Selected Papers of AoIR 2016:
The 17th Annual Conference of the
Association of Internet Researchers
Berlin, Germany / 5-8 October 2016

#MISSINGDATA: A METHODOLOGICAL INQUIRY OF THE HASHTAG TO COLLECT DATA FROM TWITTER

Evelien D'heer, Pieter Verdegem & Frederik De Grove

iMinds – MICT – Ghent University

Introduction

In this paper, we assess the Twitter hashtag from a methodological perspective. Compared to conventional methods in social science research (such as surveys and in-depth interviews), social media as a method is far from being fully understood.

We evaluate a popular sampling procedure for Twitter studies, i.e. the hashtag approach (Ausserhofer & Maireder, 2013; Bruns & Burgess, 2011; Iannelli & Giglietto, 2015). Hashtags are valuable to study the emergence and evolution of Twitter debates around particular topics. However, scholars do acknowledge not all relevant tweets are captured (Bruns & Moe, 2014; Larsson & Moe, 2012). In particular, "*we may significantly underestimate the full volume of @replies which was prompted by hashtagged tweets*" (Bruns & Stieglitz, 2014, p. 75). Is that so? What are we missing? And does it matter?

Research aims

This paper empirically examines the impact of hashtag sampling on conversation networks. Conversation networks are user-user networks, based on users that address other users via the @reply function. Our dataset includes hastagged tweets, hashtagged replies *and* non-hashtagged replies. The data allows us (1) to compare the characteristics of hashtagged and non-hashtagged responses, (2) to assess the changes in the network structure and (3) to assess the relative positions of the users in the network.

Data collection and analysis

The empirical work of this study is based on hashtag data related to the 2014 elections in Belgium. In this respect, the study is explorative, looking for a number of tendencies that are worth investigating for other events and contexts.

Data collection is based on the combination of hashtag and user streams to capture follow-up responses that do not contain the hashtag. This procedure was followed for a

D'heer, E., Verdegem, P. & De Grove, F. (2016, October 5-8). *#MissingData: A methodological inquiry of the hashtag to collect data from Twitter*. Paper presented at AoIR 2016: The 17th Annual Conference of the Association of Internet Researchers. Berlin, Germany: AoIR. Retrieved from http://spir.aoir.org.

selected period of time during the pre-election campaign (early May 2014). First, all tweets containing the dedicated hashtag (i.e. #vk14 or #vk2014) are captured via the public stream. Second, for each harvested tweet the original sender is tracked, capturing all tweets from and to this specific Twitter user. Following, the combination of the hashtag and user streams allows us to reconstruct full conversations containing both hashtagged tweets and non-hashtagged responses. In total, our sub-sample consists of 1719 tweets from 868 unique users which reflects about 10% of the pre-election debate.

We used a logistic regression to account for the differences between hashtagged and non-hashtagged responses. The independent variables are: (1) the number of included hyperlinks, (2) the number of additional hashtags, (3) the number of included @mentions and (4) the message word count.

Via Social Network analysis (SNA) we compare (1) the "hashtag only" conversation network and (2) the conversation network *including* non-hashtagged responses. We analyzed structural network characteristics and users' altering positions in the network. With respect to the users' identity, we distinguish between elites (i.e. journalists and politicians) and non-elites (i.e. citizens). The analyses were conducted in UCINET (Borgatti, Everett & Freeman, 2002).

Results

The difference between hashtagged and non-hashtagged responses

The logistic regression showed structural differences between replies that include or exclude the dedicated hashtag (i.e. #vk14/2014). In particular, the following elements show a significant positive relation with the inclusion of the dedicated hashtag: (1) additional hashtags***, (2) the presence of hyperlinks*** and (3) tweet length***.
In other words, the inclusion of the dedicated hashtag (i.e. #vk14/#vk2014) co-occurs with in inclusion of interactive, informational elements (e.g. hashtags/hyperlinks). Further, the significance of tweet length also indicates the usage of the dedicated hashtag is linked to the information value of the Twitter message.

