Data Mining and Social Network Analysis on Twitter

Jesus Silva, Noel Varela, David Ovallos-Gazabon, Hugo Hernández Palma, Ana Cazallo-Antunez, Osman Redondo Bilbao, Nataly Orellano Llinás and Omar Bonerge Pineda Lezama

Abstract  The emergence of a networked social structure in the last decade of twenty-first century is accelerated by the evolution of information technologies and, in particular, the Internet has given rise to the full emergence of what has been called the Information Age [1] or the Information Society [2]. Social media is yet another example of people’s extraordinary ability to generate, disseminate and exchange meanings in collective interpersonal communication with a massive, real-time networked system where everything tends to be connected. The analysis of the climate of opinion on Twitter is presented around the Common Core State Standards (CCSS), one of the most ambitious educational reforms of the last 50 years in USA.
Keywords Social media mining · Social media · Twitter · Social network analysis · SNA · Common core state standards

1 Introduction

Twitter is a microblogging social network that is widely used today. Collecting and protecting data has been a constant in order to recover, analyze and generate information and knowledge. Social media present the same need with a significant difference [3, 4]: the speed at which new data is produced, which the sociologist Paul Virilio anticipated when stated that “real time prevails over real space and the geosphere [5]. The supremacy of real time and immediacy, over space and surface is a fait accompli and has an inaugural value (heralds a new era)” [6], so that “the emergence of hyper-immediate social media highlights the need for new forms of real-time research” [7].

Social Media Mining (SMM) [8] can be defined as the process of extracting, storing, representing, visualizing and analyzing user-generated data with the aim of discovering significant patterns (dissemination of information or rumors, influence, homophilia, social or consumer behavior, prediction, etc.) from social interactions in Internet social media. To deepen the understanding, real and potential, both of the patterns of interactions and of the identification of those nodes with a disproportionate influence or superhubs within the studied network and their effects have become an increasingly significant issue for more fields of knowledge [9].

The objective of the research is to propose the identification of superhubs in large networks extracted from Twitter in relation to their most prominent explicit relationships.

2 Data

For presenting a new method linking social network analysis (SNA), SMM and Twitter, it was necessary to choose an object of study that fulfilled two conditions: (1) to be as significant as possible in its social impact, without necessarily having a global scale and (2) the climate of opinion must keep active over time [10, 11].

The fieldwork consisted of tracking, capturing, storing, representing and analyzing the information generated on Twitter about US President Donald Trump.

3 Method

The capture and extraction of the data presented in this research were carried out directly from the application programming interface (API) of Twitter during six
The resulting Complete Network is an addressed network of 85,324 nodes and 154,258 relationships that were filtered to obtain the subsequent networks.

- The giant component (GC) with 54,235 nodes and 125,365 relationships gives an idea of the significant relational density within the climate of opinion.
- Application of the modularity metrics to the GC (Fig. 2) which shows the existence, represented in colors, of three large structural communities.
- The 1% network by in-degree: 754 nodes; 4520 relationships (Fig. 3). Applying the key player procedure to this network the study obtains that 93.5% of the network
Fig. 2  Modularity of the giant component

would be reached with only 18 superhubs nodes (3.59%) which is a very small fraction to achieve site percolation [14].

- Network of 0.25% per in-degree: 251 nodes; 1023 relationships (0.7% of the total). Starting from the hypothesis that vaccinating/infecting a small fraction of nodes—in this case, superhubs identified by third parties as influential—and their relationships in a network, site percolation can have very significant effects on it when it comes to slowing down/expanding information, ideas or perceptions and provoking a knock-on or chain effect in the network in which the benefit of slowing down/expanding this reduced number of nodes will provoke significant changes not only in the structure of the whole network but also in the behavior of the individuals that form it. Applying the key player procedure to this network the results show that 96.3% of the network would be reached with only 9 superhubs nodes to get the site percolation.

- Network of 1% per out-degree: 789 nodes; 4250 relationships. Applying the key player procedure to this network, the results show that 95.3% of the network would be reached with only 19 superhubs nodes, which is a very small fraction to get the site percolation.

- Network of 0.25% per out-degree: 250 nodes; 920 relationships. Starting from the same hypothesis that in the network of 0.32% by in-degree and applying the
procedure of the key player to this network, the results show that 94.2% of the network would be reached with only 5 superhubs nodes to get the site percolation.

- **Geolocation by in-degree.** With those individuals who make public the information of their geolocation in the USA in their profile, a representation by their in-degree was performed (Fig. 4), which shows the activity of the two large structural communities identified throughout the USA and a certain balance in the distribution of influence of viral information and the debate about Donald Trump.

- **Geolocation by out-degree.** With those individuals who make public the information of their geolocation in the USA in their profile, the representation was obtained by their out-degree, which shows the main focuses of emission, with very few nodes exercising in practice as main broadcasters. It also reflects the geographical origin of the activity of the two large structural communities on the two coasts and in the center of the country on the debate about Donald Trump.
5 Conclusions

The intersection between collective interpersonal communication on Twitter, SMM techniques and ARS presents four key characteristics for this study:

1. A structural intuition of social relations is assumed, 2. relational empirical data are systematically captured, collected, represented and analyzed, 3. mathematical models are used for analysis along with technology and 4. visualizations of relations and patterns of interaction are created and shared, allowing the generation of meaningful structural ideas and their communication to others, which fully coincides with [14] on the development of ARS as a social discipline.

The objective to study networks in this research is the identification of the main actors or superhubs in the social, cultural and political debate about Donald Trump on Twitter, from his most significant relationships, using those data generated by the user’s communication. Twitter offers, as a singular feature, the fact that it is playing, in practice, the function of intersecting medium of the rest of the media. This is a sort of spinal column or central nervous system through which the contents of the collective interpersonal communication, facilitated by the Internet architecture, can be identified, captured, analyzed and represented.

