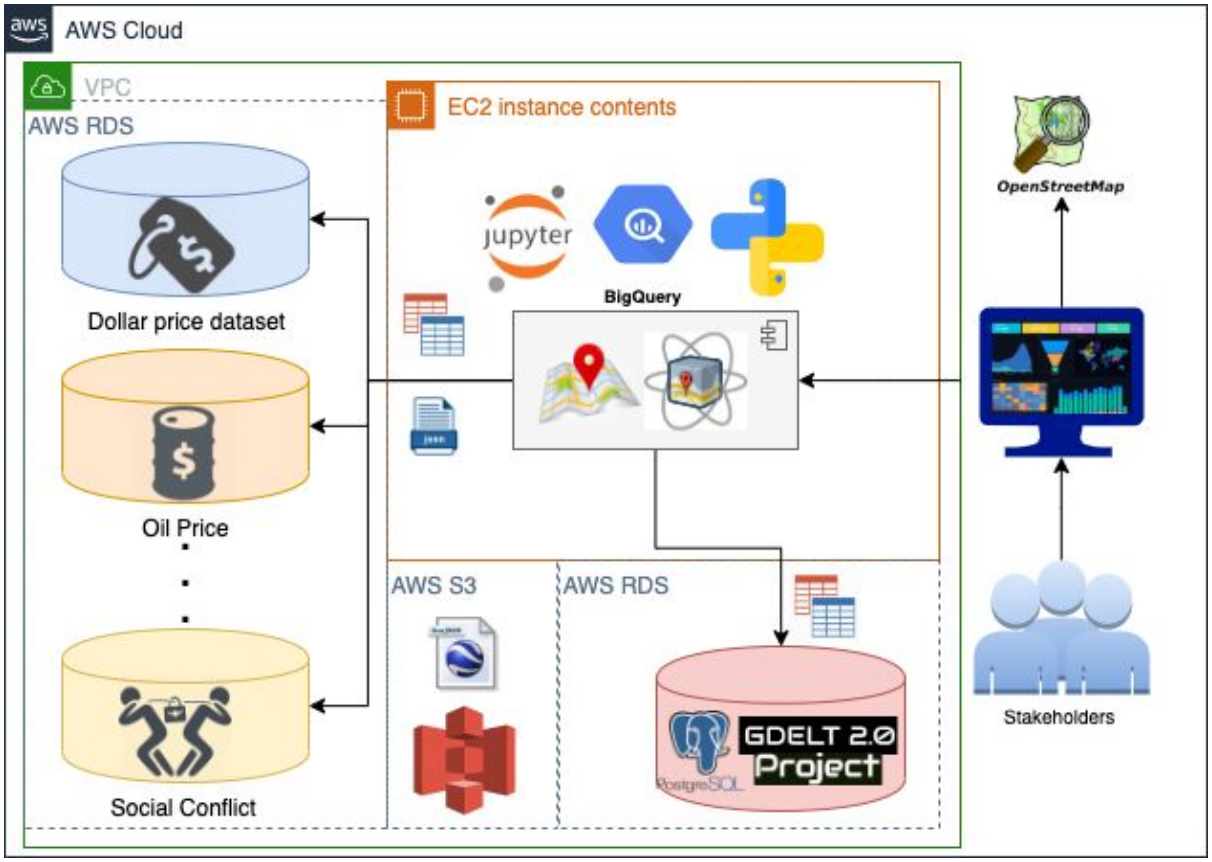
Dollar price dataset

Currency devaluation is a strategy that several countries use for reducing their purchasing power. Some countries as China use currency devaluation to gain a competitive edge in global trade, as a strategy for its trade war with the United States. In Colombia, the dollar price is closely related to the price of oil, which is our main export. However, it also responds to political phenomena that indicates the country's stability in the eyes of foreign investors.

Oil Price Dataset

The oil price can vary by decisions about output made by OPEC (Organization of Petroleum Exporting Countries), when a combination of stable demand and oversupply has put pressure on oil prices over the last five years. Other causes are related with natural disasters that could potentially disrupt production, and political unrest in an oil-producing juggernaut like the Middle East all impact pricing[4]. We found a dataset with data about crude oil prices per barrel back to 1946. This dataset is available at https://www.macrotrends.net/1369/crude-oil-price-history-chart. Another source for oil price history is https://datahub.io/core/oil-prices#resource-oil-prices_zip that has a daily value from May-1987 until October-2019.

