Defamation on Colombian Electoral Campaigns

Santiago Bacaro Bravo

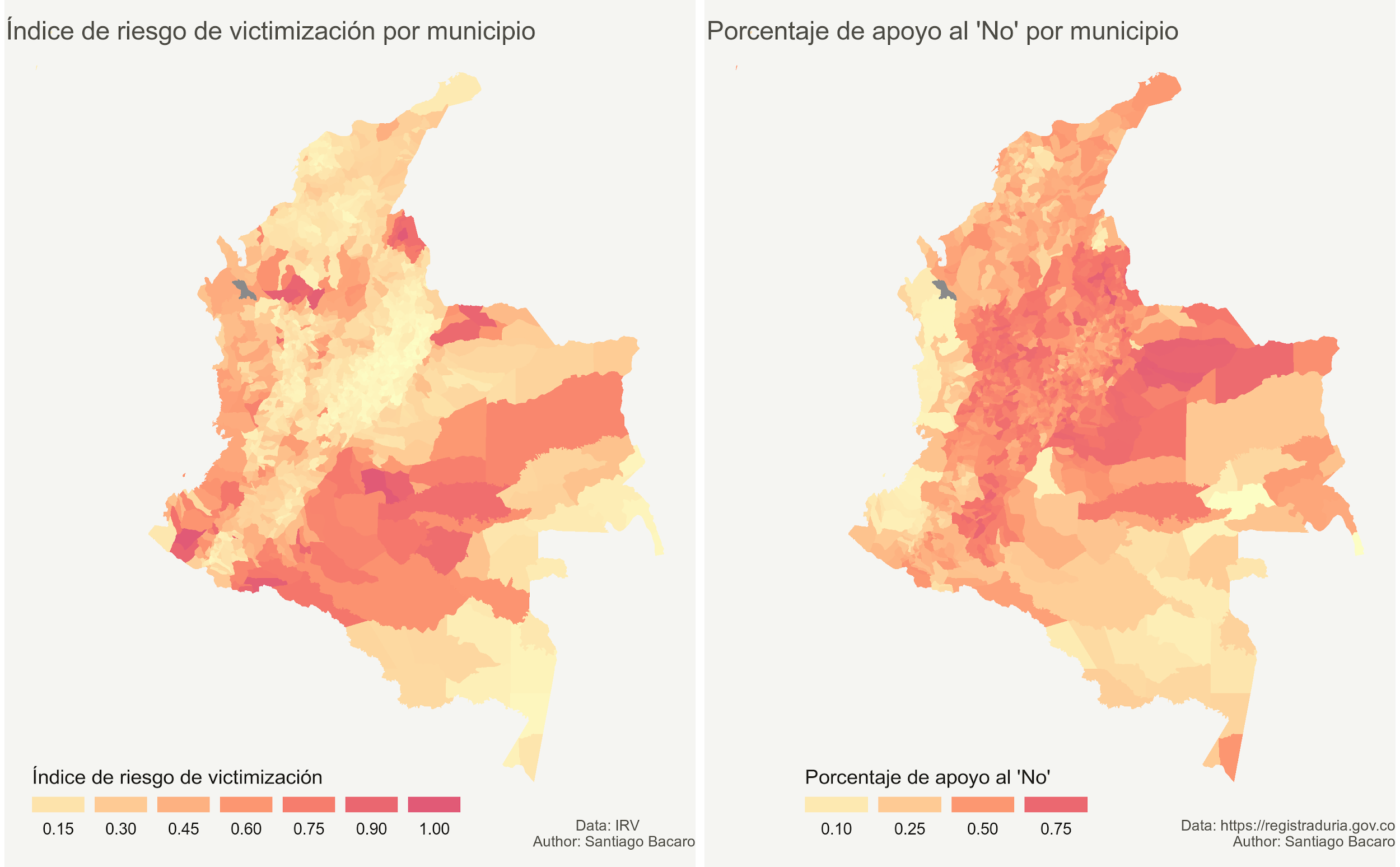
# Introduction

Colombia’s political situation has been polarized for a very long time. Historically the polarization had been between the conservative and liberal parties, but, as new parties arose, the violence was perpetuated against them.

