Perceptions of Labeling Opinion and News Articles in reddit News Feeds

Final Project for W241 Experiments and Causality | Spring 2019
Craig Fujii, Tako Hisada, Cameron Kennedy

Introduction

Americans are increasingly obtaining their news from social media sites as opposed to traditional media outlets such as television, newspaper, and radio. Social media in the form of Facebook, Twitter, Whatsapp and reddit have grown since 2008 when the term “social media” was just starting to gain traction.[1] While these platforms started out as social platforms, they have since taken on a greater role in delivering news content. In 2016, 68% of Americans reported receiving their news from social media [2], and there is growing awareness of partisan polarization attributable to news obtained from social media [3].

Social media news feeds frequently take on minimalist approach in presenting only the title of the article, a link, and usually a photo. Opinion articles are included in these news feeds mixed in with news articles without clear distinctions between the two. In general, a news article contains information about current events where the author is expected to strive for objectivity. In contrast, an opinion article reflects the author's opinion about the subject and includes editorials, columns and op-eds. It is very possible that misperception of opinions as news may contribute to misunderstanding of the news and even worse, partisan polarization.

One potential way to address this issue would be to clearly label opinion articles within a social media’s news feed. This solution follows the same intent to inform the viewer for labeling promoted posts or advertisements within a news feed. This study assesses whether labeling articles as “opinion” affects reader perception of political bias and factualness.

One social media site that may benefit from this practice of labeling is Reddit’s r/politics community where the news feed lacks a systematic delineation between news and opinion article titles. For those unfamiliar with Reddit, Will Nicol of Digital Trends describes Reddit as “a massive collection of forums, where people can share news and content or comment on other people’s posts.”[4] Reddit is growing in popularity where it is ranked in the top 10 visited sites in the United States with an estimated daily time on site at 11:28 (mm:ss) and a unique daily page view of 7.35[1].[5] When this study started, Reddit’s r/politics had 4.7 million subscribers.

[1] Metrics taken on 3/2/19 For comparison Facebook has 09:32 with 3.92 Daily Page views.

In the posts in r/politics, there is usually a link to the actual article. Sometimes the URL presented contains the word ‘opinion’ to tip the reader of the article type, but often the post is void of any information of this kind. A clear label may be helpful and the platform of Reddit provides a convenient format to test the effect of labeling opinion articles.

Research Question and Hypotheses

Research Question: Does adding a label prior to article titles within the reddit news feed to clearly identify those that are opinion based versus those that are news based change a reader’s perception of the article’s political tone and/or factualness?

Hypothesis #1: Adding an “Opinion:” label prior to opinion article titles in the reddit news feed shifts reader perception toward the article being more opinionated.

Hypothesis #2: Adding an “Opinion:” label prior to opinion article titles in the reddit news feed shifts reader perception toward the article being more politically intense (see definition below).

Experimental Design

This study utilizes a within subjects study design with features allowing for between subject comparisons of the treatment and control. The study recruited paid participants where a single survey was administered presenting article titles within a simulated reddit news feeds. Treatment consisted of placing the label “Opinion:” before the opinion articles and “News:” before the news articles. Controls did not have a label. The within subjects design permits each participant to be presented with an equal number of control and treatment article titles and an equal number of opinion and news articles. Outcome measures included perception of political bias and factualness of the title.

Recruitment and enrollment

Amazon’s Mechanical Turk (MTurk) was used for recruitment by placing an ad asking for participants to participate in a study requiring seven minutes of their time for $0.75 in compensation (see Appendix A). In MTurk parlance an assignment is called a Human Intelligence Task (HIT). MTurk workers had to qualify for this HIT where they had to reside in the United States, completed more than 5000 HITs and had at least 99% of their HITs deemed acceptable worked (i.e., Approved). In addition, the HIT was blocked on MTurk qualification of US Political Affiliation (i.e., Liberal vs. Conservative) requiring separate advertisements for each.

To canvas different times of day within the MTurk environment, the pair of ads were posted four different times using Pacific Time as a reference. The first two sought 30 and 45 participants per ad per political affiliation and were deployed on Monday April 4, at 3:50 PM and Tuesday 4/9 3:45 PM respectively. The latter two were deployed in the evenings on Tuesday 4/9 9:30 PM

and Wednesday 4/10 5:20 PM and sought 35 participants per ad. If the MTurk workers chose to participate they clicked on a link to take them to the online survey designed in Qualtrics (Provo, Utah) which can be viewed in Appendix A. The survey URL links had an id to allow tracing back to the HIT advertisement.

Upon arriving at the survey, participants were taken to a welcome page explaining the study, identifying who was conducting study, and its purpose. Description of news and opinion articles were provided. In addition to the MTurk qualifications, participants had to acknowledge they met the following eligibility criteria: they are 18 years or older, live in the United States, are able to read and comprehend English and have a basic understanding of liberal and conservative viewpoints. Participation was voluntary and participants were informed they could stop participating at any time for any reason. The survey contents are described in following sections. Upon survey completion, an eight digit random number was provided to the MTurk worker to confirm their work within the MTurk system.

Treatment

Reddit news feeds are sparse in their presentation as seen in Figure 1. The treatment is to present study participants visual stimuli in the form of a reddit news feed with a label clearly identifying the article as an Opinion (Figure 2) or News (Figure 3). To minimize potential sources of bias, the poster of the article, news feed statistics (e.g. votes, comments, etc) and source of the article were redacted with a line the same color as the underlying font.

Figure 1 Example of a reddit News Feed (Control). This is an opinion article from Newsweek.

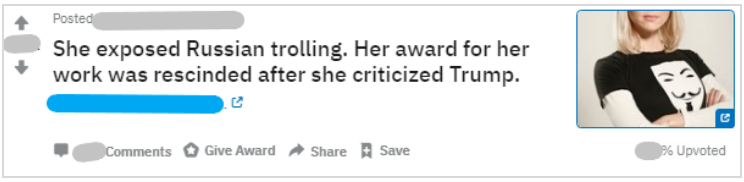


Figure 2 Treatment for Opinion Article

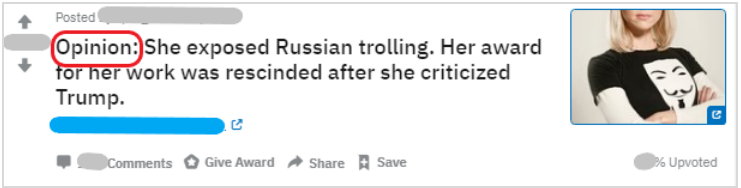
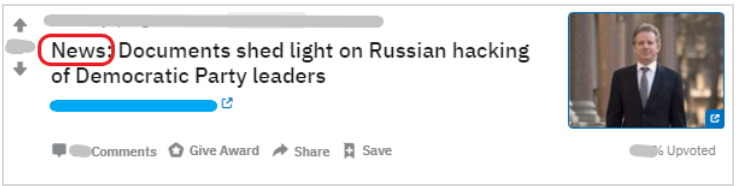


Figure 3 Treatment for News Article



Topics

For topic selection we consulted the ISideWith website, a site that presents questions pertaining to political topics and allows users to vote on how they side. Topics of interest were then searched by keywords within reddit news feeds. Each topic had to have a news article variation and an opinion variation. The process of identifying articles to include was accelerated by extracting the titles from the news feeds using reddit's API.

All selected titles of articles were from sources that had a political bias assessment available at AllSides, a website that ranks media bias by presenting articles with their media bias rating. We also adjudicated each article title for political leaning. News articles consisted of four liberal, one neutral and three conservative. Opinion articles consisted of four liberal and four conservative. This covariate was labeled "Article Tone" in the model, not shown to users, and had values of either "Liberal" or "Conservative."

Each title was assessed for ease of recognition as factual or opinionated using the Pew Research definitions for identifying factual and opinion statements [2]. Factual statements can be proved accurate or inaccurate based on objective evidence. Opinion statements are based on people's beliefs and values and lack objectivity. Two ambiguous titles were included for each article type. Specifically we included two news article titles that present like an opinion and two opinion article titles that present like a fact. As with Article Tone, this covariate was also not shown to the users, and it was termed "Article Clarity," with values of either "Clear" or "Ambiguous" as described above. Summary of titles is presented in Table 1.

Table 1 Topics selected for this study, their media outlet source, and the media outlet's political leaning as reported by AllSides. Political tone adjudication is denoted with superscript L,N,C for liberal, neutral,conservative. Articles with ambiguous clarity have an asterisk.

