Creating Social-Science Grounded Algorithms to Analyze Communication Dynamics in Big Data

Jennifer Stromer-Galley, Tomek Strzalkowski, George Aaron Broadwell, Samira Shaikh, Ting Liu, Sarah Taylor, Xiaoai Ren, Feifei Zhang, Jennifer Crowley

The growing trend in "big data" calls for new social science approaches to analyzing communication dynamics. Systematic content analysis by humans has been a key method to examine such dynamics in the past, but the bigness of social media data presents challenges requiring new methodological approaches. In this paper, we describe an effort to develop a computational algorithm that can automatically identify influence in online discussion environments (synchronous and asynchronous). Building on small group decision-making scholarship (Ellis & Fisher, 1994) as well as research by Huffaker (2010), Bracewell et al (2011), Freedman et al (2011) and others that examine leadership and influence in computer-mediated communication, we identify a variety of discourse features of influence. Then, groups of students are recruited to participate in an online chat (similar to Internet Relay Chat) for an hour (a total of 23 groups), and are given a survey before and after the discussion. The survey includes questions that ask the students to evaluate who was influential in the discussion. Then a sub-sample of the collected chats were content analyzed by humans with good levels of intercoder agreement on coding categories measuring a variety of dialogue acts (Krippendorff's alpha > .7). This coding was further used to train the algorithm. Preliminary results of the algorithm suggest a 63% agreement with participant evaluations on identifying the most influential in online discussion, and 97% agreement between the algorithm and the participants' evaluation of influential members of the discussion of at least one of the top two influential members of the discussion. Preliminary application of the algorithm to examine influence in WikiTalks has resulted in 85% accuracy of identifying the top two influential members of the discussion. Results of the project highlight the value of using social science methods to develop tools to tackle examination of communication processes in big data.

D. B. Bracewell, M. Tomlinson, Y. Shi, J. Bensley, and M. Draper. (2011). Who's playing well with others: Determining collegiality in text. In *Proceedings of the 5th IEEE International Conference on Semantic Computing (ICSC 2011)*.

Ellis, D. G., & Fisher, B. A. (1994). *Small group decision making: Communication and the group process*. New York: McGraw-Hill

Freedman, M., Baron, A., Punyakanok, V. & Weischedel, R. (2011). "Language Use: What it can tell us?" In *Proceedings of the Association of Computational Linguistics*. Portland, Oregon.

Huffaker, D. (2010). Dimensions of leadership and social influence in online communities. *Human Communication Research, 36*, 596-617.