Final Project Proposal

Marcus Daly (mrd2194), Anders Geil (ag4290)

October 23, 2020

1 Introduction

Domain: Social networks
Task: Comparing user engagement across fake, factual, corrective, and regular news on Reddit

The motivation for this project is a recent focus on "astroturfing" in social science circles. One definition of astroturfing, offered by Zhang et al., 2013 is as follows: "Online astroturfing refers to coordinated campaigns where messages supporting a specific agenda are distributed via the Internet. These messages employ deception to create the appearance of being generated by an independent entity."

Here, we are specifically concerned with user engagement on fake news stories on Reddit. With the increasing prevalence of bots and fake accounts on social media, we hypothesize that the growth of a fake news story may be different from the 'organic' growth of other types of news posts. In this project, we want to (1) compare the total engagement of fake and other news stories on Reddit, and (2) understand what drives the growth in engagement of a news post.

2 Data

The raw datasets consist of the complete collection of Reddit posts stretching from January 2016 to April 2020, and a large collection of ClaimReviews (fact checks) covering the period from January 2015 to October 2020. This counts a total of 151,025 ClaimReviews and 695,147,095 Reddit posts. For each Reddit post, we also have access to the associated user comments.

The working dataset is a smaller collection of matching pairs of Reddit posts and ClaimReviews, summarized in figure 1. By matching pair, we refer to a Reddit post and a ClaimReview that both link to the same external URL. We distinguish between two sets of matching pairs: Pairs linking to external URLs with disputed claims (typically news articles); and pairs linking to review articles with corrective or clarifying information (e.g. an investigative review by snopes.com). Additionally, we also collect a sample of regular Reddit news posts as a control group.

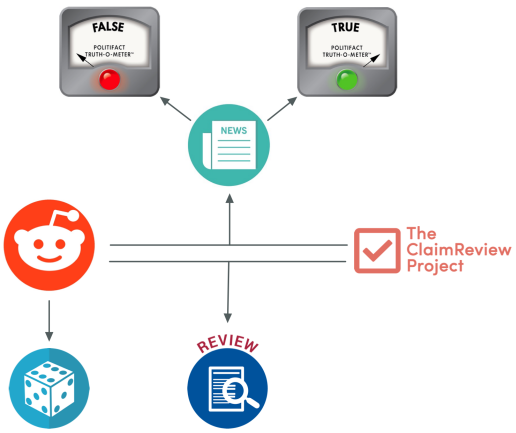

Figure 1: Diagram showing the relationship between the four different groups (random control, review articles, false news, and true news).

Figure 2 shows the distribution of our data. In total, we have 12,079 matching pairs linking to disputed claims, representing 2,289 unique stories. This highlights the fact that the same external news story can be crossposted to multiple subreddits. In contrast, we have 27,213 matching pairs linking to corrective news, representing 12,918 unique corrections. Note that under this methodology, there exists a corrective news article for each disputed news article. But there is no guarantee that the disputed and the corrective news articles relating to the same news story both have been posted to Reddit.

3 Model

3.1 Definitions

First, we define the "engagement" of a story s as the total number of comments over all posts that share the same story.

Next, note that each story can be categorized into one of four different groups:

0. "Natural", undisputed story (control)
1. Factual, disputed story
2. Fake, disputed story
3. Corrective story

Each story has exactly one value of a categorical variable $t \in T = \{0, 1, 2, 3\}$ corresponding to the type of the story as defined by the numbers in the above list.

3.2 Bias

The first question we wish to answer is the following: Is there a systematic difference in how fake, factual, and corrective news stories build engagement compared to "natural" news stories? To answer this question, we must find the bias of the engagement for any one group. i.e., the difference in expected values of engagement for any non-control group versus for the control group: $\forall t' \in T' = \{1, 2, 3\}$,

$$EngagementBias_{t'} = \mathbb{E}[Engagement|t = t'] - \mathbb{E}[Engagement|t = 0] \quad (1)$$

If we do find that the level of engagement is biased in some group, we will be interested in explaining why. If not, we will be interested in explaining what drives user engagement in general.

