European Commission

Social media data to indicate active citizenship at local level: the case study of attendance to demonstrations

by Cristina Rosales,

Introduction

Data availability is a persistent constraint in the generation of indicators for decision-making processes. Web 2.0 technologies offer new potentials of data sources that need to be investigated. To contribute to that field of research, we analyze the potential of social media to obtain indicators of active citizenship at a local level.

Active citizenship is a relevant indicator of social capital, classically measured based on survey results, which are costly and time consuming. Social media data, instead, is generated in enormous quantities at any moment and, in some cases, is publicly accessible through APIs. Our research focuses on Twitter and Instagram data. Twitter data can be retrieved in real time by using the Twitter Streaming API, while access to Instagram data is restricted and only possible through third-party services.

To obtain indicators at a local level, the geolocation of the data is crucial. Due to privacy concerns, very few Twitter users reveal their location in their posts, while in Instagram, pictures are frequently published with exact location. These and other limitations are also analyzed in the present study, that focuses in attendance/support to demonstrations, one variable contained in the Active Citizenship Composite Indicator (ACCI). Here we present some early findings of our ongoing research that centers the case study of Spain.

Method

We develop a method that allows to measure the variable ¨attendance/support to demonstrations¨ by using data from the social networks Twitter and Instagram (Fig.1). The method includes the following steps:

1. **Data collection** of posts containing terms related to demonstrations. In such a way, a set of keywords is settled and data is collected during March 2017.
2. **Data cleaning**. Polysemic keywords induce to false positives that need to be removed from the dataset.
3. **Users profiling**. We remove posts from users related to media, collectivities, organizations, corporations or brands, and keep just posts generated by individual citizens. We establish a set of keywords that allows to distinguish between different users' typology.
4. **Posts geolocation.** When data from Twitter and Instagram does not include geographical information, we can make use of the spatial information included in the metadata attributes. We use the online geocoding services of OpenStreetMap to parse the textual words and phrases in the metadata attributes and assign to them geographic identifiers, i.e. latitude and longitude.

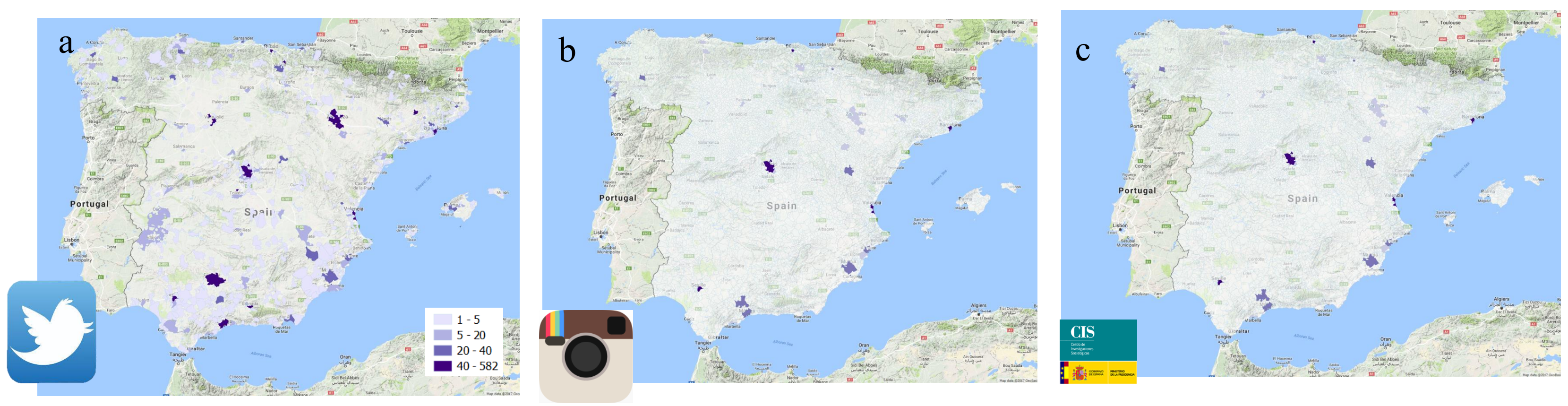

Figure 2. Number of Twitter users (a) and Instagram users (b) sharing information on demonstrations by city in Spain during the period of data collection. c) Percentage of citizens that participated in a demonstration in the last 12 months

Results

- 85% of the tweets collected are related to demonstrations. Instagram data shows lower results (75%) due to a high number of hashtags are commonly used and not always relevant to the topic, producing false positives.
- The method for profiling users offers good results for Twitter and Instagram data, with 91% and 89% accuracy, accordingly.
- Just the 0.4% of the collected tweets include coordinates. The rest of the tweets are located by geoparsing the tweeting location (present in 1% of the collected tweets) and the location established in the user profile (present in 62% of the tweets). The tweeting location are toponyms that follow standard formats, easing the geocoding process. While the location from the user profiles are open fields, thus the accuracy of the geoparsing method is not so high, 65%.

  The collected data from Instagram present coordinates in a 31% of the cases. Toponyms found in the pictures' captions allow to geolocate another third of the Instagram posts.
- The spatial granularity of the results is variable, but mainly to city level (Fig. 2)..
- High temporal granularity. Tweets are collected in real time, while Instagram data are obtained searching in the posts from the past.
- Results from the survey show the percentage of citizens that participated in a demonstration during the last 12 months. Survey results from cities with less than 100.000 inhabitants are anonymized to keep privacy. Thus, spatial , and also temporal, granularity are low.

