Does labeling opinion pieces impact their persuasiveness

Jay Zuniga, Armand Kok, Younus Ahmed and Kevin Pang

UC Berkeley, w241-2 December 11, 2018

Abstract

The purpose of our experiment was to see if labeling articles written like factual news articles would impact its persuasiveness. Our goal in doing this was to see if this could be part of a solution with combatting disinformation. To do this, our group picked a neutral article and had participants in our experiment a version of this article. For the control group, the article was unlabeled while for the treatment group, the article was labeled as an opinion piece. We then measured the persuasiveness of the article by asking both groups in a survey to rate the articles on the dimensions of factualness, effectiveness and how likely they were to recommend it. While a significant treatment effect was detected on the effectiveness dimension, there were no significant treatment effects detected on the other dimensions. Nonetheless, there remain areas that would be interesting to come back to.

Introduction

The current level of discourse in the country is toxic and polarized. One of the reasons that this situation has gotten is the fact that it is very easy to find online content that suits and reinforces one's biases and beliefs, regardless if one's beliefs are based on facts or based on 'alternative facts'.

While disinformation has always existed, the group wanted to study if there are simple things that could be part of a larger solution to combat this. Using ideas in a proposal made by Kyle Chuang "Does labeling opinions as such influence the perception of the readers?", the group thought of testing the hypothesis that the simple labeling of opinion articles (where it was not clear they were opinion articles) could impact their persuasiveness. If this were true, a general solution of labeling online content that was not 'factual' could decrease the strength of their message and sap some of the power of 'fake news'.

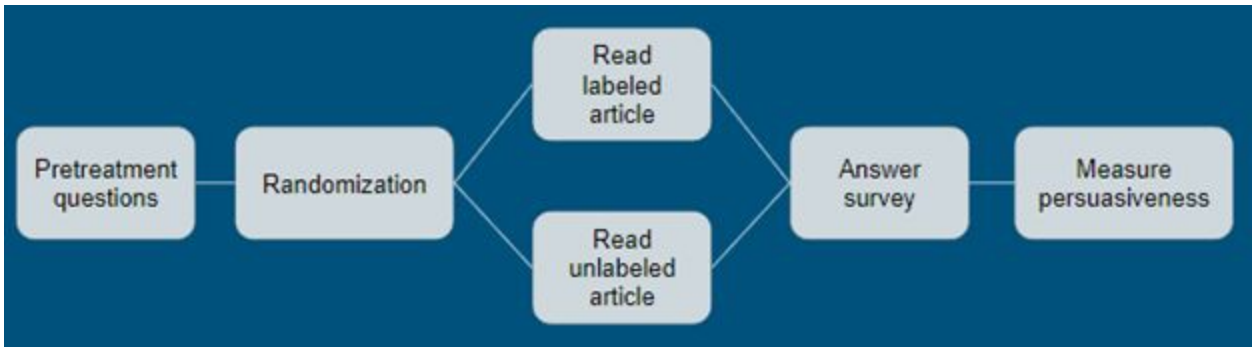
Related Work

According to Duke Reporter's Lab [1], news organizations aren't doing enough to help readers understand the difference between news, analysis and opinion. Readers often come to articles from links in social media and don't know if an article is published in a news or opinion section unless it is labeled. The findings are significant because journalists and educators are focusing on article labels as one way to address the decline in trust of the news media. Labels help readers distinguish between news and opinion so they better understand different forms of journalism and can assess allegations of bias.

Hypothesis

Our group's hypothesis is that the ambiguity of whether an article is fact-based or opinion strengthens the persuasiveness of its position. Author opinion can be misinterpreted as fact making its core argument more believable. The group wanted to test if taking out this ambiguity by labeling an opinion article as such would decrease the persuasiveness of the article.