#	Topic	**News Article Title** *(source - political lean)*	**Opinion Article Title** *(source - political lean)*
1	Green New Deal	Ocasio-Cortez's 'Green New Deal' Fails In Senate. NO Senators Supported It [C] *(Daily Wire - Right)*	We Need a President Who Will Support The Green New Deal [L] *(Newsweek - Slightly Left)*
2	Manafort	Edward Snowden rips light Manafort sentence: 'Your sentence derives from your proximity to power' [L] *(Washington Examiner - Slightly Right)*	Paul Manafort judge deals sharp blow to Robert Mueller's credibility* [C] *(Washington Examiner - Slightly Right)*
3	Russian Trolling	Documents shed light on Russian hacking of Democratic Party leaders [L] *(Washington Post - Slightly Left)*	She exposed Russian trolling. Her award for her work was rescinded after she criticized Trump.* [L] *(Washington Post - Slightly Left)*
4	Joe Biden	Joe Biden's Biggest 2020 Problem Is Joe Biden* [C] *(Huffington Post - Left)*	The Trouble With Biden [C] *(New York Times - Left)*

5	Border/ Wall	Trump To Request An Additional $8.6 Billion To Complete Border Wall [N] *(The Daily Caller - Right)*	Border Patrol union leader: Trump is right, Congress is wrong National emergency on border is real [C] *(Fox News- Right)*
6	Climate Change	Five Things a Democratic President Could Do By Declaring a National Emergency Over Climate Change* [L] *(Mother Jones - Left)*	Students are striking for action on climate change - a truancy everyone should applaud [L] *(LA Times - Slightly Left)*
7	Capitalism	Ocasio-Cortez Blasts Capitalism as an "Irredeemable" System [C] *(Bloomberg - Center)*	Capitalism Needs Elizabeth Warren [L] *(New York Times - Left)*
8	Gun Control	Study: where gun laws are weaker, there are more mass shootings [L] *(Vox - Left)*	An Armed Citizenry Is A Free Citizenry, Which Is Why Liberals Want You Disarmed [C] *(Townhall - Right)*

* Adjudicated as ambiguous. News article title presents like an opinion. Opinion article title presents like a fact

[L] political tone of article title adjudicated as liberal

[C] political tone of article title adjudicated as conservative

[N] political tone of article title adjudicated as neutral

Factorial Design

Each participant was presented 16 reddit article titles consisting of the eight political topics and each article type (opinion or news) per topic; however, half are labeled (treatment) and half are unlabeled (control). In order to ensure every combination is tested in equal amounts, four groupings of article title presentation is required.

Article Title Presentation

Reddit news feeds are stacked one right after another allowing the reddit user to scan titles of articles in a rapid fashion. While the presentation of the visual stimuli in this study could not completely replicate the reddit news feed style, the importance of neighboring news feed was recognized and preserved in this study. We chose to present a “flight” of four titles in a web survey page consisting of two news article titles and two opinion article titles. The control stimuli consist of the unlabeled titles and treatment stimuli consist of prepending the title with ‘News:’ for news articles and ‘Opinion’ for opinion articles.

The permutations of presenting four articles with and without labeling gives rise to four flights: 1) no labeling; 2) only news articles labeled; 3) only opinion articles labeled; and 4) both news and opinion articles are labeled. Figure 4 demonstrates the need for four groups (I,II, III, IV) of these four flights (color coded) intermixed to ensure a balance in the frequency of presentation. This designs allows for a between subject comparison at the topic-article type level since participant is randomized to one group. For example, unlabeled topic 1 - Opinion (1_O) is presented to participants assigned to Groups II and IV and serves as the control. The corresponding treatment, labeled topic 1 - Opinion (1_O_Lab), is located in Groups I and III.

Figure 4 Flight Combinations for Survey Administration

	4 Groups (I,II, III, IV) with four flights per Group			
Group Details				
I	**X1-A**	**X2-BR**	**X3-B**	**X4-AR**
	1_N	5_O_Lab	5_N	1_O_Lab
	2_N	6_O_Lab	**6_N**	2_O_Lab
	3_O	7_N	7_O_Lab	3_N_Lab
	4_O	8_N	8_O_Lab	4_N_Lab
II	**X1-AR**	**X2-A**	**X3-BR**	**X4-B**
	1_O	1_N_Lab	5_O	5_N_Lab
	2_O	2_N_Lab	6_O	6_N_Lab
	3_N	3_O	7_N_Lab	7_O_Lab
	4_N	4_O	8_N_Lab	8_O_Lab
III	**X1-B**	**X2-AR**	**X3-A**	**X4-BR**
	5_N	1_O_Lab	1_N	5_O_Lab
	6_N	2_O_Lab	2_N	6_O_Lab
	7_O	3_N	3_O_Lab	7_N_Lab
	8_O	4_N	4_O_Lab	8_N_Lab
IV	**X1-BR**	**X2-B**	**X3-AR**	**X4-A**
	5_O	5_N_Lab	1_O	1_N_Lab
	6_O	6_N_Lab	2_O	2_N_Lab
	7_N	7_O	3_N_Lab	3_O_Lab
	8_N	8_O	4_N_Lab	4_O_Lab

Randomization scheme

The Qualtrics platform allows for randomization of groups of flights (I,II,III,IV), flight order (X1,X2,X3,X4) within a group, and article presentation within each flight using a choose without replacement method which is repeated once the choices are exhausted. Figure 5 contains the ROXO grammar of the experiment design, including the three levels of randomization as described above.

Figure 5 ROXO Grammar

N R_G R_F [R_A(_O)(_O)(_O)(_O)] [R_A(_O)(_O)(XO)(XO)] [R_A(XO)(XO)(_O)(_O)] [R_A(XO)(XO)(XO)(XO)]

Legend:

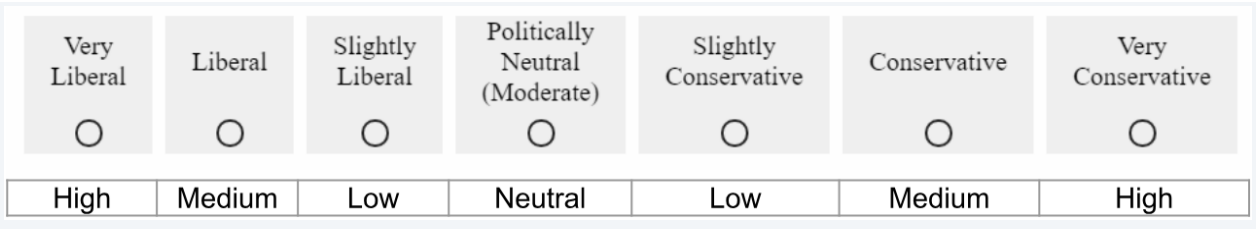
R = Randomization (G=Group Level; F=Flight Level, flights denoted with colors; A=Article Level)
N = Non-equivalent groups
_ = Control
X = Treatment
O = Observation

Outcome Measures

Political Tone and Political Intensity

To measure a person's perception of political perception or bias in an article, the participant was asked to rate the political tone of the article title using a seven point ordinal scale from Very Liberal to Very Conservative with Politically Neutral in the middle (see Figure 6). Since it is possible that different article titles may be polarizing in opposite directions, this rating is converted to an ordinal 'political intensity' rating which measures the degree away from Politically Neutral on the political tone scale. In political intensity, Slightly Liberal and Slightly Conservative are classified as Low Intensity, Liberal and Conservative are classified as Medium Intensity, and Very Liberal and Very Conservative are classified as High Intensity. To prevent polarizing views in one direction potentially cancelling polarizing views in the other direction, all statistical modeling was conducted on this political intensity variable as opposed to political tone.

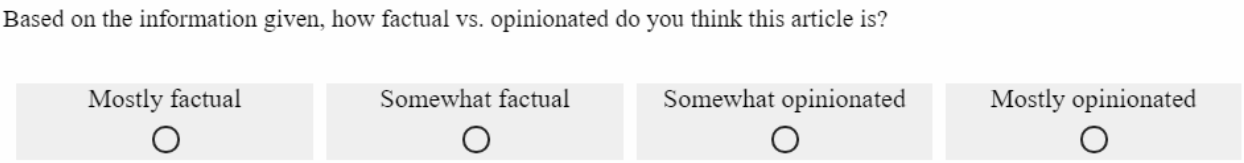
Figure 6 Political Tone Question and Corresponding Political Intensity Rating.



Factualness

Factualness is measured on a four point ordinal scale ranging from Mostly factual to Mostly opinionated but does not have a neutral category (see Figure 7).