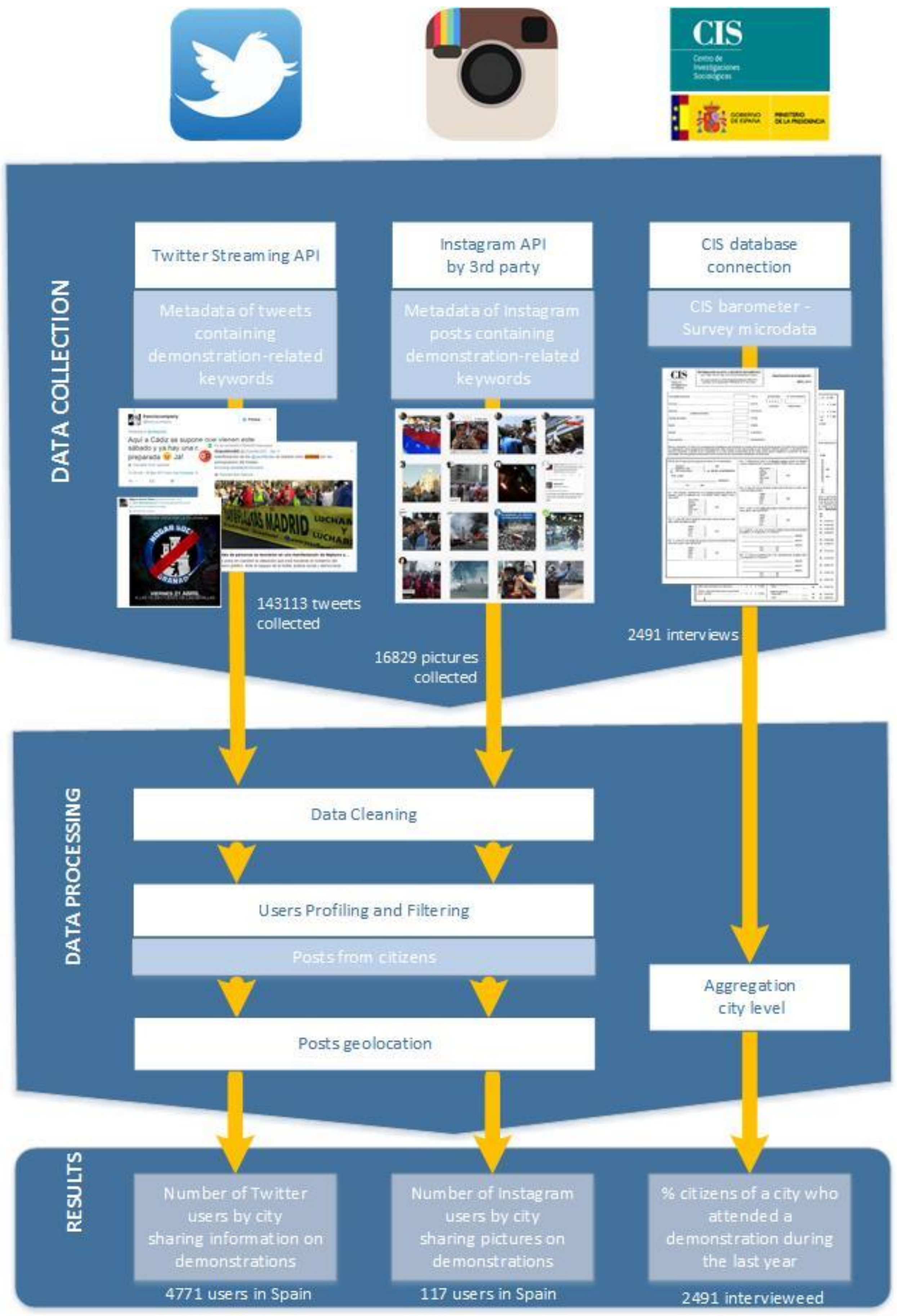
CIS
DATA COLLECTION
Twitter Streaming API
Metadata of tweets containing demonstration-related keywords
143113 tweets collected
Instagram API by 3rd party
Metadata of Instagram posts containing demonstration-related keywords
16829 pictures collected
CIS database connection
CIS barometer - Survey microdata
2491 interviews
DATA PROCESSING
Data Cleaning
Users Profiling and Filtering
Posts from citizens
Posts geolocation
Aggregation city level
RESULTS
Number of Twitter users by city sharing information on demonstrations
4771 users in Spain
Number of Instagram users by city sharing pictures on demonstrations
117 users in Spain
% citizens of a city who attended a demonstration during the last year
2491 intervieweed

Figure 1. Workflow to obtain indications of the variable "attendance/support to demonstrations" considering three different data sources: Twitter data, Instagram data and survey data

Discussion and Conclusion

Attendance/support to demonstrations is shown in Twitter and Instagram data. Based on those data, our method allows to obtain an indicator that is more timely and less costly than official surveys. Also, data from Twitter and Instagram offer results with lower spatial and temporal resolution. In such a way, data from Twitter and Instagram can be used for detailed temporal analysis of the citizens attending/supporting demonstrations, while surveys are useful for interannual analysis. Twitter and Instagram data are as well convenient for studies of small urban areas, where sample sizes of official surveys are not big enough to offer representative values of the population. However, results from the indicator based on Twitter and Instagram should be taken with caution, in the sense that the data is not representative of the whole population and that the method present limitations, and coments on social networks about demonstrations is not comparable to the answers from the survey, i.e. participating in a demonstration. Special attention needs to be taken to the keywords considered to obtain the data from Twitter and Instagram, specially in the later where the general use of many hashtags induce high number of false positives.

https://ec.europa.eu/jrc

Joint Research Centre

Cristina Rosales Sánchez
European Commission • JRC / Wageningen University
rosales.cris@gmail.com