Experimental Design



Data Collection Process

To get the necessary participants for the group to measure the effect, the group considered a number of options, including:

- Using our combined personal networks
- Social Media Ads, in particular Facebook Ads
- Amazon Mechanical Turk

We decided on Amazon Mechanical Turk mainly for the following reasons:

1. The speed of the data collection. Our research led us to believe that Amazon MT would have a faster rate of data collection since people are being paid to fill out the survey. As opposed with Facebook Ads we were concerned that the data collection process would be much longer since people would have to voluntarily fill out the survey after clicking on the ad.
2. With Facebook Ads, there is a much higher risk of the participants attritioning out of the experiment, as they can freely exit out of the survey mid-way without suffering any consequences. As opposed with Amazon MTurk, the attrition risk is lower, because participants who do not complete the survey do not get compensated for their time.
3. Amazon Mechanical Turk has competitive pricing, with $.25-.50 per participant. While the pricing was similar to Facebook Ads, our research convinced us that Amazon

Mechanical Turk has a geographically broad reach, which the group used to target all of the United States, in an effort to get very wide representation on the survey.

4. Amazon Mechanical Turk has a lot of options for blocking in case we needed to do so.

Treatment and Control Group

We estimated of needing 100 subjects in each group via a power calculation in R (i.e. power.t.test(delta=.15,sd=0.37,power=.8)). The treatment group are the subjects who received the article with the opinion label printed near the title of the article. The control group on the other hand are the ones who received the article without the opinion label near the title of the article.

Randomization

We utilized qualtrics to help randomize the assignment of treatment. Right after the subjects finishes their pretreatment question at the beginning of the survey, they are randomly given either the control or treatment article (50/50 split).

Treatment Article

The labeled article, received by the treatment group.

We Know How to Conquer Tuberculosis

Why aren't outbreaks in poor countries treated the same way as those in rich ones?

By The Editorial Board

The Editorial board represents the opinions of the board, its editor and the publisher. It is separate from the newsroom and the Op-Ed section

Sept. 26, 2018

Why aren't outbreaks in poor countries treated the same way as those in rich ones?
In 1962, the renowned epidemiologist George Comstock had a realization that would help rid modern America of one of the world's enduring scourges. Despite the advent of antibiotics, tuberculosis had remained endemic in parts of the country. Those miracle drugs were good at curing individual cases of TB, but people could pass the disease on to others long before they developed obvious symptoms, received proper diagnoses or were effectively cured.

Control Article

The unlabeled article received by Control group.

We Know How to Conquer Tuberculosis
Why aren't outbreaks in poor countries treated the same way as those in rich ones?

Sept. 26, 2018

Why aren't outbreaks in poor countries treated the same way as those in rich ones?
In 1962, the renowned epidemiologist George Comstock had a realization that would help rid modern America of one of the world's enduring scourges. Despite the advent of antibiotics, tuberculosis had remained endemic in parts of the country. Those miracle drugs were good at curing individual cases of TB, but people could pass the disease on to others long before they developed obvious symptoms, received proper diagnoses or were effectively cured.

Measurement

Outcome Variables

After the article section of the survey, the subjects are then asked four multiple choice questions in order to measure the effect of the treatment, specifically:

- Question 1: Was the article factual?
    - Choices: Yes, No, Maybe
- Question 2: How would you grade the article's effectiveness with arguing for Anti-TB initiatives?
    - Choices: 1-5 (1-not very effective, 5-very effective)
- Question 3: How likely are you to support Anti-TB initiatives given the chance after reading the article?
    - Choices: 1-5 (1-not likely, 5-very likely)
- Question 4: How likely are you to recommend or share this article with others?
    - Choices: 1-5 (1-not likely, 5-very likely)

Other Variables

There are several several covariates that we collected as part of the experiment so that they can be used to determine never takers and/or improve the precision of our treatment effect when included in regression. Those variables specifically are:

- Time on the article page
    - This variable informs us about how many seconds each subject is taking on the article page.
    - We used this mainly to determine who the never takers are in both the treatment and control group (i.e. users who spend an extremely short amount of time on this page are considered never takers)
- Basic Demographics information
    - Gender

    - Political affiliation
    - Location
    - Age bracket

Experiment Phases

Pilot

We did a pilot study consisting of 20 observations to ensure that the integration between MTurk would work seamlessly. As part of the placebo study, we were specifically wanted the ensure that the following things worked well:

- Data output:
    - Ensuring that the data that our survey produced are acceptable for the purposes of the analysis.
    - Ensuring that we can connect the data that is gather through Qualtrics for each subject can be matched against the one we get from MTurk
- Timing:
    - We were not sure how long it would take to collect the information through MTurk
    -
- Payment amount:
    - We were not sure how much we need to pay in order to gather the necessary data collection.
    - We started with 10 cents.