Figure 7 Factualness Scale



Survey

The survey consisted of a welcome page to obtain the participant's consent, a section to collect baseline information, a section to experiment with the article titles and a conclusion page to allow for comments and whether the participant noticed any difference between the four groupings of flights. A link to the full survey is in the Appendix B.

Demographic and Baseline Characteristics

The survey next obtains information about gender, age, voter status, ethnicity, household income, and level of education, all of which made for potential covariates to include in studying the primary treatment effects as well as to explore heterogeneous treatment effects. The participants select their political view on a seven point scale very similar to the political tone in Figure 6, identify which political party is most aligned with their views, and rate their interest in politics. The participants' political knowledge and awareness are gauged with a battery of six basic questions. This section of baseline information concludes with a question on how frequent the participant uses reddit and another on how frequently they use other social media.

Experimentation

The survey then pivots to the experimental portion where participants are presented four to five article titles over four web pages. To identify participants or automated bots that randomly select responses without reading the article title, two attention checks were inserted into flights X2 and X3 as a 5th article title to assess. The attention check utilized an actual reddit news feed where instructions on how to answer the questions were inserted after the third or fourth word of the title (see Figure 8).

Figure 8 Attention check with instructions on how to respond within title

Survey Conclusion

The survey concludes with a question asking the participants if they noticed any differences between the 4 blocks of reddit news feeds and if so, to identify the differences noticed.

Pilot Studies

To help inform the design of the study, numerous pilot studies were conducted between 3/30/19 through 4/6/19 to help make improvements to the study design, to the survey, and in study execution. Pilots ranged from one to thirty people per iteration and utilized ourselves, friends and family, and MTurk workers as participants. The inclusion of attention checks resulted from recommendations from MTurk workers. The inclusion of blocking on MTurk qualification US Political Affiliation (i.e. Liberal vs. Conservative) was due to the observation that MTurk pilot studies attracted more self reported Democrats by 4 to 1 versus self reported Republicans.

Pilot analysis indicated that the treatment effect would most likely be detected in opinion articles rated on factualness and possibly also in political intensity. Given the small effect sizes relative to the statistical power calculations performed earlier, the team decided to test only two hypotheses, avoiding a large Bonferroni adjustment that may have negated actual findings.

Statistical Analysis Approach

Significance Threshold Adjustment

Since there are two hypotheses, an *a priori* Bonferroni correction was applied necessitating a p-value of 0.25 to be considered statistically significant.

Proportional Odds Logistic Regression

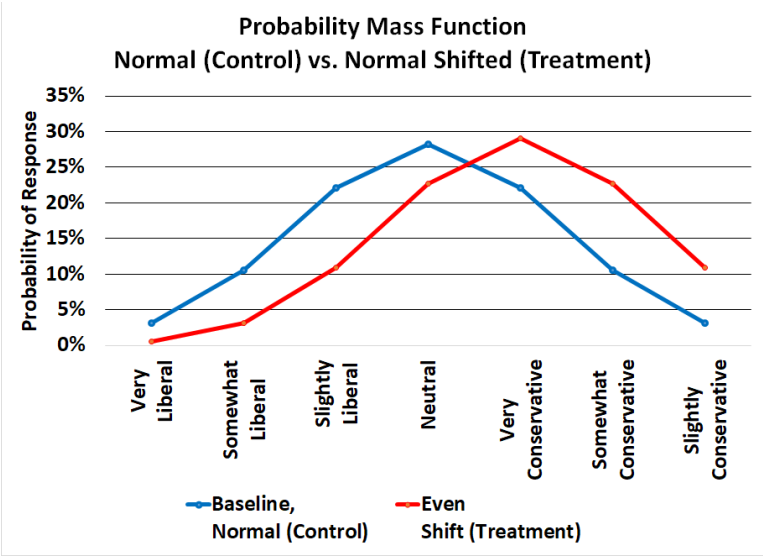
Because the outcome measures are ordinal, results were analyzed using proportional odds logistic regression, specifically using the `polr` function within R's `MASS` library. It is beyond the scope of this paper to describe this method in detail, but at high level, it estimates the log odds of being in a higher vs. lower category for each of the "cutpoints" between categories, and detects overall shifts of moving to higher or lower categories. This log odds is provided as a slope coefficient that can be transformed to estimate the probability of being in the lower category along with standard errors (SE) and t-values to detect statistical significance. These binary probabilities can be further used to estimate the probabilities of being in each ordinal category which are easier to interpret than the log odds and can be presented visually and numerically.

Statistical Power and Sample Size

Power calculations were conducted prior to running the experiment to learn roughly what size of an effect would be needed to be detected with 80% power and a significance level of 0.05. Because the underlying distribution of responses was unknown, the team experimented with shifting three different baseline distributions of responses. After evaluating different baseline distributions of responses, the most and least conservative shifts were selected to gather the full range of estimated required sample size: 1) An even shift of uniformly distributed responses, where an equal percentage of responses in each category moved one category higher; 2) An even shift of a normal distribution, where the mean was shifted such that one unit of shift represented moving the mean by one full category, and the standard distribution remained constant.

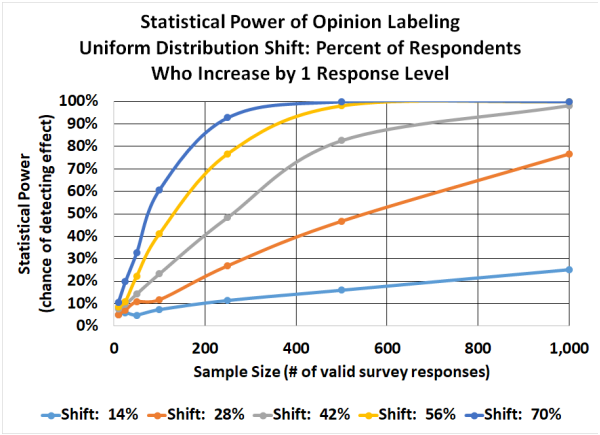
Figure 9 illustrates a one-unit (one category) shift in response for the normal curve described in option 2 above.

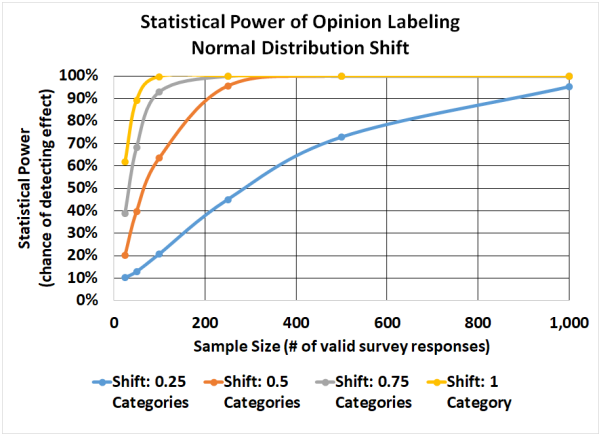
Figure 9 Illustration of One Category Shift in Normal Distribution



The power curves shown in Figure 10 are derived from simulating the results in a proportional odds logistic regression model using various numbers of survey responses (N sizes) and various effect sizes.

