Twitter Generated Information Cascades

Kerria Pang-Naylor

Introduction

The bounded-confidence information dissemination model was developed to fit with social media message spread. Treating users as scalar-valued opinion nodes on a graph, this model simulates the spread of messages through a bounded-confidence mechanism. This mechanism imitates our bias towards similar opinions with a 'confidence-bound' c that limits agents' interaction. In this dissemination model, a message may only spread to neighboring nodes whose opinions are within c of the message opinion.

This project starts the process of fitting the bounded-confidence dissemination model with Twitter data. This involves (1) writing scripts to generate and analyze the structure of data-based social networks and dissemination trees, and (2) creating a rudimentary classification model to assign opinion values to nodes and messages.

Materials and Methods

Datasets were extracted using an Academic Access License for the Twitter API through Python with the Tweepy library.

- We generated underlying followership graphs for a closed set of users by either (1) scraping the full relationship of every combination pair within a set of users or (2) scraping a full following list for every user. The second method works faster with larger user sets, and it produced the following graph[BHJ09]:

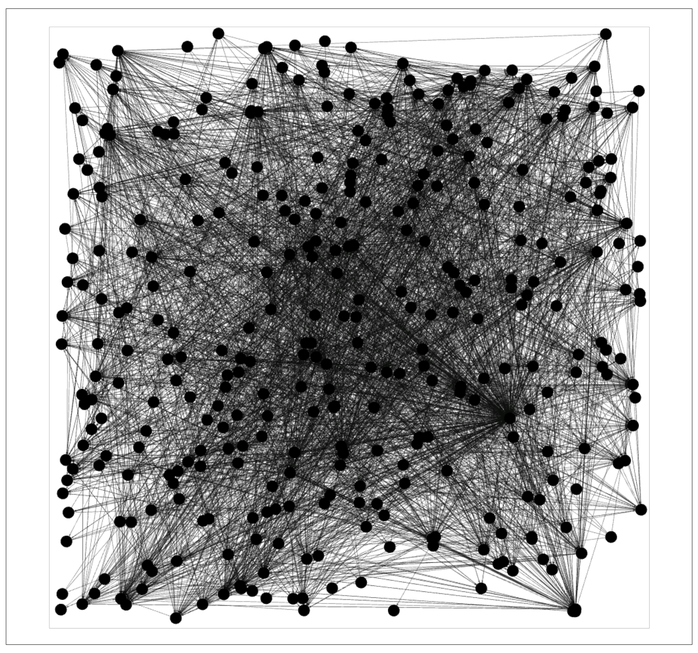

Figure 1: 356 node followership graph of the @HMCAlumni following

- Data-based message cascades originally stemmed from like and retweet data. Later, we based message cascades on the spread of common keywords or sentiments of tweets within a followership network.
- The data-generated opinion-space is based on the pro-life and pro-choice Twitter reactions to the Supreme Court Overturn of Roe v. Wade. This allowed our opinion space to consist of only two discrete values, which simplified data gathering and labeling for our classification model.
- We simultaneously extracted and labeled twenty-thousand tweets and ten-thousand Twitter bios as training data. We first pre-labeled two sets of users who followed either @NARAL or @March_for_life as pro-choice or pro-life (respectively). We then extracted and correspondingly labeled any tweets from these users within 10 days after the overturn that contained the keywords: "roe", "wade", "abortion", or "unborn".

Results

Message Cascade Structure Analysis

By treating a Twitter 'like' or 'retweet' as an outside agent adopting a message, we fit our like and retweet count data to an estimated distribution curve for the out-edge degree of each node in a dissemination tree. A linear approximation function is fitted to the logarithmically-scaled distribution curve, as shown below.

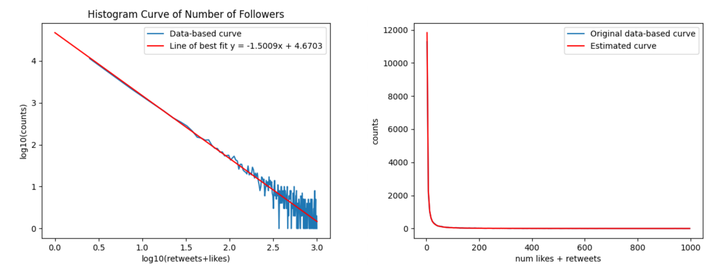

Figure 2: Logarithmically-scaled and normal scaled distribution curves of likes and retweets under 1000.

Opinion Value Classification Models

We built two binary text classification models for our tweet and bio datasets using a linear regression classifier. Unexpectedly, both models perform more accurately on raw uncleaned text.

((a)) Bio classification confusion matrix.

((b)) Tweet classification confusion matrix.

Figure 3: Confusion matrices (data balanced by duplication).[WPS+20]

Conclusions and Future Work

We soon hope to fit a bounded-confidence mechanism with real data. It will enable accurate simulations of online polarization, echo chambers, and the spread of misinformation. But to incorporate a bounded confidence mechanism, we must first train a multi-label opinion classification model.

That said, the current binary text classification model along with other methods created through this project can analyze opinion dynamics in real Twitter networks. The tools developed in this project may offer an interesting data-driven method of studying polarization and homogeneity in different online communities.

References

[BHJ09] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: an open source software for exploring and manipulating networks. In *Third international AAAI conference on weblogs and social media*, 2009.

[WPS+20] WomenWhoCode, Jayeeta Putatunda, Rishika Singh, Ruchika Singh, and Sumana Ravikrishnan. Wwc introduction to natural language processing, Jul 2020.

Funded by NSF DMS-2109239.

Advisor: Heather Zinn-Brooks