The changes in the conversation network when including non-hashtagged responses

The "hashtag only" network contains 161 users and 156 relations, whereas the network including non-hashtagged responses contains 518 users and 1082 ties (see Figure 1 below).

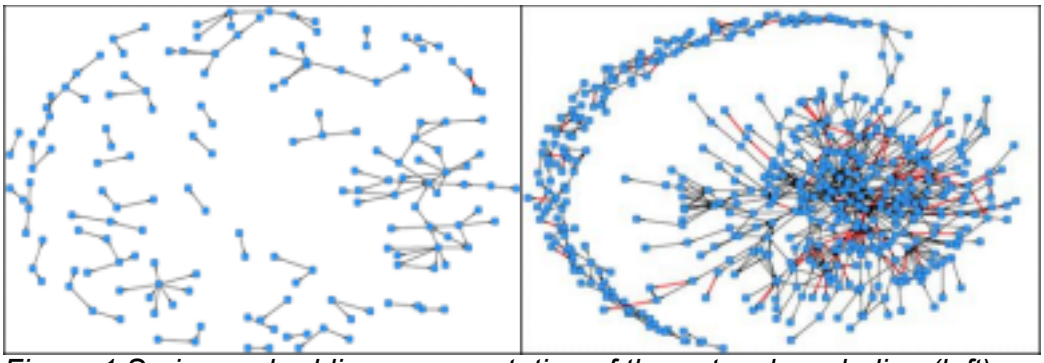

Figure 1 Spring-embedding representation of the network excluding (left) and including (right) non-hashtagged responses

Network reciprocity increased from .79% to 14.69%. Reciprocity operates at the "micro level" (Bruns & Moe, 2014); as it measures to what extent reciprocal relations between two users occur. Reciprocal ties are represented by the thicker red lines in Figure 1. Further, the "main component" or largest component of connected nodes increased in size from 37 nodes to 359 nodes. Hence, more users are added and more connections between users occur in the network. However, these connections occur amongst a relatively limited set of users, rather than across the entire network.

Since the network grew in size and new users are included, existing users' positions altered. For about half of the users, we found the relative amount of messages they send or receive, increased when we include non-hashtagged responses. Taking a closer look at the identity of the users, we found differences between elites (i.e. politicians and journalists) and non-elites (i.e. citizens). It is predominantly the former that strengthen their position. The inclusion of non-hashtagged responses further confirms the insights we receive from "hashtag only" studies on the political debate on Twitter, i.e. the popularity and dominance of elites in the network.

References

Ausserhofer, J., & Maireder, A. (2013). National politics on Twitter: Structures and topics of a networked public sphere. Information, Communication & Society, 16(3), 291–314.

Borgatti, S.P., Everett, M.G. & Freeman, L.C. 2002. Ucinet 6 for Windows: Software for Social Network Analysis. Harvard, MA: Analytic Technologies.

Bruns, A., & Burgess, J. (2011). #Ausvotes: How twitter covered the 2010 Australian federal election. Communication, Politics & Culture, 44(2), 37–56.

Bruns, A., & Moe, H. (2014). Structural layers of communication on Twitter. In K. Weller, A. Bruns, J. Burgess, M. Mahrt, & C. Puschmann (Eds.), Twitter and society (pp. 15–28). New York: Peter Lang.

Bruns, A., & Stieglitz, S. (2014). Metrics for understanding communication on Twitter. In K. Weller, A. Bruns, J. Burgess, M. Mahrt, & C. Puschmann (Eds.), Twitter and society (pp. 69–82). New York: Peter Lang.

Iannelli, L., & Giglietto, F. (2015). Hybrid spaces of politics: the 2013 general elections in Italy, between talk shows and Twitter. Information, Communication & Society, 18(9), 1006–1021.

Larsson, A. O., & Moe, H. (2012). Studying political microblogging: Twitter users in the 2010 Swedish election campaign. New Media & Society, 14(5), 729–747.