Figure 10 Statistical Power Curves





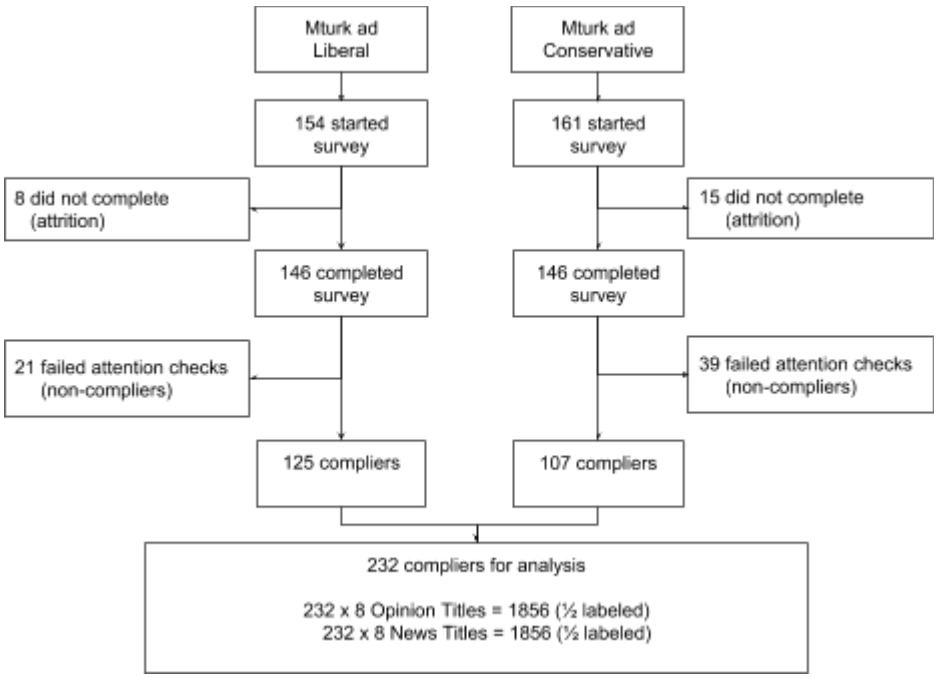
Within these power curves, examples that influenced the survey design included a need for N≅500 to have an ~80% chance of detecting ~42% of responses shifting one category higher under a uniform distribution, but only N≅150 to have an ~80% chance of detecting a shift of ~0.5 categories higher under the normal distribution. These two approaches provided different ways of estimating sample size and proved useful because the team was unaware of either the underlying distribution or the rough magnitude of effect the experiment might reveal. The project budget afforded enrolling 290 participants with 8 opinion observations to yield 2,320 observations with half in control and half in treatment providing 1,160 observations to compare. The sample size calculations provided confidence that a treatment effect may be observed if one exists.

Results

Participant Flow and Demographics

The first two pairs of MTurk HIT assignments were posted on weekday afternoons Pacific time and completed within two hours. In contrast, the second two pairs of HIT assignments were advertised in the evening and required 2.5 to 7.5 hours to complete. This difference is likely due to less Eastern time zone MTurk workers online in evenings. Figure 11 shows the participant flow. There were 315 MTurk participants who agreed to participate where 23 did not complete the survey (8 liberal and 15 conservative; Fisher's Exact Test, p =0.1956). For each MTurk worker that does not complete the survey, the advertisement remains open until fulfilled. This allowed for the 145 spots to be fulfilled in each political affiliation plus one more in each group where someone completed the survey but did not provide their random code to complete the HIT. Conservatives disproportionately did not complete the study compared to liberals at 39 vs. 21 which was statistically significant per Fisher's Exact Test ($p<0.05$). We did state our affiliation with UC Berkeley which may have contributed to this difference given the university's reputation.

Figure 11 Participant Flow



The demographics between the liberal and conservative were not statistically different for gender, number registered to vote, income, education, political interest, the correctness on the political battery questions and social media use. Other than the obvious statistically significant differences such as political party, the participants in the liberal group tended to skew younger

and use reddit more often than conservatives. The conservative group were more white than those identified as liberal. Results can be found in Appendix C.

Primary Findings

In testing hypothesis #1, treating opinion articles by prepending their titles with the “Opinion:” label showed a statistically significant effect on the factualness measure, with subjects in treatment perceiving them to be more opinionated and less factual. For hypothesis #2, this same treatment within opinion articles showed no statistically significant effect on the political intensity measure. Additionally, though not official hypotheses, the treatment showed no statistically significant effects on either outcome measure for news articles.

Primary Model Specifications

Four model specifications with various covariate choices are calculated for each of the four combinations of article type (news or opinion) and outcome measure (political intensity and factualness): 1) Treatment only (no other covariates); 2) Treatment plus Article Clarity, Article Tone, and Subject’s Political View as listed in Mechanical Turk; 3) Treatment plus Article Topic; 4) Treatment, all covariates in Model 2, and numerous others: the subject’s interest in politics, the subject’s political party as indicated in the survey (as opposed to Mechanical Turk), age, gender, income, reddit usage, social media usage, and subject (userID).

The first model was chosen to observe the pure relationship between the treatment and the outcome measures without influence from any other covariates. Models 2 and 3 were chosen to see the specific effects that both article covariates (clarity, tone, and topic) and the subject’s political view covariate -- the key blocking variable -- had on the model. Note that both article clarity and article tone are perfectly multicollinear with article topic, which is why they are omitted from Model 3. Model 4 was chosen to observe the treatment effect and standard error while controlling for the variance of numerous other demographics. Finally, the team confirmed that there was no statistically significant relationship between the order of flights and model outcomes, so flight order was not included as a covariate.

Model specification designations shown along the top of Table 2 contain the number corresponding to the options above, plus O or N for opinion or news articles, respectively, and I or F for political intensity or factualness, respectively. For instance, model 1OI represents model 1: treatment only (no covariates), opinion articles, and factualness.

Table 2 Statistical Findings for Opinion Articles

Opinion Articles, Political Intensity and Factualness

	Dependent Variable: Political Intensity				*Dependent Variable: Factualness*			
Model Specification:	(1OI)	(2OI)	(3OI)	(4OI)	(1OF)	(2OF)	(3OF)	(4OF)
Treatment-Labeled	-0.011	-0.001	-0.004	0.047	0.490***	0.526***	0.533***	0.569***
	(0.085)	(0.086)	(0.088)	(0.104)	(0.090)	(0.092)	(0.094)	(0.112)
ArticleClarity-Clear		0.981***		1.147***		1.293***		1.519***
		(0.098)		(0.103)		(0.102)		(0.110)
ArticleTone-Liberal		0.353***		0.431***		-0.053		-0.034
		(0.086)		(0.091)		(0.091)		(0.097)
SubjectMTurkPolView-Liberal		0.127	0.133	-1.603		0.047	0.065	2.165*
		(0.086)	(0.088)	(1.319)		(0.092)	(0.093)	(1.261)
TopicName-Capitalism			-0.434**				1.071***	
			(0.175)				(0.193)	
TopicName-Climate Change			-0.135				0.056	
			(0.177)				(0.178)	
TopicName-Green New Deal			0.521***				1.050***	
			(0.180)				(0.192)	
TopicName-Gun Control			1.083***				1.261***	
			(0.190)				(0.200)	
TopicName-Joe Biden			-1.825***				0.697***	
			(0.177)				(0.184)	
TopicName-Manafort			-1.564***				-0.478***	
			(0.179)				(0.175)	
TopicName-Russian Trolling			-0.917***				-0.818***	
			(0.172)				(0.173)	
Data Subset, Article Type	Opinion	Opinion	Opinion	Opinion	Opinion	Opinion	Opinion	Opinion
Fixed Interest in Politics	No	No	No	Yes	No	No	No	Yes
Fixed Surveyed Political Party	No	No	No	Yes	No	No	No	Yes
Fixed Age	No	No	No	Yes	No	No	No	Yes
Fixed Gender	No	No	No	Yes	No	No	No	Yes
Fixed Income	No	No	No	Yes	No	No	No	Yes
Fixed Reddit Usage	No	No	No	Yes	No	No	No	Yes
Fixed Social Media Usage	No	No	No	Yes	No	No	No	Yes
Fixed Subject (userID)	No	No	No	Yes	No	No	No	Yes
Observations	1,856	1,856	1,856	1,856	1,856	1,856	1,856	1,856

Note: *p<0.1; **p<0.05; ***p<0.01

The reference categories, i.e., the baseline values within a covariate that are not shown and against which all shown values are compared, are "Ambiguous" for article clarity, "Conservative" for both article tone and the subject's political views as provided in Mechanical Turk, and "Border / Wall" for article topic.

Table 2 shows a highly statistically significant effect of the treatment on perception of factualness, with average treatment effect sizes ranging between 0.490 and 0.569 and standard errors ranging from 0.090 to 0.112. The range between the largest and smallest treatment effect is smaller than smallest standard error among the four model specifications, demonstrating the consistency of treatment effect across all model specifications, helping to confirm there are not harmful relationships between other covariates and the treatment as well as the effectiveness of the randomization process.

