Challenges in Social Media Analytics: Are We Studying Bots or Humans?

Anatoliy Gruzd, PhD
Ted Rogers School of Management, Ryerson University
350 Victoria Street, Toronto, ON, Canada, M5B 2K3
gruzd@ryerson.ca

Social media is a rich source of behavioural data that can reveal how and why we interact with each other online and what that might mean for our society as we continue to speed towards an increasingly computing-mediated future. Computational techniques such as social media analytics have been extremely effective in turning user-generated data into useful insights for benefits of individual users in applications such as recommender systems [1], [2] and businesses in applications such as customer relation management [3]. However, due to the rise of social bots that are becoming increasingly more sophisticated [4] and the rise of algorithmic filtering which guides online users to make certain choices and view certain content [5], it begs the question whether we are modeling human behavior in social media or simply reverse engineering how bots and other algorithms operate.

It is true that some previous research has already highlighted the influence of user interfaces and other affordances of social networking websites on how people interact online. For example, due to the public nature of Twitter and its users' inability to control who has access to view what content, Twitter users are more likely to post and share positive messages [6]. This is in part to avoid being associated with negative news. Hogan [7] refers to such behavior as following the 'lowest common denominated', suggesting that people post the content that they believe is the most acceptable to their perceived online audience. Hogan observed a similar behavior among Facebook users.

While it is crucial to understand how user interfaces and other affordances, as shown above, influence how people interact with one another and what content they share; the rise of aforementioned bots and algorithmic filtering may have an even bigger impact on how users behave online and how researchers interpret online data. This is because computational techniques used to analyze social media data is often blind to biases or noise in data unless we specifically model it. There are a lot of emerging work on bot and spam detection [8]–[10], but the challenge is how do we get these techniques to masses of researchers who are actively relying on social media data to study different demographic groups and their behavior online and offline. The most susceptible to this growing issue are social scientists and others who often rely on ready-to-use applications to mine social media data, but who might not have background or resources to develop customs scripts and run modelling to limit the influence of bot-like accounts and take into account the use of algorithmic filtering by a studied platform. Just think how many times, have you seen a section on how social media data was cleaned from bots and other outliers in social media literature, not counting the papers that specifically focused on this topic? Probably not a lot.

This issue is complicated by social media platforms likely being reluctant to pursuing removal of suspicious accounts aggressively as it may affect their user growth rate (often equated to a platform's worth by investors). At the same time, social media platforms like Facebook and Twitter are wrapping up their efforts in experimenting with algorithmic filtering in attempt to combat information overload and customize user experience by showing users only content that the 'algorithm' thinks relevant to them [11], [12]. Putting aside the discussion about whether social media companies should or should not be doing this, the main point for this piece is that they are doing it and that it affects what people see and do in social media. And this has a direct impact on studies that rely on social media data, such as studies related to information diffusion modeling in social media. The issue is less salient if a study is examining how information flows in a particular system, but it would be highly problematic if it relies on data to model human behavior.

The increasing number of bots in our datasets and the increasing use of algorithmic filtering by social media giants is essentially widening the gap between online and offline, and between computer-mediated and algorithm-driven communication. This in turn makes the online data less reliable, at least for those of us studying human behavior. Therefore, there is an urgent need to better understand the nature of bots and algorithmic filtering, and their influence on users' online interactions, not just from a computational, but also from sociological perspective.

I want to conclude this position paper by calling social media researchers in computation fields to develop and share strong principles, protocols, tools and techniques around handling and cleaning social media data. We also need to develop stronger partnerships across social media-related fields (and especially with social science researchers) to start discussing how to properly handle bot-like accounts and the influence of algorithmic filtering once detected. For example, should such accounts be removed from the datasets or kept and treated like any other agents in our models? We may consider as a safe practice (from a research perspective) to remove a group of marketing-related bot accounts that are part of an activist online group, if these bots do not interact with anyone else in the group but just there to increase their following base, like those observed in [13]. While at the same time, we might want to keep automated Twitter accounts designed to repost certain news stories as they may play an important information propagation role by transmitting bits across different online communities. The answer would vary and likely depend on many factors such as the study focus, the nature of bots and their impact or lack of on online participants. But getting to the 'answer' would also require more empirical and social science-driven work to be introduced in the computational modelling arena and vice versa; thus, my call for closer partnership among qualitative and quantitative social media researchers.

REFERENCES

[1] W. Feng and J. Wang, "Retweet or not?: personalized tweet re-ranking," 2013, p. 577.

[2] N. Nizam, C. Watters, and A. Gruzd, "Link Sharing on Twitter during Popular Events: Implications for Social Navigation on Websites," 2014, pp. 1745–1754.

[3] M. Oliveira, A. Guerreiro, and J. Gama, "Dynamic communities in evolving customer networks: an analysis using landmark and sliding windows," *Soc. Netw. Anal. Min.*, vol. 4, no. 1, Dec. 2014.

[4] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The Rise of Social Bots," *ArXiv14075225 Phys.*, Jul. 2014.

[5] D. Lazer, "The rise of the social algorithm," *Science*, vol. 348, no. 6239, pp. 1090–1091, Jun. 2015.

[6] A. Gruzd, "Emotions in the Twitterverse and Implications for User Interface Design," *AIS Trans. Hum.-Comput. Interact.*, vol. 5, no. 1, pp. 42–56, Mar. 2013.

[7] B. Hogan, "The Presentation of Self in the Age of Social Media: Distinguishing Performances and Exhibitions Online," *Bull. Sci. Technol. Soc.*, vol. 30, no. 6, pp. 377–386, Nov. 2010.

[8] A. H. Wang, "Don't follow me: Spam detection in Twitter," in *Proceedings of the 2010 International Conference on Security and Cryptography (SECRYPT)*, 2010, pp. 1–10.

[9] M. Fazeen, R. Dantu, and P. Guturu, "Identification of leaders, lurkers, associates and spammers in a social network: context-dependent and context-independent approaches," *Soc. Netw. Anal. Min.*, vol. 1, no. 3, pp. 241–254, Jul. 2011.

[10] S. Yardi, D. Romero, G. Schoenebeck, and D. Boyd, "Detecting spam in a Twitter network," *First Monday*, vol. 15, no. 1, Dec. 2009.

[11] K. Kokalitcheva, "Twitter's New Algorithmic Filtering Is Here and Optional," *Fortune*, 10-Feb-2016.

[12] V. Luckerson, "Here's How Facebook's News Feed Actually Works," *Time*, 09-Jul-2015.

[13] A. Gruzd and K. Tsyganova, "Information Wars and Online Activism During the 2013/2014 Crisis in Ukraine: Examining the Social Structures of Pro- and Anti-Maidan Groups," *Policy Internet*, vol. 7, no. 2, pp. 121–158, Jun. 2015