Letter to the Editor

The Research Topic Defines "Noise" in Social Media Data – a Response from the Authors

Yoonsang Kim[1], PhD; Jidong Huang[2], PhD; Sherry Emery[1], PhD

[1]NORC at the University of Chicago, Health Media Collaboratory, Chicago, IL, United States

[2]Georgia State University, Health Management and Policy, Atlanta, Georgia

Corresponding Author:
Yoonsang Kim, PhD
NORC at the University of Chicago
Health Media Collaboratory
55 E Monroe St, 30th Floor
Chicago, IL, 60603
United States
Phone: 1 312 357 3878
Fax: 1 312 759 4000
Email: kim-yoonsang@norc.org

Related Articles:
Comment on: http://jmir.org/2016/8/e219/
Comment on: http://jmir.org/2016/2/e41/

(J Med Internet Res 2017;19(6):e165) doi: 10.2196/jmir.6824

KEYWORDS

automated tweets, noise, social media data

We provide a response to Allem and Ferrara [1], who recently commented on our article, "Garbage in Garbage Out: Data Collection, Quality Assessment and Reporting Standards for Social Media Data Use in Health Research, Infodemiology and Digital Disease Detection," which was published in JMIR in February 2016 [2]. In their comment, published in JMIR in August 2016, entitled "The importance of Debiasing Social Media Data to Better Understand E-Cigarette-Related Attitudes and Behaviors," Allem and Ferrara discuss the importance of removing bias in social media data. They claim that automated tweets are noise that injects bias into the data, and thus should be removed before applying the framework we proposed [1]. We believe they misunderstood our intent. In addition, their discussion misinterprets the key messages of our article; the implication of their comments, which suggests that automated tweets are garbage, is highly misleading. A formal response is provided here to articulate accurately the main focus of our article and present a different view about the "noise" in social media data.

The objective of our paper was "to develop and apply a framework of social media data collection and quality assessment, and to propose a reporting standard," as stated in the abstract. The e-cigarette-related tweet data were used as "a real-world example" to demonstrate how to apply this framework to develop a search filter, and how to estimate the measures of data quality under different conditions. The objective of our paper was not to understand e-cigarette-related attitudes and behaviors expressed on Twitter.

The definition of the "noise" in social media data by Allem and Ferrara, as any tweets produced from an account identified as a social bot, is narrow and oversimplifying, and may even be misleading in some cases. Organic and commercial tweets are not isolated in the Twittersphere. Many organic tweets are retweets or replies to commercial tweets, of which a large number is generated by bots. Whether automated contents generated by bots should be considered as noise depends on the research topic at hand. Although it may be important to remove bot tweets and focus solely on organic contents for certain research topics, it is equally important to measure the amount of these bot tweets and the content of (mis)information in these tweets for many other research topics [3]. For example, a study that examines the commercial advertising on e-cigarette should include the tweets generated by bots. The automated social media messages are not unique to the topic of e-cigarettes. For many other research topics, including other tobacco products, pharmaceutical products, dietary supplements, etc., automatically-generated marketing content is common. In fact, one of the studies that Allem and Ferrara cited to justify removing automated tweets discussed the value of "understanding the effect of promotionally marketing vaporization products" on social media using "cyborgs to mimic organic users" because of their importance to public health and policy [4]. This underscores the importance of being able to

identify and quantify such automated messages in order to understand their impact on the marketplace and individual attitudes, beliefs and behaviors.

Allem and Ferrara also briefly discussed the inherent bias in social media data due to the fact that social media users are not a representative sample of the general population. However, this itself does not limit the value of social media data, and it can be used as an advantage to study hard-to-reach populations such as young adults, and ethnic, racial, and sexual minorities. Social media can serve as a good alternative or complementary data source to understand behavior and intentions among these understudied and hard-to-reach groups.

Removing automated contents and applying other approaches to remove noise can be considered in the stage of developing search filters if it is deemed appropriate for the research topics in study. However, it is not a necessary component to be considered for all research using social media data. This point underscores the main thesis of our paper: that clear disclosure about data cleaning and processing (e.g. whether bot tweets are included or not) is important.

Acknowledgments

This study was supported by the FDA Center for Tobacco Products under Award Number P50CA179546. The content is solely the responsibility of the authors and does not necessarily represent the official views of the FDA.

Conflicts of Interest

None declared.

References

1. Allem JP, Ferrara E. The importance of debiasing social media data to better understand e-cigarette-related attitudes and behaviors. J Med Internet Res 2016 Aug 09;18(8):e219 [FREE Full text] [doi: 10.2196/jmir.6185] [Medline: 27507563]
2. Kim Y, Huang J, Emery S. Garbage in, garbage out: data collection, quality assessment and reporting standards for social media data use in health research, infodemiology and digital disease detection. J Med Internet Res 2016 Feb 26;18(2):e41 [FREE Full text] [doi: 10.2196/jmir.4738] [Medline: 26920122]
3. Gruzd A. Who are We Modelling: Bots or Humans? In: Proceedings of the 25th International Conference Companion on World Wide Web. Geneva: International World Wide Web Conferences Steering Committee; 2016:551.
4. Clark EM, Jones CA, Williams JR, Kurti AN, Norotsky MC, Danforth CM, et al. Vaporous marketing: uncovering pervasive electronic cigarette advertisements on Twitter. PLoS One 2016 Jul;11(7):e0157304 [FREE Full text] [doi: 10.1371/journal.pone.0157304] [Medline: 27410031]

Edited by G Eysenbach; this is a non–peer-reviewed article. Submitted 11.11.16; accepted 08.05.17; published 02.06.17.

Please cite as:
Kim Y, Huang J, Emery S
The Research Topic Defines "Noise" in Social Media Data – a Response from the Authors
J Med Internet Res 2017;19(6):e165
URL: http://www.jmir.org/2017/6/e165/
doi: 10.2196/jmir.6824
PMID: 28576756