In 1957, the “Pacto de Benidorm” was signed, which gave way to the “Frente Nacional”. This agreement stated that the power was to be equally distributed between the liberal and conservative parties. Meaning that for one period the president would be from the Conservative party, and for the subsequent period, they would be from the Liberal party.

This enforced discrimination against all other parties, which led to the constitution of guerrillas. One of the guerrillas that arose during that period, was the “Fuerzas Armadas Revolucionarias de Colombia (FARC)”, which until very recently was the oldest guerrilla in Latin America. A couple of years ago, a peace agreement was proposed between the State and this guerrilla. Fundamentally, the members of the guerrilla wanted to be able to participate in politics, as well as be pardoned for some of the crimes they committed as a guerrilla. This, however, did not include humanitarian crimes, or crimes against the fundamental human rights.

Surprisingly enough, this deal, which would mean the end of war with this particular guerrilla, was turned down by the people. Here are the results:



This visualization compares the percentage of votes against the peace agreement, with the “Índice de Riesgo de Victimización” (IRV). The IRV, is an index that indicates how likely are the people from the different municipalities, to suffer from violence caused by the internal war. According to its description, it takes into account actions against life, personal freedom, personal integrity, safety, and “free circulation”[[1]](#footnote-0).

Ironically, the two maps seem as though they were a negative of each other. The municipalities with the lowest IRV scores, were the ones that voted against the peace deal the most. Whereas, the municipalities most afflicted by the internal war, had the fewest percentage of votes against the agreement.

The hypothesis for this project is that the results of that voting were influenced by misinformation, defamation and fake news. And its purpose is to identify if the same happened with the presidential elections of 2018.

The presidential candidates who made it to the final stage of the elections were:

* Gustavo Petro: a leftist candidate, who promotes the rights of the poorest and least favoured. He is a former sympathizer of the government of Hugo Chavez, and he has some socialist ideas himself.
* Iván Duque: a right-wing candidate. This candidate came out of nowhere. His previous work had been at the “Banco Interamericano de Desarrollo” (BID). And he had not made any political career whatsoever. He is, however, the candidate of the “Centro Democrático” party, led by Álvaro Uribe Vélez. Uribe is a far-right politician, who has been linked with paramilitarism and whose presidential candidates have been linked with corruption scandals, such as the one with Odebrecht. Álvaro Uribe was the president of Colombia for two consecutive periods and wanted to be re-elected once more, but the constitution prevented it, which he also tried to change. Juan Manuel Santos was then elected, he was Uribe’s minister of defense, and turned his back on Uribe after winning the elections.

# Methodology

This project contains the following stages:

* Collection of data of the results of the Plebiscito por la Paz.
* Analysis and visualization of the results, comparing them with the IRV scores.
* Collection of Twitter data for the period on which the electoral campaigns took place
* Preprocessing of the data
* Analysis of the data