3.3 Baseline model

Our initial model, shown in Figure 3, assumes that engagement is generated as a function of not just the posts, but also the news story linked to the post. In this way, we can imagine the level of engagement for a distinct post, y_{pst}, to depend on the story-level effects on engagement, θ_{st}, specific to the associated news article.

Figure 2: Number of ClaimReviews (top), matching news articles (upper middle), matching corrections (lower middle), and Reddit posts (bottom) over time.

The models we will use for both this baseline model and the extended model are based on different multilevel models as described in Gelman and Hill, 2007.

Additionally, we also assume that news stories are predisposed to a certain level of engagement based on their type; i.e. fake, factual, corrective, or control. Similar to how we expect engagement to depend on story-level effects, we expect the story-level effects θ_{st} to differ based on the type-level effects ϕ_t, specific to the type of story. These type-level effects are, in turn, generated by the shared hyperparameter $\eta \sim f(\alpha)$.

We model this by ϕ_t at the news type-level, and let this parameter be generated as a function of the shared parameter η which is selected by the fixed hyperparameter α.

The nested structure of our model reflects the idea that each story and post belongs to the same type, t. Similarly, each post is associated with one story, s. And each story may also generate multiple posts, p. This model then lends the following factorized joint distribution:

$$p(\alpha, \eta, \phi, \theta, x, y) = p(\eta \mid \alpha) \sum_t p(\phi_t \mid \eta) \sum_s p(\theta_{st} \mid \phi_t) \sum_p p(y_{pst} \mid \theta_{st}, x_{pst}) p(x_{pst})$$

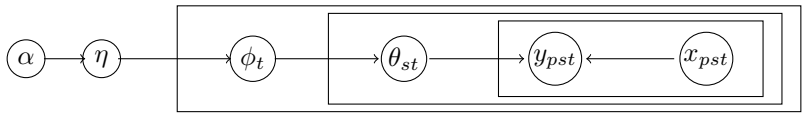

Figure 3: Baseline model. The fixed hyperparameter α induces the hyperparameter η shared between each type of story. The type-specific engagement levels ϕ_t affect the story-specific engagement levels θ_{st}. The parameter θ_{st} in turn, along with the data x_{pst}, influences the engagement of a post y_{pst}.

3.4 Extended model

In a more extensive model, shown in Figure 4, we also acknowledge the hierarchical nature of the Reddit platform itself. In particular, we may expect that the engagement of a post is also dependent on the subreddit it is posted to. Since each subreddit may differ in its effects on engagement (e.g. due to number of regular subreddit users), we incorporate the parameter ρ_r to account for subreddit-level effects. This parameter is again drawn from a hyperparameter, $\tau \sim g(\beta)$, shared across all subreddits.

In this view, we can intuitively think of each post as belonging to a subreddit. At the same time, we can also acknowledge that a subreddit may contain multiple posts. This results in another nested model, which, when combined with the story and type-specific effects, produces two hierarchies, overlapping at the post-level. We can factorize the corresponding joint distribution as follows:

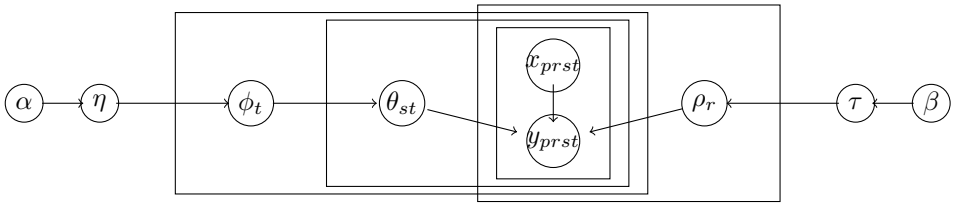

Figure 4: Extended model. The fixed hyperparameter β now induces the shared hyperparameter τ over each subreddit, setting the subreddit-level parameters, ρ_r. In unison, the subreddit-, story-, and post-level data now produce a number of total comments, "engagement," captured by y_{prst}.

