The Argentine economy on Twitter

J. Daniel Aromí[1,2] and Sergio Andrés De Raco[1]

[1] Universidad de Buenos Aires. Facultad de Ciencias Económicas. Buenos Aires, Argentina. Universidad de Buenos Aires. Instituto Interdisciplinario de Economía Política de Buenos Aires. Buenos Aires, Argentina
[2] Pontificia Universidad Católica Argentina

Abstract. We propose and implement a methodology for data collection and analysis of Twitter discussions linked to the Argentine economy. Starting with a list of "seed users" later expanded based on following-follower relationships, we build a network of interactions and fetch their tweet timelines. Then, we use a community detection model to compress the structure of underlying relationships and a standard topic model to represent the latent issues discussed in each community. Results suggest that this strategy is able to learn a useful organization and to summarize the contents of social media exchanges of the Argentine economic tweetosphere. Potential applications could be to characterize the links between different economic sectors and to construct community-level indicators of opinions.

Keywords: Economics · Social network analysis · Community detection · Topic modeling.

1 Introduction

Social media data provide information about human activity and reaction to events, usually grouping opinions in actors communities. The analysis of communities on social media has proved useful in studies of diverse subjects such as public debate [1], recommendation dynamics [2] and marketing [3]. This data can also prove useful in the study of economic modeling and prediction [4]. With this objective, in this work we propose and implement a methodology to identify an Argentine economic tweetosphere.

2 Methodology and Data

We acquire data from selected sites on the internet and Twitter API calls to build a database of curated user accounts and a corpus of tweets. Next, we use this relational data to build a network and analyze community structure between the users involved. Finally, we use topic modeling techniques to look for latent structure in the corpus of tweets harvested.

Firstly, we define a set, D, of internet pages with directories of potentially relevant macroeconomic actors like chambers of commerce and its corporate

members [3], labor unions [4], and public institutions and regulators [5]. The selection criteria was deliberately wide, with international replication in mind. Then we scrap these sites and fetch a list of available homepages locators, L, which we further scrap to look for their institutional Twitter accounts. The resulting list of users, S, is then stored and enriched with user's metadata using Twitter API. In our experiment, S comprises 245 Twitter accounts. Next, beginning with each seed user s friends' list, $\overrightarrow{F_s} \in \overrightarrow{F}$, we grow the network in n successive generations adding in each step G_i users to the original seeds, $U = S + \sum_i^n G_i$, using metadata for each unique account according to a "representativeness" criteria. We evaluate relevant candidates $f \in \overrightarrow{F}$ to meet two joint conditions: 1. (absolute) f has to be at least in the top-k rank of relevant users' friends, $\overrightarrow{F}$; and, 2. (relative) at least an α_i share of f followers ($\overleftarrow{f}$) has to belong to the set of reference users, U, in the i-th generation. For this experiment we considered $n = 2$ growth generations, with the set of parameters $\{k = 5000, \alpha_1 = 0.2\%, \alpha_2 = 0.4\%\}$. The final network consisted of 3,745 accounts with 654,197 undirected links between them, which presents dense interaction for this kind of empirical networks, usually very sparse. This way of growing the network induces connectivity properties that can add to its largest (weakly) connected component, which we use to analyze the community structure. Secondly, to build a corpus of texts, T, we use Twitter API to get the tweets timeline for each user, $u \in U$, and compose one document per user, $T_u \in T$, consisting of her last 2,000 tweets. The corpus built consists of approximately 7M tweets grouped in 3,745 documents representing the accounts of reference.

Network community structure refers to a certain meso-structure that relates nodes of a complex network of entities in groups where the members of each subset of nodes connect more between them than with members of another subset. Depending on the specific type of network and research question of interest there are many algorithms for community detection [5], that can be broadly classified according to the type of node subsets they generate regarding the existence or not of intersections between them (v.g.: partitioning or overlapping). It has been well documented that social networks, like Twitter, tend to present overlapping community structure reflecting multiple "circles of interests" of actors (nodes) involved. We tested different algorithms and choose DANMF [6], which uses non-negative matrix factorization techniques on the (undirected) network's adjacency matrix to learn a low dimensional representation of the node embeddings with neural networks, because the communities detected were diverse and interesting in the sense of economic interpretation.

A well known problem in the field of natural language processing refers to the identification and extraction of abstract themes or topics in the content of text corpora. Topic modeling is a type of statistical model for discovering these

[3] Chambers of Commerce at *Argentina* portal, https://www.argentina.gob.ar/trabajo/camarasempresarias, and corporate partners at *Union Industrial Argentina*, https://uia.org.ar/socios/.
[4] Labor unions at *Sindicatos Argentinos*, http://www.sindicatosargentina.com.ar/.
[5] Public institutions at *Mapa del Estado*, https://mapadelestado.jefatura.gob.ar/organismos.php.

abstract themes for which there are a variety of sound techniques available. In this work we use Latent Dirichlet Allocation (LDA, [7]).