Placebo Test

After a successful pilot study, we wanted to ensure that there is covariate balance in our observations, and that our data collection mechanics works at a larger scale. We also wanted to ensure that the randomization process used by Qualtrics was legitimate by confirming that there was no significant non-zero treatment effect between treatment and control. The placebo test aimed at collecting 100 observations in total.

Primary

After the Placebo test found that the covariates were balanced and our tools scaled out properly, we went ahead and did our primary data collection. During this phase, we collected 200 observations altogether for our analysis.

Data

In addition to `labelling the article as opinion` for treatment/control assignment, the features collected can be split into three main categories:

- Outcomes: outcome variables as stated in earlier measurement section
    - Article Effectiveness: 5pt likert-scale
    - Article Factual: yes, no, maybe
    - Support TB: 5pt likert-scale
    - Recommend Article: 5pt likert-scale
- Covariates: these features help us improve accuracy on any significant effects
    - Gender: male, female
    - Political identification: political affiliation of the subject
    - Age: age of the subject
    - Phase: whether the subject participated in the placebo or main experiment
- Meta/Comprehension: these features help us set up a cleaner experiment
    - Location Lat / Lon: geolocation provided by Qualtrics
    - Worker ID: MTurk worker identification
    - MTurk Code: code to link task to qualtrics survey
    - Time To Submit Article: seconds it took to read an article
    - Point Of Article: comprehension check
    - Finished: whether the survey was completed

While exploring the data, we wanted to understand primarily three things:

1. Location of the workers: To confirm that the 310 subjects whom we collected geolocation data are randomized and come from different parts of United States, we plotted their longitude and latitude coordinates onto a map. The result is fairly pleasing that there's randomization and MTurk drew subjects from across the United States. There were few instances (about a dozen) where the subjects reported that they are located in the U.S., but their geolocation data indicate otherwise. For these people, we think the use of VPN could have distorted their geolocations and we chose to retain the subjects.
2. Compliance: we care about both survey completion and comprehension here
    - It's fairly easy to control for completion with the survey being a task and the mechanical turk user has to complete it to claim the reward.
    - For comprehension, it's a bit a tricky. In addition to including a captcha question to filter out bots, we wanted to make sure that our subjects do not simply click through the survey without paying attention.
    - A quick descriptive analysis on article reading time in seconds helped us understand how to identify never takers and always takers so that we can calculate for CACE.
3. Covariate balance: we plotted distributions of covariates and saw pretty balanced splits between treatment and control groups.

Analysis

For this experiment, we cared about two types of treatment effects: ITT and CACE. The ITT is significant because we want to know what the impact is of a potential policy change surrounding the labeling of articles. On the other hand, the CACE is important because we would like to understand the actual treatment effect among those who actually received the treatment. Never takers in this experiment were defined as such if they fell within one of two criteria:

1) The subject quickly read through either the treatment or control article within 30 seconds or less
2) The subject answered the comprehension question with the following answers: 'Kevin Pang can conquer Tuberculosis in his spare time', or ‘Tuberculosis can be treated with diet of gummy bears and diet soda'

We also excluded subjects who did not fully complete the survey.

Placebo Test Results

There were 50 observations in control and 53 observations in treatment. Among the aforementioned covariates, a covariate balance was plotted in order to ensure the randomization process was executed properly. The results showed that there were 16 never-takers in control and 13 never-takers in the treatment group.