Factualness categories are ordered from most factual to most opinionated, and intensity categories from low to high, meaning that positive coefficients indicate a shift toward being more opinionated or of higher intensity, and negative coefficients a shift toward more factual or of lower intensity. The numerical effect size is difficult to interpret because its units are the natural log of the odds ratio (probability of being in a lower vs. higher category) between each category. Instead, a more interpretable outcome is shown visually for factualness:

Figure 12 Primary Finding - Illustration of Shift in Factualness

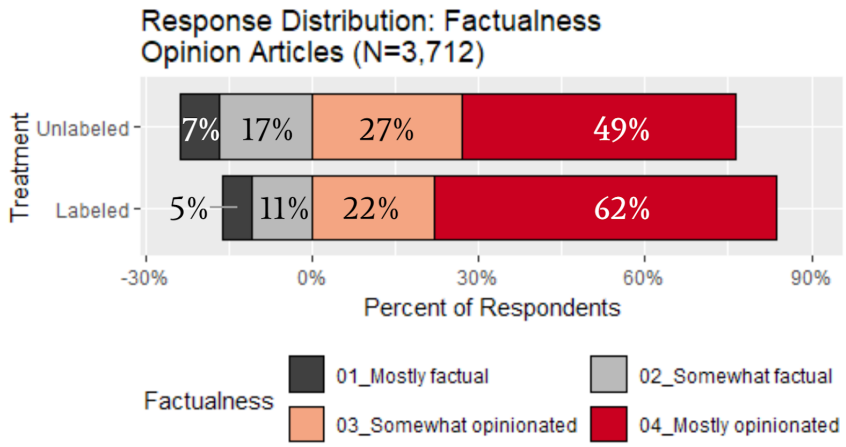


Figure 12 shows the percent of responses in each of the four factualness categories for unlabeled (control) and labeled (treatment) responses, clearly demonstrating that the treatment causes a shift toward more opinionated. Though both groups perceived opinion articles to be more opinionated overall, adding the treatment increased the total number of "somewhat opinionated" and "mostly opinionated" responses by 8 percentage points, from 76% when unlabeled to 84% when labeled. While it would be mathematically incorrect to say the average treatment effect is this 8 percentage point difference, this amount does give a more interpretable rough approximation of the effect size.

The treatment showed no significant effect on political intensity as evidenced in Table 2 by the near zero coefficients for treatment in all models.

The covariates shown in this model reveal a few interesting additional findings, even if not causal. First, articles whose titles were more clear (less ambiguous) were significantly more likely to be seen as both more opinionated and more politically intense. Next, neither the article's tone nor the subject's political views (from Mechanical Turk) correlated with perceptions of factualness, though liberal articles are perceived as more politically intense than conservative ones. Finally, the article itself was also an important contributor of perceptions of both factualness and intensity, the coefficients for which represent the differences compared to the baseline article about the border wall issue. Given the notably different coefficient sizes among all the articles and their relatively small standard errors, there is additional evidence that the effects are article dependent.

Table 3 Statistical Findings for News Articles

News Articles, Political Intensity and Factualness

	Dependent Variable: Political Intensity				*Dependent Variable: Factualness*			
Model Specification:	(1NI)	(2NI)	(3NI)	(4NI)	(1NF)	(2NF)	(3NF)	(4NF)
Treatment-Labeled	-0.06	-0.041	-0.0004	-0.077	0.002	-0.002	0.005	0.002
	(0.084)	(0.086)	(0.087)	(0.102)	(0.083)	(0.086)	(0.087)	(0.103)
ArticleClarity-Clear		-0.116		-0.15		-1.155***		-1.345***
		(0.098)		(0.103)		(0.103)		(0.109)
ArticleTone-Liberal		0.006		-0.006		-0.431***		-0.530***
		(0.090)		(0.094)		(0.092)		(0.096)
ArticleTone-Neutral		-1.878***		-2.196***		-2.186***		-2.598***
		(0.173)		(0.184)		(0.160)		(0.171)
SubjectMTurkPolView-Liberal		-0.352***	-0.364***	-2.881**		-0.228***	-0.244***	-1.497
		(0.086)	(0.087)	(1.187)		(0.086)	(0.087)	(0.959)
TopicName-Capitalism			2.744***				2.885***	
			(0.201)				(0.188)	
TopicName-Climate Change			2.731***				2.812***	
			(0.198)				(0.187)	
TopicName-Green New Deal			1.867***				1.422***	
			(0.197)				(0.183)	
TopicName-Gun Control			2.609***				1.680***	
			(0.197)				(0.184)	
TopicName-Joe Biden			1.399***				3.757***	
			(0.194)				(0.196)	
TopicName-Manafort			1.224***				2.853***	
			(0.196)				(0.188)	
TopicName-Russian Trolling			1.311***				1.193***	
			(0.195)				(0.182)	
Data Subset, Article Type	News	News	News	News	News	News	News	News
Fixed Interest in Politics	No	No	No	Yes	No	No	No	Yes
Fixed Surveyed Political Party	No	No	No	Yes	No	No	No	Yes
Fixed Age	No	No	No	Yes	No	No	No	Yes
Fixed Gender	No	No	No	Yes	No	No	No	Yes
Fixed Income	No	No	No	Yes	No	No	No	Yes
Fixed Reddit Usage	No	No	No	Yes	No	No	No	Yes
Fixed Social Media Usage	No	No	No	Yes	No	No	No	Yes
Fixed Subject (userID)	No	No	No	Yes	No	No	No	Yes
Observations	1,856	1,856	1,856	1,856	1,856	1,856	1,856	1,856

Note: *p<0.1; **p<0.05; ***p<0.01

The reference categories, i.e., the baseline values within a covariate that are not shown and against which all shown values are compared, are "Ambiguous" for article clarity, "Conservative" for both article tone and the subject's political views as provided in Mechanical Turk, and "Border / Wall" for article topic.

Table 3 presents results for news articles and shows almost no difference in perceptions of either political intensity or factualness in any model specifications. This low treatment effect shows meaningful evidence that user perceptions are not affected by labeling for news articles.

Finally, though not mentioned individually, none of the other demographic covariates yielded any meaningful, practical relationships in this study's findings as single covariates to treatment.

Non-Compliance and Attrition

The 21% non-compliers who failed attention checks were added back into the analysis, yielding a small dilutive effect on factualness, dropping the Model 1OF coefficient from a complier average treatment effect (CACE) of 0.490 shown above to the intent to treat (ITT) of 0.408. Models 2OF and 3OF showed similar drops, and Model 4OF remained the same. Because the study aims to generalize to people who actually read news articles, we chose to focus on CACE rather than ITT. There is no way to know the outcomes of those who attrited. We could apply a worse case scenario; however, given the attrition rate was small, at 7%, we did not take any action.

Additional Exploratory Analyses

The findings in this section represent analyses outside testing the experiment's official hypotheses, and therefore are subject to the potential perils of false discovery; that is, "fishing" for these results may have led to these findings by chance alone. Nevertheless, the team felt these results were sufficiently interesting to include, and they may make for good routes to explore if this study were to be extended.

Findings by Topic

Figures 13-16 are broken out by news and opinion articles, and by political tone and factualness, and depict the response distribution for treatment and control at the article topic level, yielding 32 total comparisons.

Figure 13 Response Distribution for Political Tone, News Articles, by Article Topic

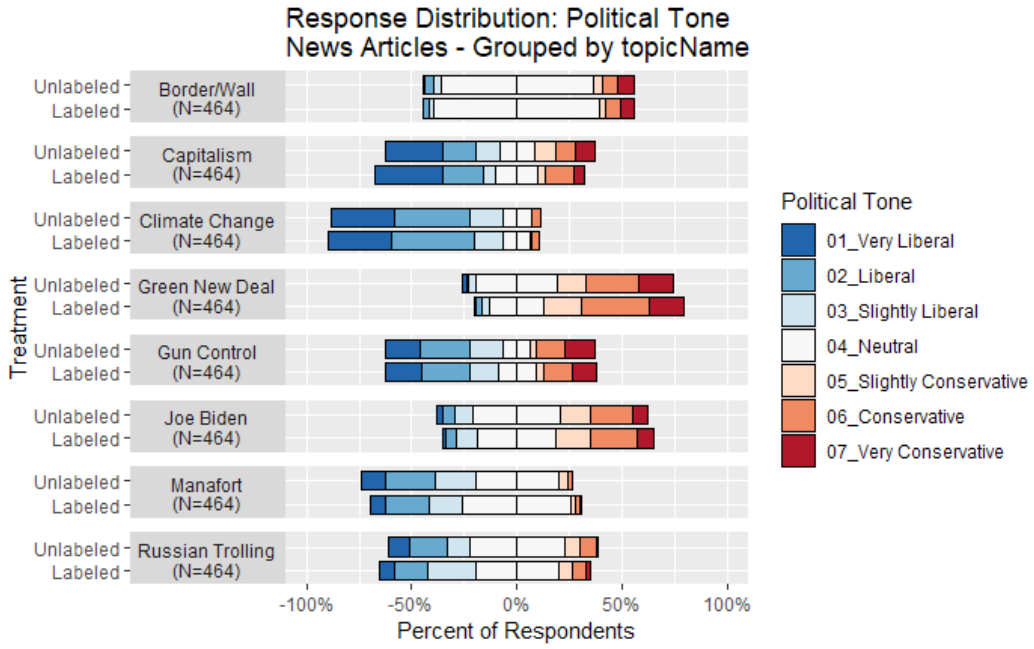


Figure 14 Response Distribution for Political Tone, Opinion Articles, by Article Topic