There is evidence that shows clear connections between the contagion/diffusion/development of infectious diseases and the diffusion of information, since both are propagated from person to person through networks, of influence or homophilia, which shows a great structural similarity [15]. This feature has led to
the diffusion of ideas being conceptualized as social contagion [16] and is applicable to Internet social media and, notably, to Twitter.

This type of data presents new opportunities and challenges to researchers, where the most interesting resides, not only in the amount of data, but also in what they can do with these large amounts of data that cannot be done with small amounts. At least two major challenges for researchers already exist:

1. The challenge of complexity or how to capture and add multidimensional data in a consistent way, not very homogeneous, not very structured and massive that are produced endlessly at any time or place and that have heterogeneous and unstable sources (that can appear and disappear) keeping the search and identification of significant patterns as a central objective.

2. The challenge of *N* = everything or how to develop methodologies that allow to work with the totality of the data that is produced, that is, how to investigate with complete universes. If sampling is a technique developed for times of information scarcity, it is a reality that was collectively abandoned already. However, the era of Big Data or the Petabyte Era predicts “a world in which vast amounts of data and applied mathematics replace any other instrument” implying that “the volume of data will obviate the need for theory, and even scientific method” [17]. In the meantime, researchers must keep the focus on analysis processes and correct decision making by identifying significant patterns and being aware that “the implicit promise of Big Data is that the solution to information overload passes through greater amounts of data” [18].

An obvious limitation of this type of research would lie in the lack of access to the meanings that circulate through the networks. To this end, there are emerging methodologies such as netnography [19], a proposal between sociology and anthropology that responds to the need for interdisciplinary approaches to unite the study of structure and access to meanings.

References

1. Hahsler, M., Karpienko, R.: Visualizing association rules in hierarchical groups. J. Bus. Econ. 87, 317–335 (2017)
2. Yuan, M., Ouyang, Y., Xiong, Z., Sheng, H.: Sentiment classification of web review using association rules. In: Ozok, A.A., Zaphiris P. (eds.) Online Communities and Social Computing. OCSC 2013. Lecture Notes in Computer Science, vol. 8029. Springer, Berlin, Heidelberg (2013)
3. Silverstein, C., Brin, S., Motwani, R., Ullman, J.: Scalable techniques for mining causal structures. Data Mining Knowl. Discov. 4(2–3), 163–192 (2000)
4. Amelec, V., Carmen, V.: Relationship between variables of performance social and financial of microfinance institutions. Adv. Sci. Lett. 21(6), 1931–1934 (2015)
5. Java, A., Finin, T., Song, X., Tseng, B.: Why we Twitter: understanding microblogging usage and communities. In: WebKDD/SNA-KDD’07 Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis (2007)
6. Naaman, M., Boase, J., Lai, C.-H.: Is it really about me? Message content in social awareness streams. In: CSCW’10 Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (2010)
7. Zafarani, R., Ali Abbasi, M., Liu, H.: Social Media Mining. Cambridge University Press. http://dx.doi.org/10.1017/cbo9781139088510 (2014)

8. Morris, M.R., Teevan, J., Panovich, K.: What do people ask their social networks, and why? A survey study of status message Q&A behavior. In: CHI'10 Proceedings of the 28th International Conference on Human Factors in Computing Systems (2010)

9. Viloria, A., Lezama, O.B.P.: Improvements for determining the number of clusters in k-means for innovation databases in SMEs. Proc. Comput. Sci. 151, 1201–1206 (2019)

10. Kamatkar, S.J., Kamble, A., Viloria, A., Hernández-Fernandez, L., Cali, E.G.: Database performance tuning and query optimization. In: International Conference on Data Mining and Big Data, pp. 3–11. Springer, Cham (2018)

11. Mayer-Schönberger, V., Cukier, K.: Big data: a revolution that will transform how we live, work, and think. Mariner Books. http://dx.doi.org/10.2501/ija-33-1-181-183 (2014)

12. Caglierio, L., Fiori, A.: Analyzing Twitter user behaviors and topic trends by exploiting dynamic rules. In: Behavior Computing: Modeling, Analysis, Mining and Decision, pp. 267–287. Springer (2012)

13. Erlundsson, F., Brodka, P., Borg, A., Johnson, H.: Finding influential users in social media using association rule learning. Entropy 18, 164 (2016)

14. Meduru, M., Mahimkar, A., Subramanian, K., Padiya, P.Y., Gunigur, P.N.: Opinion mining using twitter feeds for political analysis. Int. J. Comput. (IJC) 25(1), 116–123 (2017)

15. Abascal-Mena, R., López-Ornelas, E., Zepeda-Hernández, J.S.: User generated content: an analysis of user behavior by mining political tweets. In: Ozok, A.A., Zaphiris, P. (eds.) Online Communities and Social Computing. OCSC 2013. Lecture Notes in Computer Science, vol. 8029. Springer, Berlin, Heidelberg

16. Dehkharhgan, R., Mercan, H., Javeed, A., Saygin, Y.: Sentimental causal rule discovery from Twitter. Expert Syst. Appl. 41(10), 4950–4958 (2014)

17. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363–370 (2005)

18. Viloria, A., et al.: Integration of data mining techniques to PostgreSQL database manager system. Proc. Comput. Sci. 155, 575–580 (2019)

19. Guandong, X., Lin, L., (eds.). (2013). Social media mining and social network analysis; emerging research. Reference and Research Book News 28(3). Portland: Ringgold Inc. http://dx.doi.org/10.4018/978-1-4666-2806-9