Exploring the content of misinformation from multiple perspectives

Abstract for SBP 2017 Data Challenge

Dian Hu, David A. Broniatowski

Department of Engineering Management & Systems Engineering, The George Washington University, Washington, DC, 20052, USA
{hudian,broniatowski} @email.gwu.edu

1 Introduction

In recent years, rumors, unverified news and other types of misinformation have gradually become a common phenomenon online (Del Vicario et al., 2016). Misinformation has effects on political campaigns (Allcott & Gentzkow, 2017), public health communication (Broniatowski, Hilyard & Dredze, 2016), modern markets and other areas (DiFonzo & Bordia, 2007). In this data challenge project, we are interested in how the internal content of misinformation would affect online sharing behavior. In particular, we want to understand which internal factors of an article, if any, are predictive of how often the article is shared on Facebook.

2 Data Explored

As per the requirement of the data challenge, we will analyze news data from the GDELT database gathered using the keyword "vaccine". We will also explore news data from a major anti-vaccine news website from which we have previously gathered data.

3 Method, Designs and Variables:

3.1 Main Method and Dependent variables:

Our approach is to use regression analysis to explore which of the following factors, as independent variables, are predictive of the number of Facebook shares of articles. In this project, we gather Facebook share statistics using the Graph API provided by Facebook. The details of the API can be found at (https://developers.facebook.com/docs/graph-api).

3.2 Controlled variables:

Prior studies have found several factors are significantly associated with information cascades (Szabo & Huberman, 2010; Hofman, Sharma, & Watts 2017). These factors

include, but are not limited to: (1) the popularity of the article's topic at the time of its sharing; (2) the article's length; (3) the popularity of the news platform. To control these known significant factors respectively, (1) we select news articles about a fixed topic: vaccination; (2) we only consider news published on and after Jan 1st, 2016, and (3) we analyze data from each news site separately.

3.3 Independent variables

(a) **Five Measurement using Moral Foundations Dictionary**

The Moral Foundations Dictionary is a dictionary of linguistic tokens that are derived from the study of Moral Foundations Theory (Graham, Haidt & Nosek, 2009; Graham et al., 2012). The dictionary highlights five moral dimensions: Authority, Fairness, Harm, Loyalty, and Purity. In this project, we implement the dictionary by calculating the normalized frequency of the tokens in each category used in each article. Specifically, we measure each of the 5 moral foundation dimensions: $M_1 \ldots M_5$, each representing the weight of each dimension using this method.

(b) Readability Score, Article Length, and LSA Overlap Score

Many different metrics have been proposed in the field of psycholinguistics to characterize the various genres and patterns used in written communications. In this project, we will focus on three metrics that have been shown to be related to an article's causal coherence:

An article's readability score is calculated by assessing the reading difficulty of a written passage. The most commonly used formula is the Flesch Reading Ease Score (Flesch, 1947). We will also measure an article's length: the number of tokens in the article. Although the Flesch Reading Ease Score includes the article's length as a measurement, readers are may be less likely to complete a longer article, thus making it less likely that they will comprehend its central gist. Therefore, we examine the article's length separate from its readability

Finally, we will examine a metric based on Latent Semantic Analysis (Deerwester et al., 1990) – a tool that provides measures of semantic overlap between sentences. It's often used to measure the internal connections of a written passage. LSA provides a measure of semantic coherence (e.g., Foltz, Kintsch, & Landauer, 1998) that is frequently used as a proxy for causal coherence. In this project, we would use the online version of the Coh-Metrix suite (Graesser et al., 2004) to generate measurement of these variables.

(c) New Measurement of causal situation models

The measurements discussed in prior sections do not directly measure causal situation models that have been shown to underlie recall of text (e.g., van den Broek, 2010). We will compare the performance of these measures to graph-theoretic metrics derived from situation models gathered using the crowdsourcing tool presented in a companion paper in this year's SBP-BRIMS conference (Hu & Broniatowski, 2017), including network density, average degree, and betweenness of these graph-based models.

4 References:

1. Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. National Bureau of Economic Research.
2. Broniatowski, D. A., Hilyard, K. M., & Dredze, M. (2016). Effective vaccine communication during the disneyland measles outbreak. Vaccine, 34(28), 3225–3228.
3. Clark, H. H., & Clark, E. V. (1980). Psychology and language. An introduction to psycholinguistics.
4. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391.
5. Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., … Quattrociocchi, W. (2016). The spreading of misinformation online. Proceedings of the National Academy of Sciences, 113(3), 554–559.
6. DiFonzo, N., & Bordia, P. (2007). Rumor psychology: Social and organizational approaches. American Psychological Association.
7. Flesch, R. (1948). A new readability yardstick. Journal of Applied Psychology, 32(3), 221.
8. Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, *25*(2–3), 285–307.
9. Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. Behavior Research Methods, 36(2), 193–202.
10. Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2012). Moral foundations theory: The pragmatic validity of moral pluralism.
11. Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. Journal of Personality and Social Psychology, 96(5), 1029.
12. Hofman, J. M., Sharma, A., & Watts, D. J. (2017). Prediction and explanation in social systems. Science, 355(6324), 486–488.
13. Hu, D., Broniatowski, D. A. (2017). Measuring Perceived Causal Relationships Between Narrative Events with a Crowdsourcing Application on Mturk. Washington, DC, US.
14. Szabo, G., & Huberman, B. A. (2010). Predicting the popularity of online content. Communications of the ACM, 53(8), 80–88.
15. Van den Broek, P. (2010). Using texts in science education: Cognitive processes and knowledge representation. Science, 328(5977), 453–456.