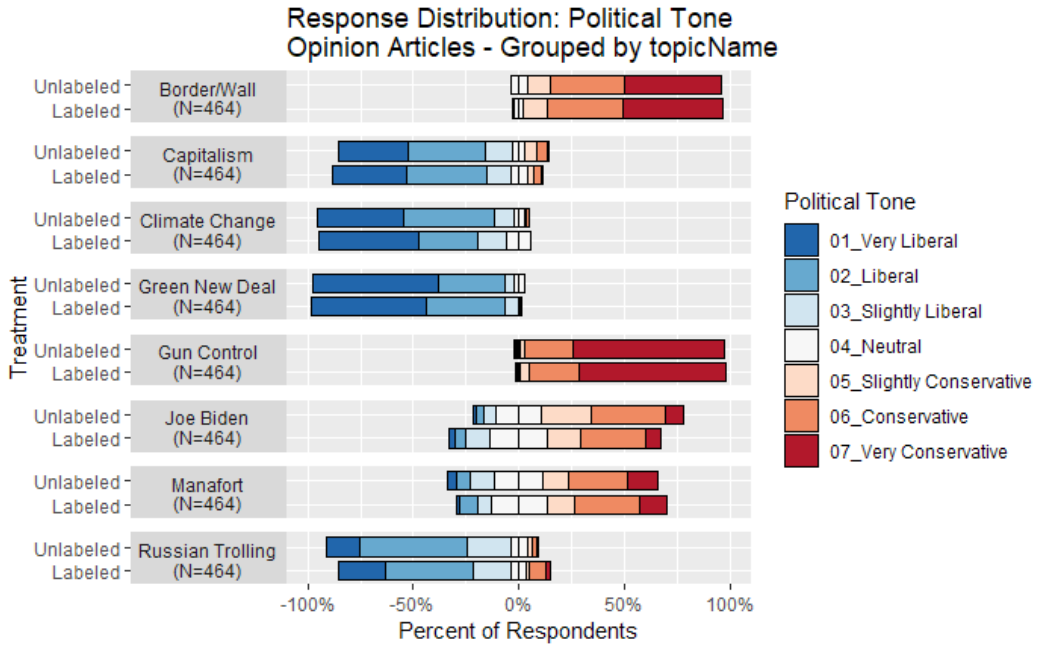


Figure 15 Response Distribution for Factualness, News Articles, by Article Topic

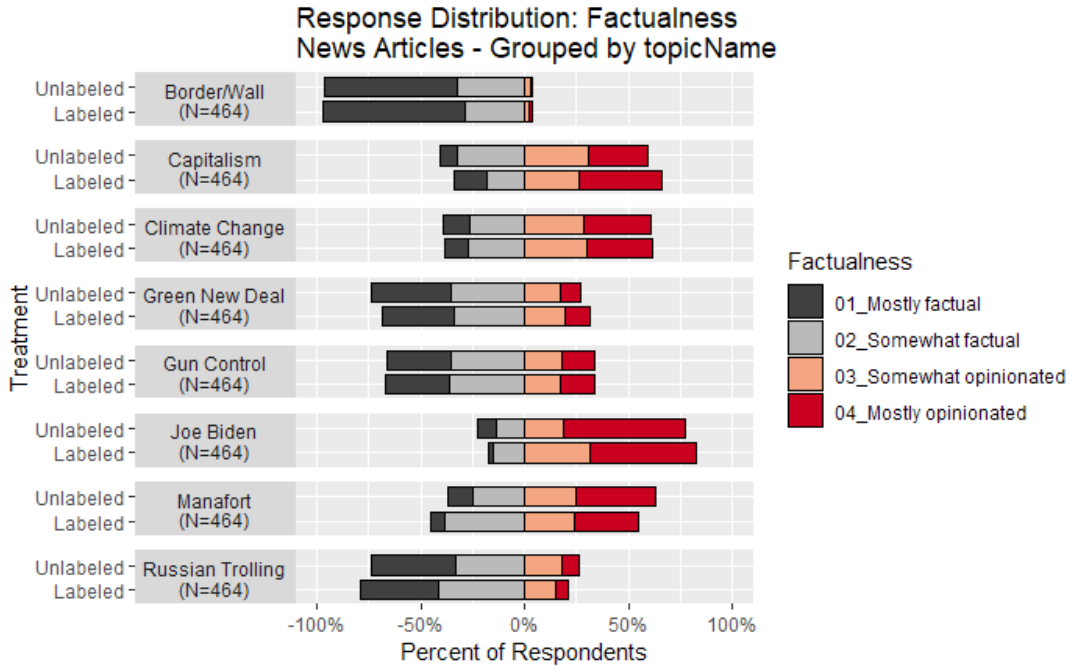
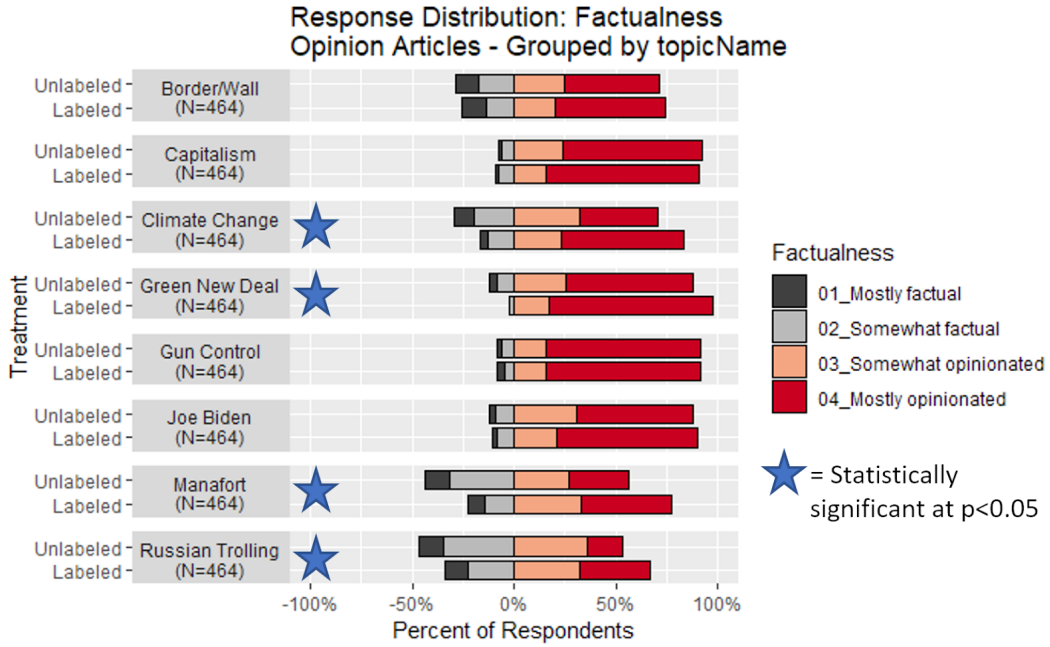


Figure 16 Response Distribution for Factualness, Opinion Articles, by Article Topic



These visualizations mostly show the negligible differences between treatment and control for nearly every combination of article type, outcome measure, and topic. In the political tone graphs (Figures 13 and 14), it shows the clear political leaning, liberal or conservative, of each article, especially for opinion articles (Figure 14). For factualness (Figures 15 and 16), they show how news articles tend to be perceived as much more factual than opinion articles regardless of treatment. While this finding is not revolutionary, it does help confirm the baseline expectation that opinion articles are indeed perceived as more opinionated. Finally, for opinion article's factualness measure (Figure 16), some of the results show noteworthy shifts in response distribution between treatment and control, as marked with stars, which helps answer the question of which articles contributed to the overall finding that adding the "Opinion:" label shifts perceptions. Testing using subsets of the data on these articles revealed that four topics showed statistically significant treatment effects ($p<0.05$): Climate Change, Green New Deal, Manafort, and Russian Trolling. Finally, among political intensity in opinion articles and both factualness and political intensity in news articles, at the article level, none showed statistically significant results between treatment and control.