$$p(\alpha, \beta, \eta, \phi, \rho, \tau, \theta, x, y) = p(\eta \mid \alpha)p(\tau \mid \beta)$$
$$\sum_t p(\phi_t \mid \eta) \sum_s p(\theta_{st} \mid \phi_t) \sum_r p(\rho_r \mid \tau) \sum_p p(y_{prst} \mid \theta_{st}, \rho_r, x_{prst})p(x_{prst})$$

4 Inference

To perform inference, we plan to use MCMC to approximate the posterior distributions of our regression variables. Our inference should be fairly straightfoward: after defining our regression model as above, we will use Pyro to calculate the necessary gradients and perform MCMC to find the posterior distribution of each regression variable. Based on the output of MCMC, we will be able to interpret the regression coefficients accordingly.

5 Evaluation

To evaluate our process, we will use various posterior predictive checks to make sure that key statistics of data simulated by our process look like the data we actually observe. Namely, we plan to perform checks on at least the following statistics:

1. Maximum engagement, to see if our process accurately models how much engagement the most popular stories receive.
2. Engagement percentiles (e.g. 10^{th} percentile, median, 90^{th} percentile), to make sure our process accurately models the distribution of engagement.

As we see the natural distributions of our data, we will likely find additional checks which may be particularly pertinent to our data and add these on as we come across them.

Additionally, we want to get a sense for the explanatory power of our models. To accomplish this, we will use various metrics of "fit" such as R^2 and Adjusted R^2. As we introduce additional complexity to the model, we would hope that fit will rise accordingly.

6 Drawbacks

One drawback in our current choice of model is that we cannot capture dependencies from one post to another. In reality, it is possible that popularity would travel from one posting of a story to the next. That is, if the first posting of a story is particularly popular, some of the users who commented on that first post might be looking for more places that may have posted this story and come across a post of the same story on another subreddit, thereby increasing the popularity of this second post due to the high popularity of the first post. Although our model can capture variation that comes from the story itself, it cannot capture this temporal dependency from post to post. Still, we are able to roughly capture this phenomenon by including post-level attributes such as the time since the first posting of story.

7 Possible Modifications

7.1 Temporal dependency

A common phenomenon on Reddit is crossposting. That is, posting the same news article to multiple subreddits in sequence. In this way, the engagement of earlier posts may well inform the engagement of later posts. A desirable modification would be to add something to the model to directly capture this temporal dependency. One possible idea is to consider the current post as "inheriting" from each of the prior posts in a story, introducing a multiple-membership aspect to our model, where each recent post "belongs" to multiple ancestor posts. A similar model was proposed as a Multiple Membership Multiple Classification (MMMC) model in Browne et al., 2001.

7.2 Communities & Topics

One major benefit of using a multi level model with a flexible inference tool is that we can easily include additional levels to our model. One level that may be especially interesting to include for interpretability is a subreddit cluster level. A group of subreddits may share some underlying topics of interest and common users. Together, this group can be thought of as forming a "community". This means that there may be some underlying differences in engagement that arise from not just the subreddit level, but even the subreddit cluster level. Therefore, by finding a meaningful way to cluster subreddits together into relevant community-level groups such as "Social Right" or "Fiscal Left," we may be able to more easily interpret the engagement of different communities.

In a similar fashion, each claim is associated with a small set of tags representing the general topic of the story. We could also imagine clustering these tags via topic modeling to form a separate topic level at the story side of our model. It would be interesting to see how certain topics may attract more user engagement than others.

References

Browne, W. J., Goldstein, H., & Rasbash, J. (2001). Multiple membership multiple classification (mmmc) models. http://www.bristol.ac.uk/media-library/sites/cmm/migrated/documents/xcmmrev2.pdf

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.*

Zhang, J., Carpenter, D., & Ko, M. (2013). Online astroturfing: A theoretical perspective.