[4] https://www.macrotrends.net/1369/crude-oil-price-history-chart



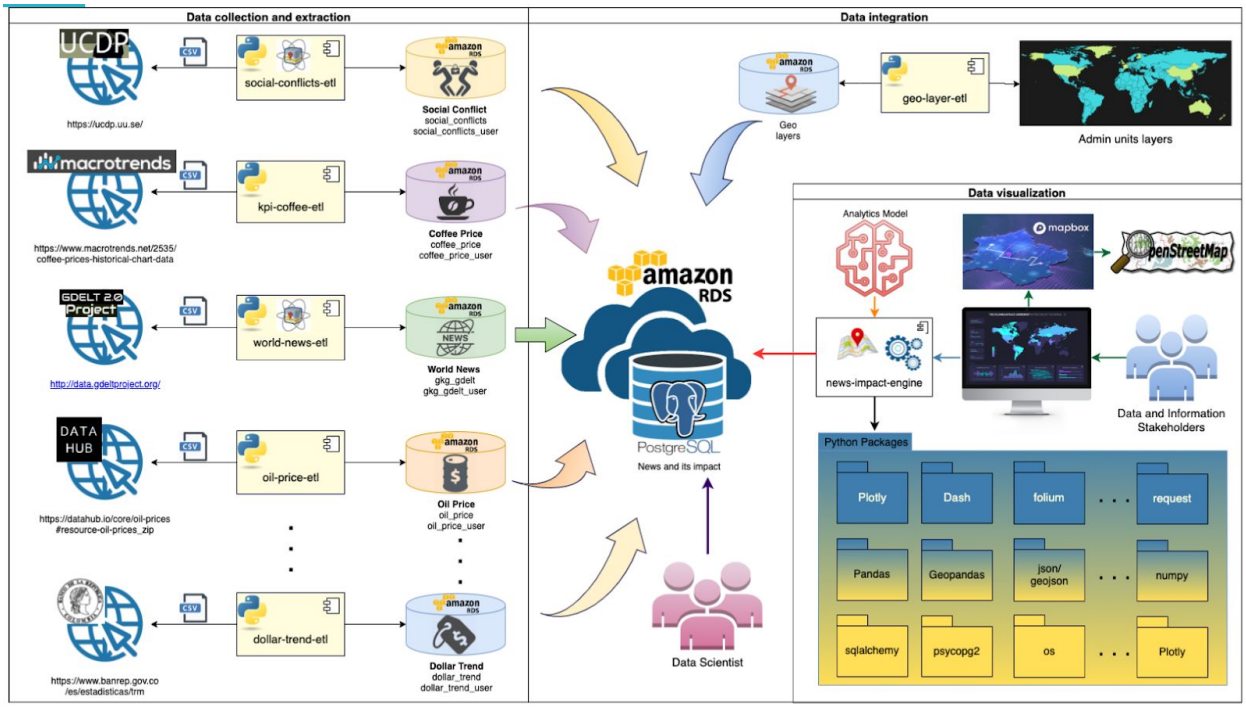
4.3 AWS-hosted Data Analysis & Computation

We downloaded almost a hundred gigabytes of data and filtered it to remove unrelated news. We automated this process with a Python script that updates the database every day.

Some of the filters we applied to make sure that the data is actually related to the armed conflict in Colombia are:

- Human rights
- Inequality
- Terror group
- Justice
- Security

One of the biggest challenges we had was the fact that we were working with big data and we wanted to make an interactive dashboard that process huge amount of data in real time. About 1 million articles are being updated daily in gdelt and today our own DB have 134 GB of data processing. Even though we have an amazon cloud instance of 8 cores our queries were taken about 30min. However, we could reduce the time to less than a second per query. Using full-text search applying natural language processing and partitioning data techniques.



We have different data sources, including GDELT, macro trends, and others. GDELT has its data available on BigQuery or on csv files and since the first one was expensive and we had the resources for the second one we went for it. This was a trade off, considering that this last one is a costly process and we ended up with 134 GB of data processing. We created a data lake that currently has 50 GB and it is being updated on a daily basis. Finally we created an interactive dashboard that is running on an amazon cloud instance so that people can see and interact with our app.

5. Conclusion and Answer to Initial question

We found that in the world, countries have a different perception of the issue of armed conflict in Colombia. Through word cloud analysis we identify which are the relevant issues for each country.

We also noticed that the overall tones of the articles seem to be more negative when the events jeopardize the peace agreement or exacerbate the war.

We built an app that could be useful to any citizen that has an interest in exploring how global news communicate the armed conflict in Colombia, but this tool could have the same impact in any country regarding any social conflict.

Also, this could allow decision-makers to know the local narratives around social problems and address them with concrete solutions.

And our most important finding: however volatile economic indicators can be, the topics, tone and number of articles around global news can definitely guide an analysis of these indicators. Of course, further work needs to be done to reach a more reliable result.

We found through multivariate linear regression, a statistically significant negative relationship between the tone and the price of the dollar.

We found by anomaly detection (or outlier detection) that the number of articles is an indicator of relevant events.