Heterogeneous Treatment Effect Exploration

In nearly all cases, regression models both using subsets of the data for each value within each covariate and using interaction effects between treatment and each covariate showed no statistically significant results. Additional visual inspection helps confirm this finding, and also shows no practically significant effects in most cases.

The one demographic covariate that showed some statistically significant differences was age. In particular, the treatment had a statistically significant effect on younger respondents. Subsets of the data with respondents age 20-29 had a treatment effect of 0.911 (SE 0.243) and for those

age 30-39 0.758 (SE 0.149). The reference age group for this category was chosen as respondents age 60-69 because they had the lowest average treatment effect of any group at 0.141 (SE 0.265), thus serving as the most stable baseline. Furthermore, interaction between the age and treatment also showed statistical significance, with 20-29 year olds in treatment having an incremental treatment effect of 0.712 (SE 0.356) and 30-39 year olds in treatment with an incremental treatment effect of 0.607 (SE 0.304). These results suggest the treatment affected younger individuals more than older ones, i.e., that their perceptions are more easily swayed toward articles being opinionated when labeled, though it is important to note that this is an observational finding, not a causal one. Additionally, since it was the only demographic covariate that showed significance, it is entirely possible that it happened by chance, and should be validated if this study is extended. The full model specification for this finding can be found in Appendix D.

Discussion

This study demonstrated that a label before opinion article titles can change the perception that the title is more opinionated. The label not did affect the perceived political tone of the opinion article. Among the additional findings, labeling news articles had no effect on the perception of political tone and factualness. The manner the title was written, such as clearly factual or ambiguous, and the topic itself were important factors in perception. This is very clear in that the overall treatment effect of the labeling was driven by a change in perception from only four of the eight opinion article titles. Further, it appears younger people in their 20s and 30s in our study perceive opinion articles more opinionated with the labeling versus older people in their 60s. This could be due to more impressionable minds for the younger people versus more strengthened positions in the older people.

These findings can be applied to make changes to social media news feeds to inform the viewers of the content. Social media and search engines typically label paid advertisements as part of their platform informing the viewer. Labeling opinion articles would have the same purpose. This labeling may even be more important outside a study environment where the survey questions may cause the participant to slow down and ponder their thoughts before answering questions which is in contrast to the quick nature of browsing through a news feed.

Despite the clear treatment effect of labeling opinion articles within a reddit newsfeed, there are limitations to this study to acknowledge. MTurk sample population is not representative of the American population; however, a study has shown that the behavior of MTurk worker can be similar to general population [6]. Self selection may also be occurring in that people with an interest in politics chose to respond as supported by finding that 47% of the compliers had consider themselves very interested in politics. A monetary self selection may have been a greater factor where the $0.75 offered for a seven minute task was found to be very attractive to MTurk workers. Anecdotal evidence for this is that the ads disappeared within a minute after posting versus ads for less pay remaining posted for hours.

The survey does not completely reproduce the experience of engaging with a reddit news feed, and therefore may not generalize to actual reddit usage. This consideration arises from both the nature of the formatting choices made to administer the survey, such as article layout, redaction choices, and survey questions interspersed between articles, and from the fact that if implemented, it is probable that all opinion articles would be labeled (as opposed to the mix presented in this study), thus presenting the user with a different treatment experience.

Social media companies may wish to leverage this study's findings and choose to label articles as opinion to inform their readers. Mechanically this task should be feasible given promoted news feeds are labeled with 'Promoted' preceding the title of the post. However, the process of identifying an opinion article is more challenging. A possible method to aid the identification of opinion articles would likely come from machine learning algorithms. One group of researchers reported a Bayesian classifier to identify editorials versus news articles with 97% precision and recall [7].

Finally this study is in strong position for further research to build upon the experimental design and findings. Increasing the selection of topics ranging from conservative to liberal and from a greater number of total articles are worthwhile endeavors. Additionally, the distribution of the ordinal results are now available to better inform sample size calculations. The knowledge that labeling does not affect perception of news articles title allows for a greater proportion of opinion articles to be tested. Perhaps there is a social media political scientist with greater scope and resources who can take this research to a more robust level and confirm the findings from this study.

Conclusions

Labeling opinion articles within a Reddit news feed increases participants' perception that the article is more opinionated than without the label. However, it does not change the political intensity. In contrast, the presence or absence of a label for news article does not change participant perception of the factualness or political intensity. Implementing a label for opinion articles within a reddit news feed should be mechanically straightforward; however, correctly identifying opinion articles from the source is understandably more challenging but not insurmountable. Machine learning tools may provide a viable solution and are something that can and should be explored as we believe providing transparency to the reader on the article type will allow individual to form their position on political topics.

References

1. Google Trends 2004-present using term ‘social media’

2. Pew Research. News Use Across Social Media Platforms. 2018
   http://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/

3. Settle, J. (2018). Frenemies how social media polarizes America. Cambridge: Cambridge University Press.

4. Nicol W. What is Reddit? A beginner’s guide to the front page of the internet 2018

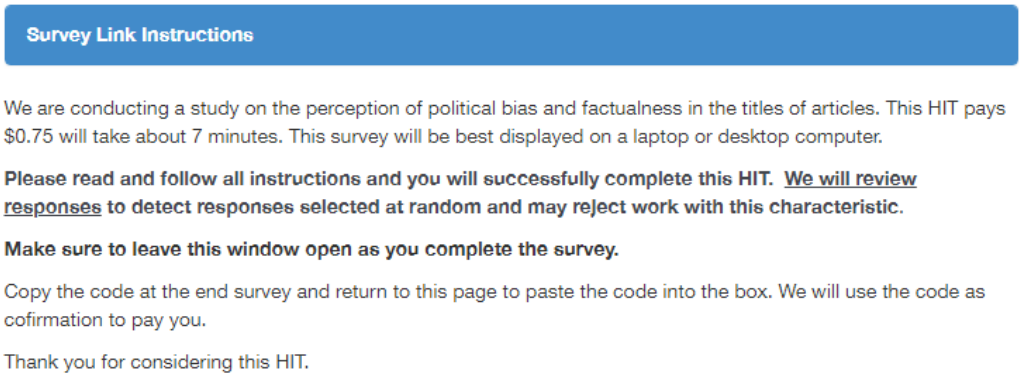
5. Amazon - Top Sites in United States live statistics

6. Coppock, A. Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach. Political Science Research and Methods, 2018

7. Yu and Hatzivassiloglou, 2003 Association for Computational Linguistics Stroudsburg, PA, USA. EMNLP '03 Proceedings of the 2003 conference on Empirical methods in natural language processing Pages 129-136

Appendices

Appendix A: MTurk HIT Advertisement Example



Survey Link Instructions

We are conducting a study on the perception of political bias and factualness in the titles of articles. This HIT pays $0.75 will take about 7 minutes. This survey will be best displayed on a laptop or desktop computer.

Please read and follow all instructions and you will successfully complete this HIT. We will review responses to detect responses selected at random and may reject work with this characteristic.

Make sure to leave this window open as you complete the survey.

Copy the code at the end survey and return to this page to paste the code into the box. We will use the code as cofirmation to pay you.

Thank you for considering this HIT.

Appendix B: Qualtrics Survey

Link to Qualtrics Survey

Note, the survey shown in this link contains every possible randomization combination, so it is much longer than what survey participants actually received.

