Big Data to study Polarization and Political Trust

Big Data team: Carlos Arcila Calderón (team leader), Félix Ortega Mohedano, David Blanco-Herrero (University of Salamanca) TRI-POL research project and José Manuel Robles (team leader, UCM), Rafael Caballero (UCM) and Juan Antonio Guevara (UCM) PRODEBAT research project.

The application of big data and computational techniques is one of the most relevant trends in the field of study of Social Sciences. That is why this project, like many others in the current scenario, complements the use of well-established research methods, such as surveys or experiments, with a big data approach. This offers the possibility of analyzing a larger amount of information, reaching longer periods of time, and reducing the effort of the research team by automating some of the tasks of collection and analysis.

This is why Tri-Pol (UPF / USAL), in collaboration with PRODEBAT (UCM), focuses on the use of Big Data strategies to directly collect and study contents from digital and social media in order to observe their effects on polarization, political trust and incivility.

With this, it is expected to specifically test some of the hypotheses of the research proposal:

H3: Direct exposure to illiberal and identity-conflictual messages produced by political elites increases the individual-level of affective polarization and political distrust;

H4: Anti-system populists and illiberal parties contain a greater number of illiberal and identity-conflictual messages;

H5: Exposure to social media of anti-illiberal and populist political parties increases affective polarization and political distrust;

H6: Exposure to like-minded outlet media and social media increases affective polarization and political distrust.

H7: *Incivility* is one of the most relevant transmission mechanisms of polarization. This communication strategy is more common among populist parties.

H8: The way in which the public debate is structured in social networks has an measurable effect on affective and ideological polarization.

H8.1: When the debate arises spontaneously, the probability of political polarization increases.

H8.2: When the debate is structured through some exogenous pattern, the probability of political polarization is reduced.

To test these hypotheses, both teams are in charge, in a coordinated manner, of the automatic collection and analysis of the content of social and digital media that will allow the projects to achieve the following objectives:

O2: Study the effects of political communication by political actors on political trust and affective polarization;

O3: Assess the effects of concrete measurable individual exposure to mass and social media and its diverse content on political trust and affective polarization;

O4: Operationalize the concept of incivility to analyze its use by political parties in online political debates;

O5: Analyze cases in which debates arise spontaneously and cases in which the debate is structured to measure polarization;

O6: Estimate the effects of discussions in social media on political trust and affective polarization.

Development of the tasks

A) Big Data Team activities

Six initial tasks were designed for the Big Data team. To this point, most of the preliminary and design work has been conducted in order to prepare the collection and analysis. During this first year, the most relevant activities of the Big Data team have been: agreeing on the specific sources and contents that will be collected and designing and preparing the collection processes.

First, together with the coordination of the project, it was decided that two types of contents were needed in each country that is present in the project: the messages posted in social media by political elites –and, specifically, the messages posted by the main political leaders and a selection of public institutions in Twitter– and contents from media outlets –particularly, editorial and opinion texts about national politics from the websites of the main printed and digital news outlets–. In total, 73 private and institutional accounts and 49 news outlets will be analyzed. Once these contents were determined, the most adequate way for their collection was designed. Two strategies have been established, one for the social media accounts and one for the news outlets.

Regarding the social media accounts, all the Twitter handles[1] of the selected accounts have been identified so that they can be used to conduct the downloads. Additionally, the script needed for the automatic download has been updated, as a new Twitter's API for researchers has been implemented recently; this API will allow the access to the complete archive of historic posts, making it possible for the team to reach all the tweets and retweets posted by the selected accounts during a period of one year[2]. This will offer large amounts of data, which allows a broader and more accurate study of the tweeting patterns of each account, therefore allowing more specific comparisons that with shorter periods of analysis would not be possible. Beside these large-scale downloads, a shorter download of the two weeks before and the two weeks after the fieldwork of the experimental surveys conducted within the project starts, testing the contents posted in those accounts at the specific time that this part of the study took place. At the same time of these downloads, the contents posted by the panelists participating in the experiment will be also downloaded; the reason to do this is that these panelists will be required to follow some of the previous Twitter accounts and this download will allow us to test whether this has any effects on the tweeting pattern of the panelists.