## Collection of data of the results of the Plebiscito por la Paz

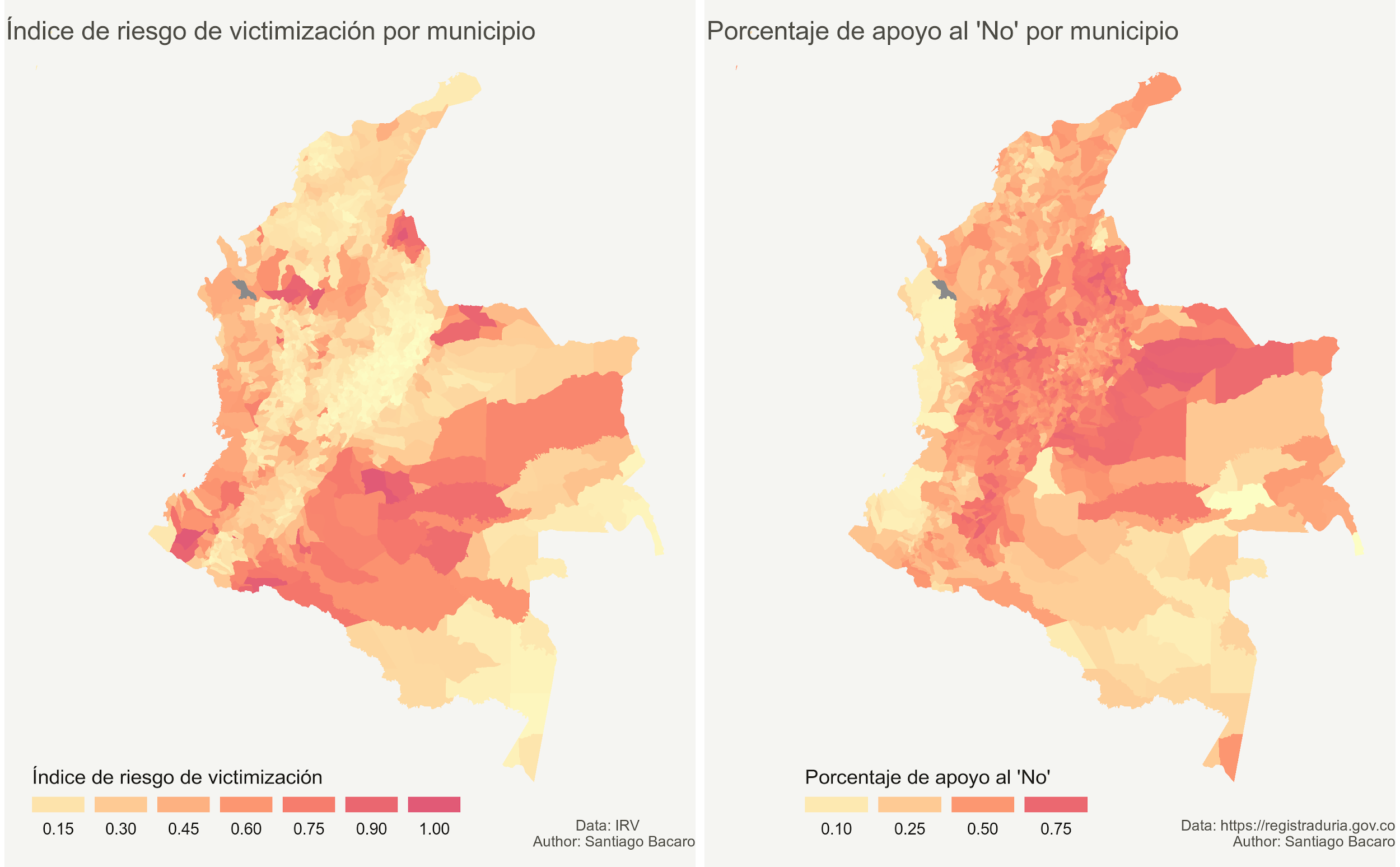
This collection of data was done manually, from the website of the results, available at: <https://www.registraduria.gov.co/?page=plebiscito_2016>. However, you will find the extracted data in the data/ directory in a file called plebiscito\_mpios.csv.

