Project Proposal

Pedro A. Arroyo, Daejin Kim, Arun Pandian, Di Tong

10/17/2019

Introduction

With the move of political discourse to digital platforms, the United States is increasingly coming into contact with computational propaganda - i.e. the use of algorithms and automation to disseminate false information or to manipulate user activity on social media (Samuel C. Wooley and Philip N. Howard 2017). Researchers, politicians, and policy experts have recognized the emergence of these strategies as a threat to democratic institutions. Concern has grown over the corrosive effects of strategic competitors using distorted, misleading, or fabricated information to affect political behavior. For example, the 2016 election saw pro-Trump campaigns leverage computational techniques to generate pro-Trump twitter accounts. By election day, the number of pro-Trump twitter accounts outnumbered pro-Clinton accounts by a factor of 5 to 1 (Bence Kollanyi, Philip N. Howard, and Samuel C. Wooley 2016). These techniques threaten the democratic process on two fronts: first, they insert false information into the political consciousness; second, they undermine the franchise of individual voters.

One response to this threat has come from the world of think tanks in the form of working papers and reports attempting to understand the challenge as well as frame and outline a counter-strategy. Think tank documents capture only a narrow slice of elite discourse, but they are nonetheless important in so far as they reflect and shape the conversation happening among decision makers.

In order to gain a better understanding of how think tanks approach and react to computational propaganda, we propose to analyze published texts (e.g. blogs, reports, articles) through the use of Latent Dirichlet Allocation (LDA) topic modeling (David M. Blei, Andrew Y. Ng, and Michael I. Jordan 2003). LDA works by generating a probabilistic distribution of words for each topic and then examining how closely documents are related in terms of the topic probabilities (David M. Blei, Andrew Y. Ng, and Michael I. Jordan 2003). LDA then uses clustering techniques to build sets of related documents. Using this method, we hope to discover how experts from think tanks conceptualize and respond to issues around computational propaganda.

Data Collection

The corpus for this analysis will come from English-language texts put out by think tanks dealing with computational propaganda (often referred to as *digital disinformation*, or simply *disinformation*), specifically those texts that are geared towards policy makers in the United States. There are mainly two subgroups of texts to be collected. The first deals with disinformation campaigns targeting the United States as well as texts that speak about disinformation *in general*. Because we are interested in the American response to disinformation campaigns, we will ignore texts that deal with disinformation campaigns in Brazil, India, or Iran. We will also collect the relevant literature coming out of military organizations (eg. NATO STRATCOM) and schools of government (eg. Harvard Kennedy School) and foreign policy journals (eg. Foreign Policy). We will further sample literature from blogs, when the blogs belong to one of the think tanks working on disinformation (i.e. when they show up elsewhere in the dataset), or when the blog post is directly linked from another text in our dataset.

Some think thanks, which we know to be working on online disinformation, clearly classify their reports with tags, and so we can include all reports under specific tags like *Online Disinformation*. With other sources, tags are broader (eg. *Disinformation*, which may refer to traditional offline disinformation), in which case judgement of inclusion has to be made based on metadata (title, date published, author names, and table of contents). In yet other cases, when think tanks/institutions do not tag their reports, we will examine the content of the report to determine relevancy.

In short, we expect to take an expansive approach when building our corpus and will allow discontinuities within the discursive space to emerge from the subsequent analysis.

Method

Our analysis will rely on Latent Dirichlet Allocation (LDA) topic modeling, an unsupervised learning method that has the potential to either confirm existing theories or discover unknown categories and patterns not immediately apparent to human readers (David M. Blei, Andrew Y. Ng, and Michael I. Jordan 2003; James A. Evans and Pedro Aceves 2016; Laura K. Nelson et al. 2018, 2018). Assuming that each document in a corpus contains a mixture of topics, some of which occur throughout the entire corpus, LDA is able to uncover hidden thematic structures both in the corpus as well as in individual documents. Specifically, it can identify (i) the topics occurring in a given corpus, defined by a group of words with different weights; (ii) the distribution of topics in the corpus and in each document of the corpus.

Our approach is largely exploratory, specially so at the outset. To begin, we will use LDA to establish the

shared dominant topics for all texts in our corpus and to get a general sense of what the major concerns are in regards to disinformation and related campaigns in the United States. We anticipate supplementing this with a word-frequency analysis.

Looking Forward

There is a necessary tension between (a) taking an exploratory approach to a body of text in an effort to discover latent structure, and (b) identifying a corpus of text precisely because, as researchers, we suspect that it might contain such a latent structure. When researchers approach a project through a hypothesis-testing lens, open inquiry is aided through the prior-specification of research and analysis protocol; exploratory methods, in contrast, largely foreclose that approach. As a result, exploratory methods place an extra onus on researchers to be open-minded and avoid premature judgements. In that spirit, and without taking a position on whether such structures will emerge, we are comfortable specifying some dimensions we anticipate might emerge as structuring elements in the discourse. To wit: *Does the thematic composition of the texts differ by type of author? Are there strong regional effects or cross-regional differences? To the extent that there are distinct topic clusters, are they interrelated or isolated from each other?* Understanding the answers to these questions will better help us understand the American response to a rapidly expanding strategic challenge.

REFERENCES

Bence Kollanyi, Philip N. Howard, and Samuel C. Wooley. 2016. "Bots and Automation over Twitter During the U.S. Election." Working Paper Series. Computation Propaganda Project.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993–1022.

James A. Evans, and Pedro Aceves. 2016. "Machine Translation: Mining Text for Social Theory." *Annual Review of Sociology* 42 (July): 21–50. https://doi.org/https://doi.org/10.1146/annurev-soc-081715-074206.

Laura K. Nelson, Derek Burk, Marcel Knudsen, and Leslie McCall. 2018. "The Future of Coding: A Comparison of Hand-Coding and Three Types of Computer-Assisted Text Analysis Methods." *Sociological Methods & Research*, May. https://doi.org/https://doi.org/10.1177%2F0049124118769114.

Samuel C. Wooley, and Philip N. Howard. 2017. "Computational Propaganda Worldwide: Executive

Summary." Working Paper No. 2017.11. Computation Propaganda Project.