Similarly, the links of all the selected media outlets were identified, as well as the possibility that of each of them offers to access their archives. It has been observed that not all the media have an open archive, so in some cases the contents can be accessed for free, while for others a paying account is needed; at the same time, the design of each website is different, so the strategy to collect past texts differs from one another. Given this multiplicity of media and the limitations of scraping techniques for this activity, it was decided that a manual download of the contents would be conducted. For that, a research assistant needs to be hired for six months; this person, in coordination and under the supervision of the Big Data team, will design an individual strategy for each medium –including a subscription or a purchase for those outlets that do not allow their archives to be accessed freely– and will manually collect in .txt format all those editorial and opinion texts about national politics for a one-year period, as well as during the two weeks before and after the start of the fieldwork of the experimental surveys.

All this refers to the three tasks of the Big Data team that have been already started:

- **T1**. Automated collection of Twitter accounts (main politicians / invited panelists) during field experiments in the five countries under analysis: this task has been prepared –the scripts are ready– and it can begin once the starting date

[1] The Twitter usernames that appear at the end of the unique Twitter URL of each user; in other words, it is the username that appears after the @ sign and is unique for each Twitter account.

[2] Given the capacity of the automated collection and analysis of contents, this period could be increased without a significant increase in efforts and resources; that is why, if needed, longer periods of time could be studied, but the one-year period has been established for two reasons: first, because it is considered enough to establish the tweeting patterns of an account and, second, because it will be the length used in the analysis of the news outlets.

of the fieldwork of the experimental survey is confirmed and once the invited panelists have accepted to participate and have provided their handles –those of the politicians have already been located–.

- **T2**. Automated collection of Twitter accounts (main politicians) during the last one year: this can take place immediately, given that the scripts are ready and the accounts have been located. The ending of the one-year period will be placed two weeks and one day before the starting of the fieldwork, so that the contents collected in T1 are not repeated here.
- **T4**. Automated/Manual collection of archive outlet contents (editorial contents mainly) of selected outlets during the last one year and during field experiments: the collection process is being designed for each outlet and it can start once the assistant research staff is hired.

The rest of tasks will take place once the collection is completed. Given that the analysis will be automated in all cases, the most intense activity will be the adaptation of existing scripts in order to adequately conduct the analysis of the collected tweets or editorial and opinion pieces. For all these contents, we will apply lexicon-sentiment analysis with *SentiStrength*, supervised sentiment analysis with a pre-trained model using deep learning and embeddings, topic modeling with different parameters, and Social Network Analysis (SNA) to visualize the network structure of the different spreading possibilities of the analyzed contents –share, reply_to, quote, etc.–. More specifically, these tasks, which have not been yet started, are:

- **T3**. Automated content analysis of collected tweets (lexicon-sentiment analysis, supervised sentiment analysis, topic modeling, SNA). It refers to the contents collected in T1 and T2.
- **T5**. Automated content analysis of archive outlet contents (editorial contents mainly) of selected outlets during the last one year and during field experiments (lexicon-sentiment analysis, supervised sentiment analysis, topic modeling, SNA). It refers to the contents collected in T4.
- **T6**. Automated content analysis of the news selected for the experiment (lexicon-sentiment analysis, supervised sentiment analysis, topic modeling, SNA). It refers to news specifically selected by the designers of the experiment to be included in it, and that will be later analyzed using the same automated methods as the other media contents.

B) Joint activities TriPol - UCM

A total of five additional tasks are proposed, coordinated between the two research groups.

- **T7.** Case of studies: For this collaboration, four case studies will be selected. We propose four general elections: Spain, EEUU, Italy, Portugal, Chile, Argentina, and Colombia.
- **T8.** Data: The data for the elections in Spain and the United States have already been downloaded by the PRODEBAT team. This team has data from Twitter, Facebook and Instagram although, so far, only the Twitter data has been analyzed.

  We propose that the data for the elections in Italy, Portugal, Chile, Argentina and Colombia be downloaded by the PRODEBAT group following the same criteria that were used for the elections in Spain and the United States. This is, in the case of Twitter using keywords (then the accounts of the main leaders, media and political parties can be searched and identified). In the case of Twitter, the download would be done through the API of this social network. In the case of Facebook and Instagram through the contract signed by PRODEBAT with Facebook to access their data through *crowdtangle*.
- **T9.** Definition of objectives for each case study and possible comparative analyzes based on the hypotheses indicated in this document. The two research teams would be involved in this task.
- **T10.** Operationalization of key concepts (polarization, incivility, trust, etc.) and data analysis.
- **T11.** Analysis and preparation of documents for publication. The two research teams would be involved in this task.