Using randomization inference with 10,000 repeated samples, each of the outcome variables satisfied the null hypothesis of no significant difference between the potential outcomes of the treatment and control groups. Figure 1 is exemplary of the typical output we obtained for each of the outcome variables:

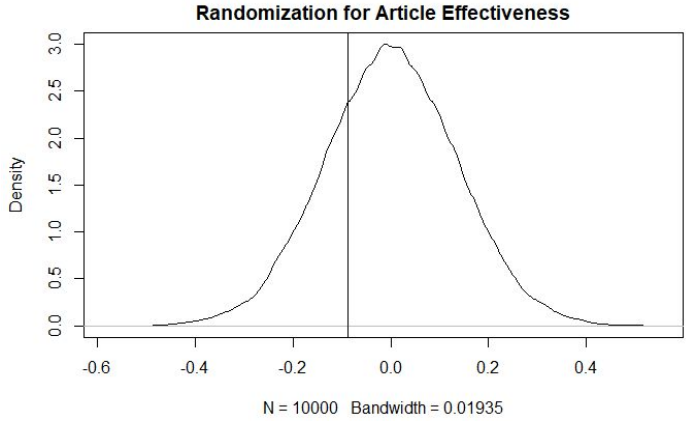


Figure 1: RI on Article Effectiveness Outcome

To verify the output of the randomization inference, we regressed each outcome variable on the treatment, for both the ITT and CACE. Each regression had an identical output to the randomization inference. The coefficient on the treatment variable for each regression failed to reject the null hypothesis that the coefficient was equal to zero. A typical regression can be seen in Figures 2 and 3.

```
#ITT calculation
model_one <- lm(article_effectiveness ~ group, data = placebo_dt)
coeftest(model_one, vcovHC)
```

```
t test of coefficients:

              Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.25490    0.10136 41.9768   <2e-16 ***
groupTreatment -0.10396    0.14465 -0.7187    0.474
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 2: ITT on Article Effectiveness in Placebo Test

```
#CACE
cace_one <- lm(article_effectiveness ~ group, data = d[d$phase == 'Placebo Test' & d$nt_ind == 0,])
coeftest(cace_one, vcovHC)
```

```
t test of coefficients:

              Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.485714   0.096401 46.5319  < 2e-16 ***
groupTreatment -0.285714   0.154985 -1.8435  0.06932 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 3: CACE on Article Effectiveness in Placebo Test

Experimental Results

There were 101 subjects in treatment and 106 subjects in control. A user in the placebo group also participated in the main experiment as well. In order to prevent any potential treatment effect from the previous experiment, this worker was discarded from the analysis. The experimental data showed that there were 13 Never takers in Control and 28 Never Takers in Treatment.

While performing analysis on the article effectiveness outcome, using randomization inference and regression on ITT, it was found that there was no significant treatment effect, as can be seen by Figure 4. However, when the never takers were removed from the analysis, both methods of analysis showed that there was a significant treatment effect, with a CACE of 0.215

and a p-value of 0.04, as shown by Figure 5. This was the only significant treatment effect that was observed in this experiment.