3 Results and Discussion

The community detection algorithm was trained for the case of eight communities. The detected communities were labelled after inspecting the set of more representative accounts, that is, accounts with a large value of the embedding dimension corresponding to that community.

As shown in Table 1, the members of the detected communities can be characterized in terms of their role, relevant economic sector and, in one case, geography. The roles associated to different detected communities are: politician, business leader, journalist, and union leader (see communities 1, 4, 5 and 7). On the other hand, the relevant economic sectors that are linked to specific communities are: agriculture, health, and transport (see communities 3, 6, 7 and 8). Finally, members of community 2 are linked to Córdoba, the second largest province in Argentina.

Table 1. Detected communities

Community	n	Selected representative users
1. Politics/Think-tanks	543	cabrerafran, andreshibarra, braunmi, fedesalvai, deAndreis
2. Córdoba	456	sol_aguirre, jmlucero, guadalupealt, fpiccato, ClusterCba
3. Agro-technology	1043	JCMolinaHafford, AgrofyNews, MinAgriCba, intaargentina
4. Business chambers	545	InfoCamarasArg, produccion_arg, RedADIMRA
5. Journalists/economists	829	gustavolcordoba, RadioCutFm, AgenciaTelam
6. Science/health/univ.	543	CONICETDialoga, INNGENIAR, PNUDArgentina
7. Transport/unions	308	NPortuarias, GlobalportsAr, FeMPINRA, SonidoGremial
8. Agriculture	626	carlosetchepare, diazdecampo, SociedadRural, bertellof

The community detection model also provides information regarding the centrality of certain users and the overlap between certain communities. Regarding centrality, only 3 accounts belonged to 5 or more communities: YPFoficial, AgenciaTelam y produccion_arg. Among significant overlaps, 228 accounts belonged to community 3 (Tecnol. Agro.) and 8 (Agriculture), 169 accounts to communities 1 (Polit./Think-tanks) and 6 (Science, health and universities), and 98 accounts belonged to communities 5 (Journalists/Economists) and 7 (Transport/Unions).

We trained a LDA model with 16 topics to characterize the issues discussed in the communities, after exploring specifications with up to 32 topics. We labeled the topics after inspecting the representative words for each topic, that is, the words that are particularly more likely in the respective topic. The trained model reflects the diversity of issues discussed.

Table 2 shows the most likely topics in each detected community. The associations provide information that validates the labels assigned to the communities and, at the same time, allow for a richer characterization. Additionally, it is worth noting the existence of two "general interest" topics (Entrepreneurship/innovation and Politics) that are among the most frequent topics in 4 out of

8 communities. In contrast there exist topics that are among the most frequent in only one community: Unions, Macroeconomics and Sustainability/Science.

Table 2. Most frequent topics by community

Community	Most frequent topics
1. Politics/Think-tanks	Polítics, Entrepr./Innov.
2. Córdoba	Entrepr./Innov., Polítics, Macroeconomics
3. Agro-technology	INTA/Agro.tech., Agriculture
4. Business Chambers	Politics, Entrepr./innov., University, Production/trade
5. Journalists/economists	Polítics, Macroeconomics
6. Science, health and univ.	Entrepr./innov., Sustain./Science, University
7. Transport and unions	Unions, Production/Trade
8. Agriculture	Agriculture, INTA/Agro-tech

Discussion This work presents and implements a method for the analysis of social media discussion related to the Argentine economy. Preliminary results are promising and could be followed by extended work to analyze robustness to changes in parameters and use this information to construct indicators of economic sentiment and performance. We consider that these novel forms of data and approaches to extract relevant information from it can contribute significantly to the understanding of complex economic phenomena.

References

1. Aruguete, N., *et al.*: Time to #protest: Selective exposure, cascading activation, and framing in social media. Journal of communication **68**: 480-502 (2018)
2. Giordano, A., *et al.*: Detección y refuerzo de comunidades de celíacos en Twitter Argentina. In: IV Simposio Argentino de GRANdes DAtos, JAIIO 47 (2018)
3. Lim, K., *et al.*: Finding twitter communities with common interests using following links of celebrities. In: Proc. of the 3rd Intl. Workshop on Modeling social media (2012).
4. Buono, D., *et al.*: Big data types for macroeconomic nowcasting. EURONA **1**: 93-145 (2017)
5. Fortunato, S.: Community detection in graphs. Phys. Rep., **486**: 75-174 (2010)
6. Ye, F., *et al.*: Deep autoencoder-like NMF for community detection. In: Proc. of the 27th ACM ICIKM, pp. 1393-1402 (2018)
7. Blei, D., *et al.*: Latent Dirichlet Allocation. JMLR, **3**: 993-1022 (2003)