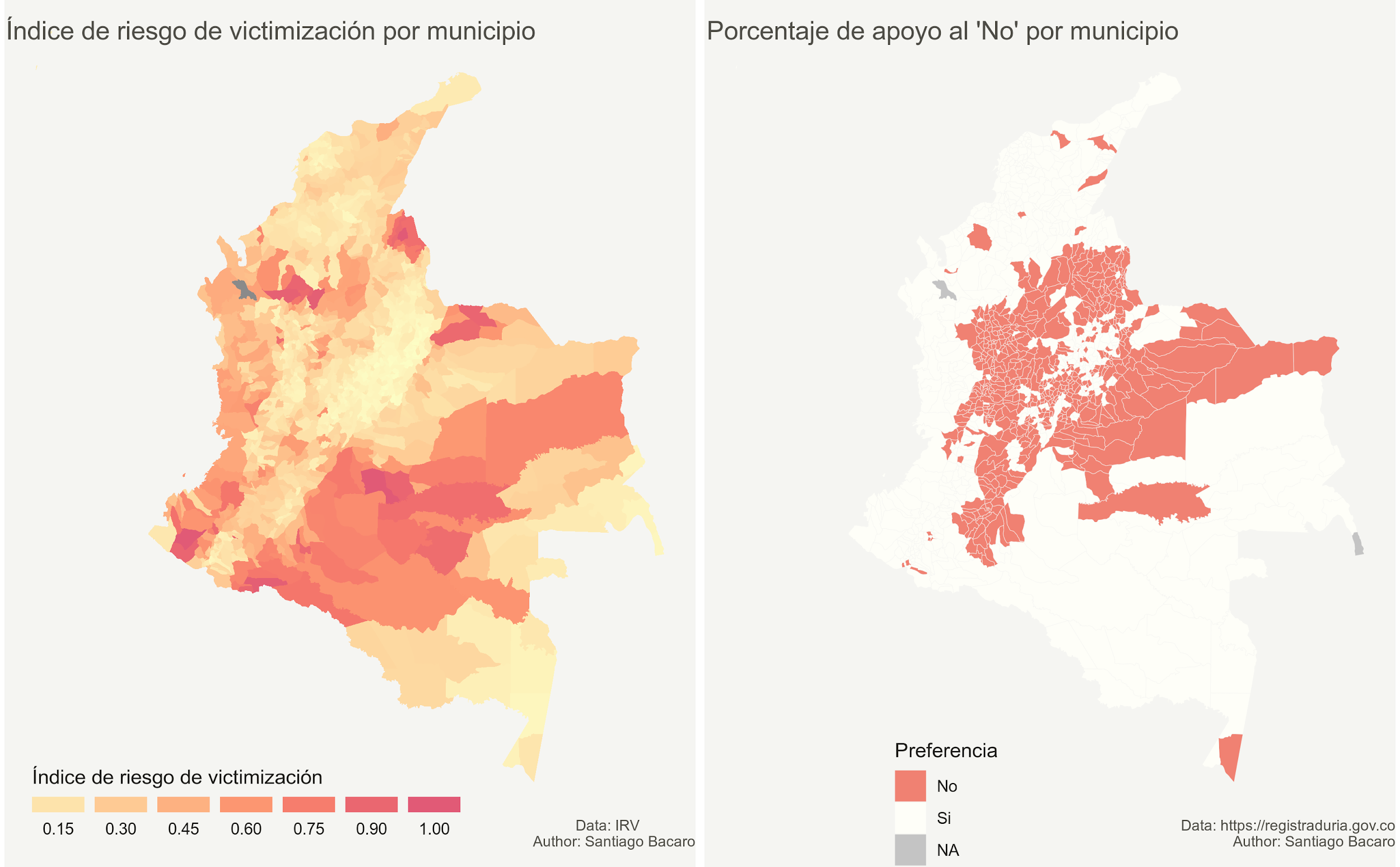
## Analysis and visualization of the results, comparing them with the IRV scores

The data for the IRV scores can be downloaded from <https://data.humdata.org/dataset/indice-de-riesgo-de-victimizacion-irv>. However, in the plebiscito\_mpios.csv file, it has already been included under the ‘IRV’ column.

You’ll find an R file in the root directory, named colombia\_irv\_map.R. That file contains R code to produce the visualization seen above, the data used by that file is all under the /data directory. You might need to move the script to that directory or update the path of the files.

