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

Abstract

Data availability is a general 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. The present study contributes to that area of research analysing the potential of Twitter and Instagram data, as examples of social media, to obtain indicators of active citizenship at a local level. More precisely the study focuses on attendance/support to demonstrations, a variable included in the active citizenship composite indicator (ACCI). The data obtained from Twitter and Instagram is filtered to obtain relevant information regarding demonstrations. Then the geolocation of the information allows generating indicators at a municipal level. Geoparsing technics are used to obtain geographic coordinates out of the spatial information included in the metadata of the Twitter and Instagram data. The results from the social media data are compared with the results from survey data, to understand the strengths and weaknesses from each social network regarding the assessment of attendance/support to demonstrations.

Keywords: social media data, active citizenship, demonstration, Twitter, Instagram, survey data

1 Framework

Huge amounts of data are generated from citizens due to the new communication possibilities offered by Web2.0 technologies. Compared with data from official sources, these data generated from citizens present particularities (related to data accuracy, quality and lack of standards) that require the generation of different methods for their acquisition and analysis.

To contribute to this field of research, our study aims at analysing the potential of data from social networks to obtain indicators of quality of life at a municipal level. In particular, we focus on active citizenship, as one of the multiple quality of life dimensions (OECD 2013; European Statistical System Committee 2011).

The traditional way to evaluate active citizenship is based on surveys, which are costly and time consuming. The present study evaluates "attendance/support to demonstrations", variable considered in the active citizenship composite indicator (Hoskins et al., 2006), considering data from the online social networks Twitter and Instagram.

2 Objective

Our study aims at analysing the potential of Twitter and Instagram data to evaluate the attendance/support to demonstrations. With that aim, we evaluate the strengths and weaknesses of unstructured data from both social networks, in comparison with data obtained from official sources.

3 Data and methods

To focus the research we center in Spain as case study. The methodology includes the following steps:

- Data acquisition from the official source CIS. The barometer held on October 2016 (CIS 2016) ask if the interviewee had participated in the last 12 months in a demonstration.
- Data acquisition from Twitter during March 2017. Connecting to the Twitter streaming API, we extract all the tweets that contain a set of selected keywords related to demonstrations.
- Data acquisition from Instagram during March 2017. In this case leveraging a third-party platform for data access. Pictures whose captions contain any of the selected keywords are selected.
- Data cleaning. Polysemic keywords induce to false positives that need to be removed from the dataset.
- Users profiling. We remove posts from users related to media, collectivities, organizations, corporations or brands, and keep just posts generated by individual citizens. We check the presence of specific keywords in the users' "bio" to determine the user typology.
- Geolocation of the data from social networks. When data from Twitter or Instagram does not include geographical information, we 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.

4 Results and conclusions

The method to collect posts from Twitter and Instagram in relation to demonstrations offers 85% accurate results in Twitter posts, and 75% in the Instagram. Polysemic words used as keywords induce to false positives, lowering the accuracy of the method. Then, users are categorized and filtered with high accuracy, around a 90% for both cases, Twitter and Instagram.

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 is an open field in which unconventional writing style make it difficult to parse (Ajao et al. 2015), thus just a 65% is geolocated by the geoparsing method. In the case of the Instagram data, a 31% of the collected data presents coordinates. Toponyms found in the pictures' captions allow to geolocate another third of the Instagram posts. In general, the spatial granularity of the results from Twitter and Instagram is variable, but mainly to city level.

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 compared with the results from Instagram and Twitter.

5 Conclusions

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 is not completely accurate in all the parts of the process. Special attention needs to be taken to the keywords considered to obtain the data from Twitter and Instagram, especially in the later where the general use of many hashtags induces high number of false positives.

6 References

Ajao, O., Hong, J. & Liu, W., 2015. A survey of location inference techniques on Twitter. *Journal of Information Science*, pp.1–11.

CIS - Centro de Investigaciones Sociológicas, 2016. *Barómetro 3156 - Octubre 2016*, Available at: http://www.cis.es/cis/opencm/ES/1_encuestas/estudios/ver.jsp?estudio=14311&cuestionario=17172&muestra=23822.

European Statistical System Committee, 2011. *Sponsorship Group on Measuring Progress , Well-being and Sustainable Development - Final Report*, Available at: http://epp.eurostat.ec.europa.eu/portal/page/portal/pgp_ess/about_ess/measuring_progress.

Hoskins, B. et al., 2006. *Measuring Active Citizenship in Europe*, Luxembourg. Available at: http://bookshop.europa.eu/en/measuring-active-citizenship-in-europe-pbLBNA22530/.

OECD, 2013. *Guidelines on measuring subjective well-being*, Available at: http://www.oecd.org/statistics/guidelines-on-measuring-subjective-well-being.htm [Accessed February 27, 2014].