Does red flagging fake news reduce its spread?

Suhas Gupta
MIDS W241
Experimental Research Proposal
Essay 2

Motivation and Background:

"Fake news" and spread of misinformation were two of the accusations levied on social media after the 2016 US presidential elections that were supported by investigations and public opinion. Social media includes the popular platforms of Facebook and Twitter along with the multitude of opinion sharing blogs platforms prevalent on the world wide web. The purpose of such spreading misinformation can range from benign humor to grave sabotage of democratic institutions around the world. Studies show the prevalence (*Silverman 2016, Gottfreid and Shearer 2017*), effect on public opinion (*Schaedel 2017, Silverman and Singer-Vine 2016*) and impact on electoral outcomes (*Allcott and Gentzkow 2017*) of such misinformation. Today, debate rages on about whether the popular social media platforms have a responsibility to flag obvious or veiled misinformation posted on the sharing medium provided by these companies. It is interesting to understand and study the tradeoffs between protecting the freedom of speech and safeguarding the democratic nature of the electorate. In order to make this tradeoff, companies need to understand if there is indeed a causal effect of *"red flagging"* fake news on reducing its spread. This experimental study proposes to measure this causal effect, if any, and contribute to a greater goal of reducing misinformative, chaotic and fake news around the globe. The research question of this study is *"Does red flagging fake news reduce its spread?".*

The aftermath of the 2016 US elections resulted in many Americans doubting the information presented to them on news channels, online newspapers and public figures. This is one of the goals of the perpetrators of fake news: i.e. reduce belief of the people in legitimate news sources and democratic institutions. Some evidence exists that flagging of misinformation can have an impact on its believability. *Ecker, Lewandowsky, and Tang (2010)* found that warnings are effective at reducing the influence of false news on beliefs, but it doesn't eliminate it completely. In this study we will investigate whether specific interventions like the ones implemented by Twitter *(Link*) are effective in reducing belief in fake news.

Experimental Design:

This study will test the effect of specific "warning" tags attached to prevalent and artificially generated fake news on the latter's believability. The null hypothesis that we will test is the following:

H_0: *Presence of specific warning tag on a misleading article have no impact on its perceived believability relative to a no warning fake article*

The alternate hypothesis in this case will be as follows:

H_a*: The presence of specific warning tag on a misleading article reduces its perceived truthfulness relative to a no warning fake article.*

A sample of 200 people will be served online surveys containing fake and true headlines collected from online social media. Some of the misinformation will also be artificially generated based on the recent events at the time the survey is served to people. This will ensure that people have an opinion or interest in the survey questions in the experiment. The sample will be divided into a treatment and control group. The treatment group will be served with surveys that have disputed or rated flags attached to the misinformation and no flags attached to the true information. The control group surveys will have no tags attached to any of the headlines.

The fake news articles or headlines served in the experiment will be drawn from a pool of headlines over the past one-year period (to ensure current affairs). There will be both pro/anti Democrat and Republican statements along with neutral statements. The statements will be tested using Amazon's Mechanical Turk with a 500 respondents and statement scoring above 90% will be placed in the appropriate category.

Covariates:

Belief in fake information can be strongly correlated with the political affiliation of the individual. This is also an example of confirmation bias which is the tendency to accept new information as evidence of one's pre-conceived notions. E.g. political affiliation with the democratic party in the US would make a respondent question news on Donald Trump and disbelief in suspicious looking headlines. A partisan republican respondent, on the other hand, might have a higher believability for the same news headline. Demographics information, economic status, age and education level are also expected to be correlated with the gullibility of an individual to fake news. Thus, demographics (like race and gender), income, age, political affiliation and education level will be recorded for each survey respondent to perform regression on the covariates and answer the causality question of this study.

Randomization, Blocking and Independence:

In order to control for imbalances in gullibility to fake news, three blocks will be created based on political affiliation, education level and gender. The question of gender specific propensity to believe in fake news is interesting. There are inconclusive studies about whether men or women are more susceptible to believe in fake news if presented by a known source or posted on the "wall of a Facebook friend". In order to ensure there is no selection bias in the experiment, I will safeguard the randomization sample by blocking on all three covariates (political affiliation, education level, gender). A pilot study will be run on a small sample of 50 people to test whether randomization has been correctly performed and whether any additional covariates are present that need to be blocked in the treatment and control assignments. The sample will be selected from a network of online acquaintances, classmates, friends and family and their extended network. The beauty of this sample acquired through online networks is that all the respondents will be familiar with the use of social media platforms and current survey

questions presented to them during the control and treatment phase. Within each block, people will be randomly assigned (using outcomes of a binomial distribution) to treatment and control groups. Another advantage of online survey is that it ensures independence among sample units since participants are unlikely to discuss survey and its results with each other.

Outcome Measures

The primary outcome variable in this study is the belief in fake news after a red flag has been attached to it.

To measure the efficacy of the treatment, respondents will be served 15 news headlines pooled from social media platforms. The headlines or stories will be varying in content from political ads, world affairs, economy and COVID-19 related news. The stories will vary in their subtleness of presenting fake information as well as in the presentation (i.e. some stories will be accompanied with photos). The source of the headlines will be presented alongside each fake or real piece of information. For each news headline two outcome variables will be measured

1: Binary (1: yes or 0: no) for whether the respondent believes the news is fake or real
2: Level of confidence in the detection of fake news (if question #1 is answered as yes)

Statistical Analysis

The null hypothesis that there is no impact of marking fake news with a flag on its believability. The average treatment effect (ATE) will be the difference in means of the treatment and control groups where tags are present and absent respectively on the same news articles. I plan to perform regression analysis on the experimental data to understand the prediction power of the covariates. We expect the partisan affiliation and education to be significant predictors in the outcome variable (belief in news regardless of the red flag present). Regression coefficients and standard errors can help determine the causal impact of red-flagging fake news articles. This regression analysis will also help understand the other predictors and their impact on the spread of fake news: literacy, gender and social status which may become part of a larger study to assess the spread of misinformation around the world.

Conclusion

Detecting and flagging fake news is an important and pertinent problem whose solution would benefit the democratic institutions around the world. The experiment proposal presented here attempts to determine the causal impact of one potential solution to the problem: “Flagging of misinformation”. Modern software and machine learning tools provide social media platforms the ability to implement this flagging infrastructure only if the causality of various covariates can be determined.

References

1. Silverman, C., & Jeremy S.-V. (2016). Most Americans who see fake news believe it; new survey says. December 6, 2016. Retrieved May 23, 2017
2. Schaedel, S. (2017). Black lives matter blocked hurricane relief? Factcheck.org, September 1, 2017. Retrieved September 26, 2017

3. Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. Journal of Economic Perspectives, 31(2), 211-236.
4. Ecker, U. K. H., Lewandowsky, S., & Tang, D. T. W. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. Memory & Cognition, 38(8), 108-1100.