These are the resulting visualizations from that file:





## Collection of Twitter data for the period on which the electoral campaigns took place

You’ll find a notebook in the root directory named retrieving\_the\_data.ipynb , as well as a script named twitter\_data\_extractor.py .

The Twitter data extractor file contains code for scraping data from Twitter, using Selenium and the Firefox webdriver. You’ll find instructions in the aforementioned Jupyter notebook to install those.

The Jupyter notebook also contains code that can be used to retrieve data from Twitter using Twitter’s API. However the data extractor is preferred, because of how limited the amount of tweets is in the free version of the API.

No wrapper Python packages were used to interact with the API, such as TweePy. This was because of the limitations it imposes. So, in the notebook you’ll find instructions for authenticating and fetching data.

## Preprocessing of the data

You’ll find a notebook named ‘generating\_and\_cleaning\_the\_dataset.ipynb’, which contains step by step instructions to perform the required preprocessing, whether you used the data extractor or Twitter’s API, and store the resulting Dataframe as a CSV file for further use.

## Analysis of the data

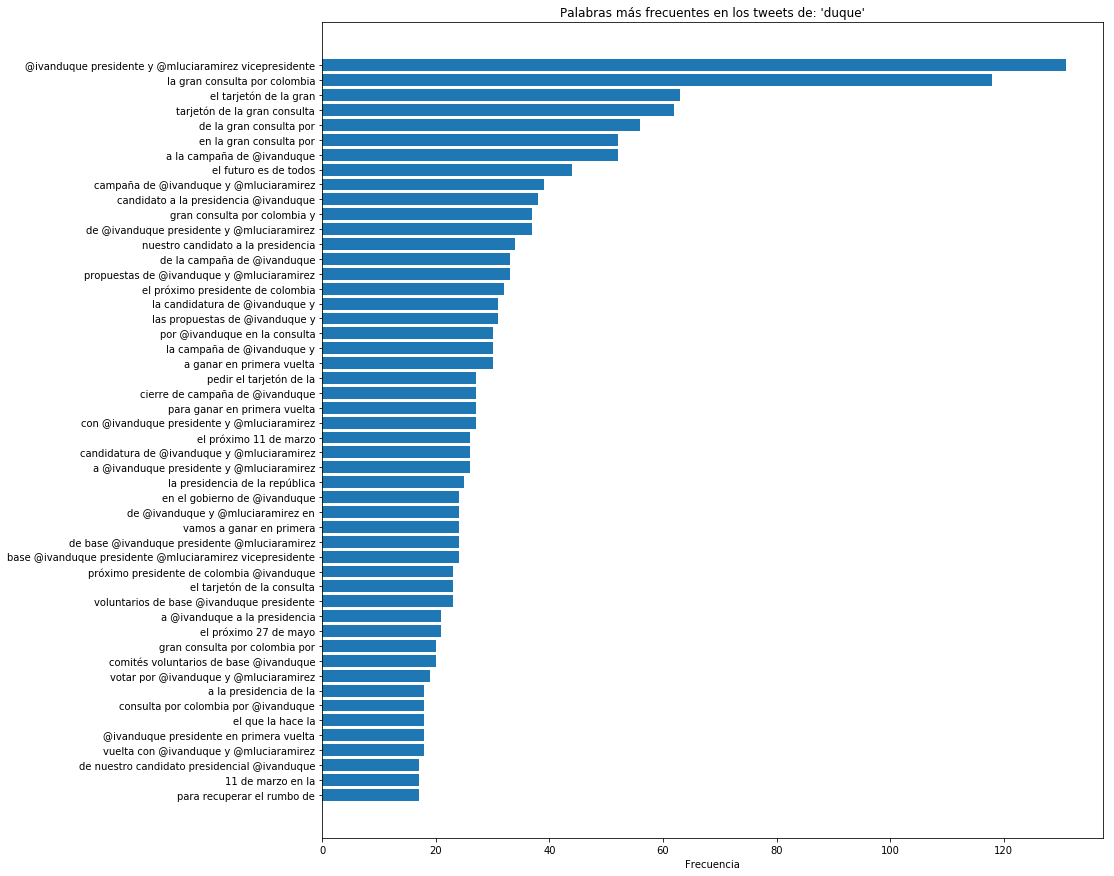
This section was carried out in two different Jupyter notebooks. The first one is named ‘analyzing\_the\_data.ipynb’, and the second one is named ‘pos\_tagging\_and\_sentiment\_analysis.ipynb’, and both are listed under the root directory.