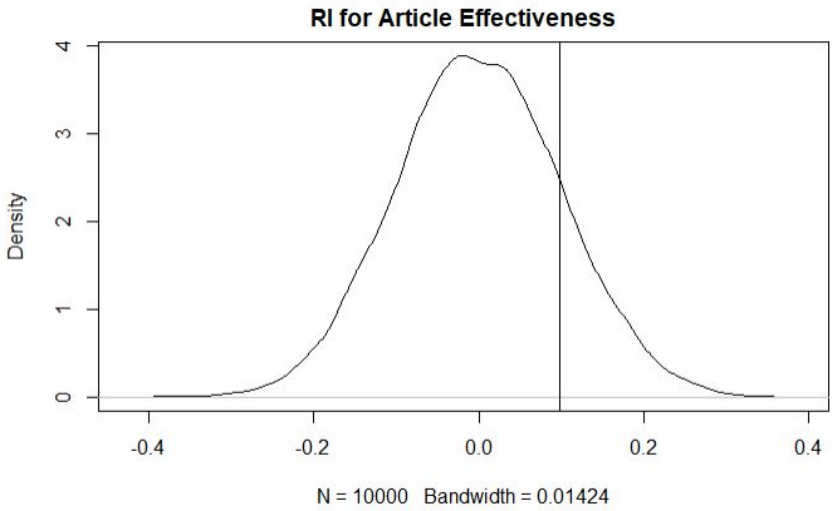


Figure 4: RI on ITT for Article Effectiveness Outcome

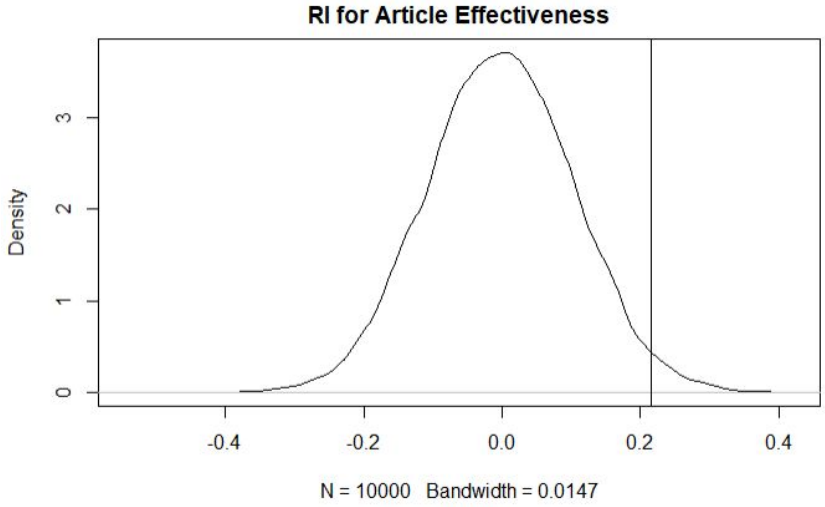


Figure 5: RI on CACE for Article Effectiveness Outcome

In order to test for heterogeneous treatment effects on political denomination, since democrats and republicans may react differently to a medical outbreak, we ran another regression with an interaction term between political affiliation and older and the treatment effect and found no statistically significant coefficient on the interaction term.

After measuring the treatment effect on the outcome of the reader's perception on how factual the article was, it was shown that both the ITT and the CACE for this outcome were not statistically significant. For this variable, three outcomes were compressed into a binary outcome. If the reader answered "yes", it would be considered a 1, and if the reader answered "maybe" or "no", it would be considered a 0.

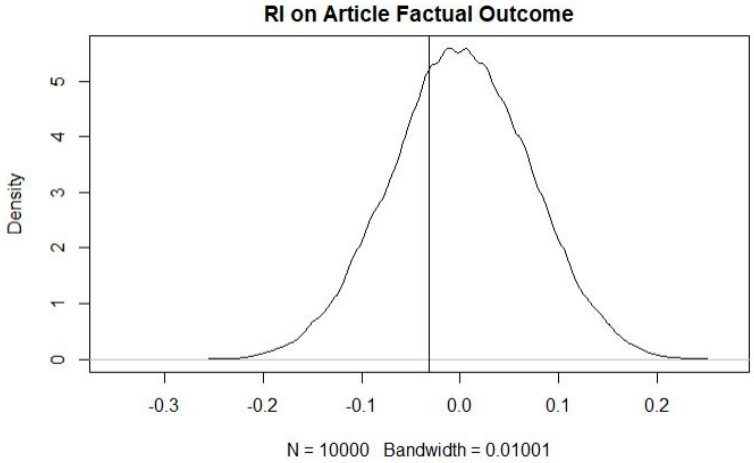


Figure 6: RI on CACE for Article Factual Outcome

It came as no surprise that there was no measured treatment effect on this outcome, given that article contains many factual claims that were true, even though the article was labeled as an opinion.

Figure 7-8 display the output of the CACE assessment of remaining outcome variables. Neither variable was shown to be statistically significant.