On those notebooks you’ll find step by step instructions to load the data from the previously created CSV file, and perform the different analyses.

### Analyzing the data

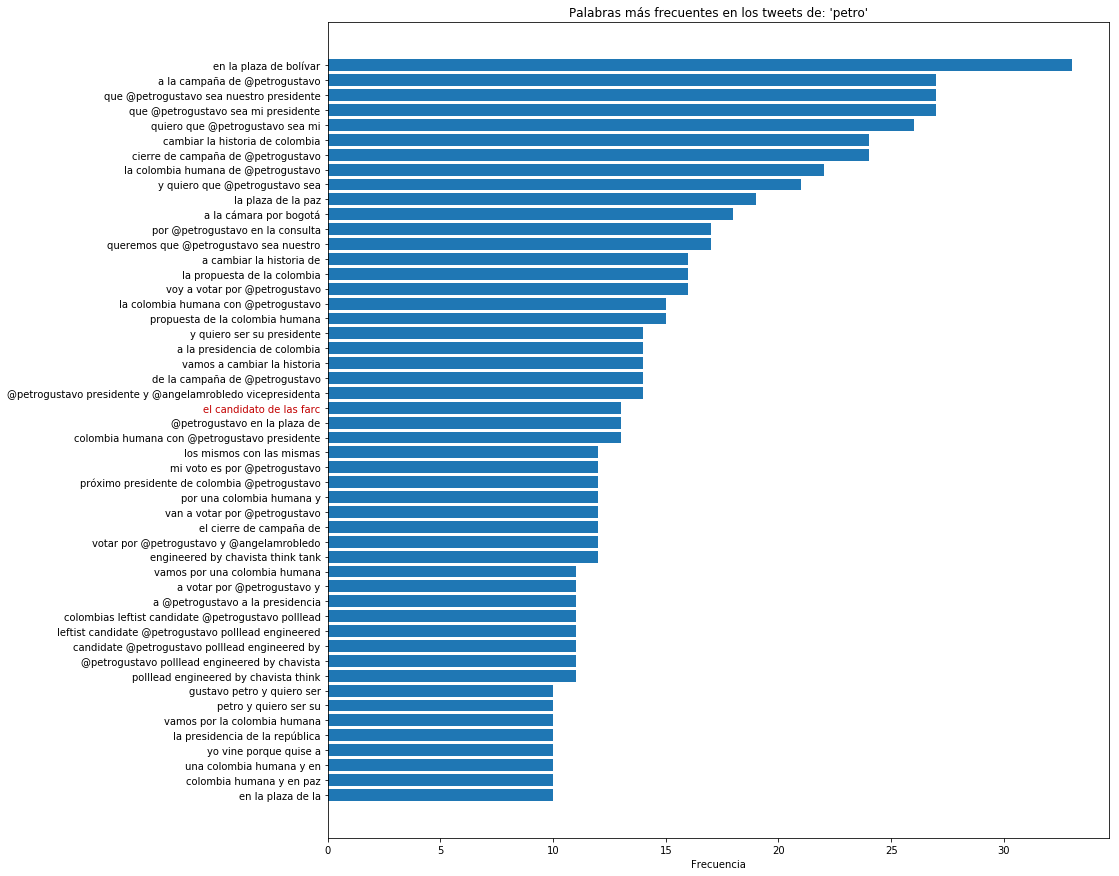
This notebook performs basic analysis of the numerical fields, an analysis of the length of the text of the tweets, and an analysis of most frequent n-grams, including unigrams, bigrams, trigrams and “pentagrams”.