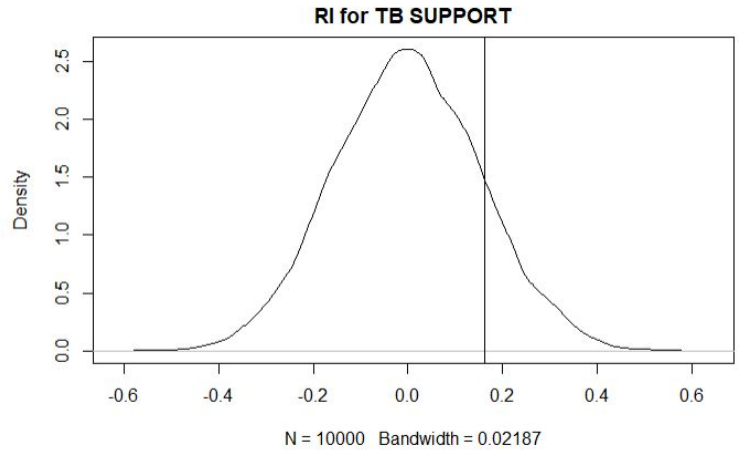


Figure 7: RI on CACE for Anti-TB Support Outcome

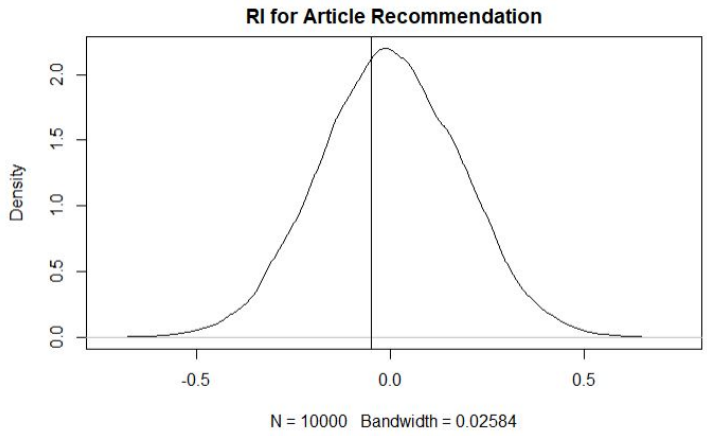


Figure 8: RI on CACE for Article Recommendation Outcome

Conclusions

The experiment revealed a significant treatment effect on the article effectiveness outcome due to the presence of an “Opinion” Label, on the CACE level. No other treatment effects were significant. Although we almost found a treatment effect on article effectiveness outcome, it turned out that most of the variation was explained by age 59 and older group. This finding is congruent with the fact that TB outbreaks were increasingly more prominent as you move backwards in time. Therefore, the older generation would be more affected by the content in the articles.

The covariate balance on both the placebo test and the experiment revealed that the randomization technique was successful and confirmed that we are producing unbiased estimates. We first tried using Google Forms but found that the randomization required some additional javascript coding so that decided to use Qualtrics instead.

Lastly, the average treatment effect may have been more effective if label was made more prominent on the treatment article or if the article contained less factual claims. Perhaps the barrage of facts in the article were made more noticeable to the reader than the opinion label at the top of the treatment article.

Next Steps

The group felt there were a couple of areas that would be interesting to come back to.

Labeling variations

Our labeling was designed to be low-key but noticeable with the placement just after the article title. It would be interesting to see if the effect would be stronger if we used a label that would be more eye-catching using one or more of the following: larger font size, text differentiation such as bolding, underlining or italicizing, different text color, varying the location of the label, using graphics or putting framing around the label, and finally, using stronger wording.

Type of article

One of the biggest debates on the team was the use of a neutral article where we expected there wouldn't be pre-existing strong positions from our participants. With the conclusion we reached, it would be interesting to re-run the experiment using an article with a polarizing subject such as politics, abortion, immigration to see what effect it would have on participants with a pre-existing strong pro or con position.

Appendix and References

1. https://www.poynter.org/ethics-trust/2017/news-or-opinion-online-its-hard-to-tell/