Here are some of the results from that notebook:



This chart shows the most frequent “pentagrams” from the tweets mentioning Iván Duque.

The most frequent one found was “@IvanDuque presidente y @mluciaramirez vicepresidente”. This, as is seen in further detail in the notebook, shows that Iván Duque’s campaign was partly based in getting him to run for president, and not Marta Lucía Ramirez, his current vicepresident. It is not very clear why this might have been this way, but it can be speculated that it has to do with Duque having no political career, and thus, no direct political enemies. As opposed to Marta Lucía, who has a long political career, and therefore, might had not have as many chances to win.



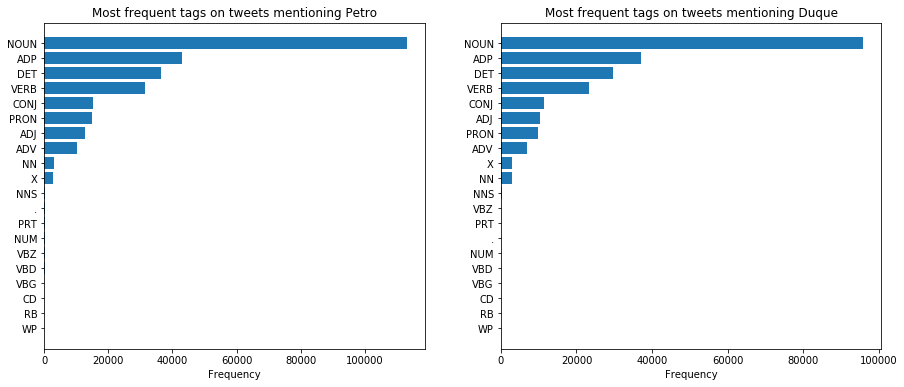
And these are the most frequent “pentagrams” mentioning Gustavo Petro. What stands out the most from these, is that one of the pentagrams was “El candidato de las FARC”. That’s part of the defamation that this project pretended to find. The followers of the Centro Democrático party said that Gustavo Petro was FARC’s candidate. Which is a plain lie, FARC, for the first time in history, were able to participate in the elections with their own candidate Rodrigo Londoño Echeverri. But many people did believe that Petro was their candidate.

You’ll find the analysis for the other n-grams in the notebook.

### POS tagging and sentiment analysis

The second notebook dealing with the analysis of the data, contains step by step instructions to perform POS tagging on the text of the tweets, as well as to create a sentiment analyzer.

Here are some of the results from that notebook:



These are the most frequent tags for the tweets mentioning each of the candidates. The only thing that is probably worth mentioning here, is that tweets mentioning Iván Duque, had more adjectives than those mentioning Gustavo Petro.

This can mean anything, it can be that the opposers criticized him more, or that his followers endorsed him more.

For the sentiment analyzer, a Support vector classifier was created. Trained with data from TASS, for which you can find links on the notebook.

However, the sentiment analyzer had a very poor performance, with an accuracy of only 67.01%. This was the main reason for which no further analyses were performed regarding the sentiment. But it is a great field for growth for this project.

## Limitations

The first main limitation found in this project was the retrieval of data from Twitter. The API, as aforementioned, had very low limits and Twitter blocked me after downloading more data from using the extractor.

The second main limitation was that there is not as much available data for training spanish sentiment analyzers as there is for training english ones. TASS provides great datasets, but because of the way the tweets are written, the initial accuracy was not as high.

1. https://repository.oim.org.co/bitstream/handle/20.500.11788/806/COL-OIM%200340.pdf?sequence=1&isAllowed=y [↑](#footnote-ref-0)