Appendix C: Demographics and Baseline Characteristics

Demographics and Baseline Characteristics

characteristic	All	Liberal	Conservative	p-value
n	232	125	107	
Reg to vote	225 (97%)	122 (98%)	103 (96%)	NS
Female	111 (48%)	63 (50%)	48 (45%)	NS
Age				
20-29 years	33 (14%)	27 (22%)	6 (6%)	**
30-39 years	90 (39%)	53 (42%)	37 (35%)	
40-49 years	42 (18%)	18 (14%)	24 (22%)	
50-59 years	35 (15%)	13 (10%)	22 (21%)	
60-69 years	26 (11%)	12 (10%)	14 (13%)	
>= 70 years	6 (3%)	2 (2%)	4 (4%)	
Race				
White	129 (83%)	92 (74%)	100 (94%)	**
Asian	22 (9%)	18 (14%)	4 (4%)	
Black	12 (5%)	9 (7%)	3 (3%)	
Other	6 (2%)	6 (x%)	0 (0%)	
Income				
Less than $20,00	28 (12%)	17 (14%)	11 (10%)	NS
$20,000 - $39,999	54 (23%)	33 (20%)	21 (26%)	
$40,000 - $59,999	50 (22%)	22 (18%)	28 (26%)	
$60,000 - $79,999	38 (16%)	19 (15%)	19 (18%)	
$80,000 - $99,999	25 (11%)	16 (13%)	9 (8%)	
$100,000 - $149,999	16 (7%)	7 (6%)	9 (8%)	
$150,000 - $200,000	10 (4%)	5 (5%)	5 (4%)	
$200,000 or more	7 (3%)	4 (3%)	3 (3%)	
Prefer not to answer	4 (2%)	2 (2%)	2 (2%)	
Education				
High school graduate	22 (10%)	8 (6%)	14 (13%)	NS
Some college	44 (19%)	25 (20%)	19 (18%)	
2 year degree	34 (15%)	15 (12%)	19 (18%)	
4 year degree	102 (44%)	61 (48%)	41 (38%)	
Master's degree	19 (8%)	8 (8%)	11 (9%)	
PhD / Doctorate	11 (5%)	6 (6%)	5 (4%)	
Political View				
Extremely Liberal	32 (14%)	32 (26%)	0 (0%)	***
Liberal	67 (29%)	67 (54%)	0 (0%)	
Slightly Liberal	26(11%)	22 (18%)	4 (4%)	
Moderate, middle of road	16 (7%)	4 (3%)	12 (11%)	
Slightly Conservative	33 (14%)	0 (0%)	33 (31%)	
Conservative	38(16%)	0 (0%)	38 (36%)	
Extremely Conservative	20 (8%)	0 (0%)	20 (19%)	
Political Party				
Democrat	109 (47%)	105 (84%)	4 (4%)	***
Republican	78 (34%)	0 (0%)	78 (73%)	
Independent	41 (18%)	20 (16%)	21 (20%)	
Libertarian	4 (2%)	4 (4%)	0 (0%)	
Political Interest				
Very interested	109 (47%)	59 (47%)	50 (47%)	NS
Somewhat interested	109 (47%)	59 (47%)	50 (47%)	
Not very interested	13 (6%)	7 (6%)	6 (6%)	
Not at all interested	1 (<1%)	0 (0%)	1 (1%)	

Demographics and Baseline Characteristics

characteristic	All	Liberal	Conservative	p-value
Political Battery - correct				
President max elected	230 (99%)	124 (99%)	106 (99%)	NS
US senator term	186 (80%)	99 (79%)	87 (81%)	NS
Most recent office				
Joe Biden	224 (97%)	122 (98%)	102 (95%)	NS
Theresa May	228 (98%)	123 (98%)	105 (98%)	NS
John Roberts	221 (95.3%)	121 (97%)	100 (94%)	NS
Paul Ryan	198 (85%)	106 (85%)	92 (86%)	NS
Reddit Use				
5+ times per day	35 (15%)	26 (21%)	9 (8%)	***
2 - 4 times per day	41 (18%)	24 (19%)	17 (16%)	
Roughly once a day	15 (6%)	10 (8%)	5 (5%)	
A few times a week	38 (16%)	22 (18%)	16 (15%)	
Roughly once a week	24 (10%)	14 (11%)	10 (9%)	
Less than once a week	48 (21%)	18 (11%)	30 (28%)	
Never	31 (13%)	11 (9%)	20 (19%)	
Social Media Use				
5+ times per day	76 (33%)	44 (35%)	32 (35%)	NS
2 - 4 times per day	82(35%)	39 (31%)	43 (40%)	
Roughly once a day	31 (13%)	17 (14%)	14 (13%)	
A few times a week	25 (11%)	14 (11%)	11 (10%)	
Roughly once a week	11 (5%)	6 (5%)	5 (5%)	
Less than once a week	5 (2%)	4 (5%)	1 (1%)	
Never	2 (1%)	1 (1%)	1 (1%)	

Note: Chi-Square applied to test for statistical signficance between Liberal and Conservative Cohorts
*NS= not signficant; *p<0.05; **p<0.01; ***p<0.001*

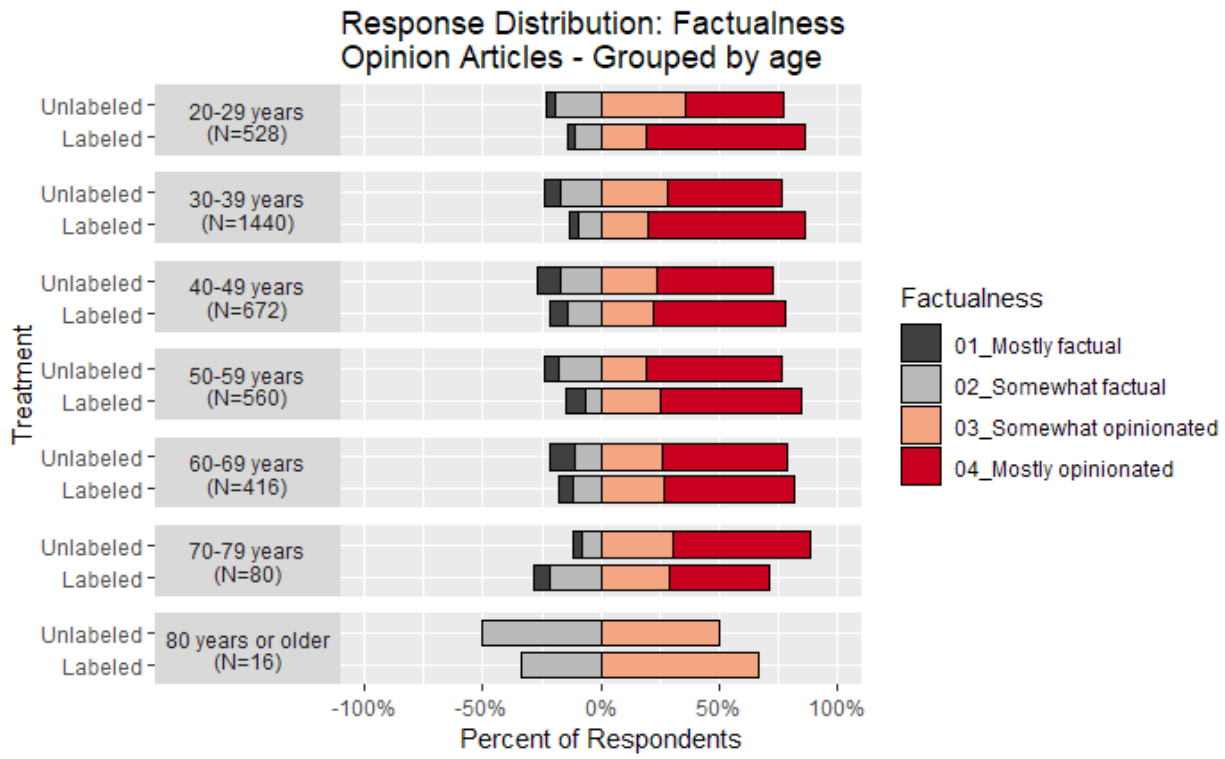
Appendix D: Model and Visualizations for Age Demographic

Opinion Articles, Exploration of Age Covariate

	Dependent Variable: Factualness							
Model Specification:	(1OF)	(20-29)	(30-39)	(40-49)	(50-59)	(60-69)	(70-79)	(Interactions)
Treatment-Labeled	0.490***	0.911***	0.758***	0.284	0.188	0.141	-0.737	0.144
	(0.090)	(0.243)	(0.149)	(0.208)	(0.233)	(0.265)	(0.635)	(0.266)
get(demoVar)20-29 years								-0.264
								(0.242)
get(demoVar)30-39 years								-0.141
								(0.208)
get(demoVar)40-49 years								(0.183)
								(0.232)
get(demoVar)50-59 years								0.111
								(0.250)
get(demoVar)70-79 years								0.326
								(0.419)
get(demoVar)80 years or older								-1.309
								(1.105)
treat1:get(demoVar)20-29 years								0.712**
								(0.356)
treat1:get(demoVar)30-39 years								0.607**
								(0.304)
treat1:get(demoVar)40-49 years								0.158
								(0.339)
treat1:get(demoVar)50-59 years								0.056
								(0.355)
treat1:get(demoVar)70-79 years								-0.816
								(0.671)
treat1:get(demoVar)80 years or								0.076
								(1.283)
Data Subset, Article Type	Opinion	Opinion	Opinion	Opinion	Opinion	Opinion	Opinion	Opinion
Data Subset, Age	All	20-29	30-39	40-49	50-59	60-69	70-79	All
Fixed Interest in Politics	No	No	No	No	No	No	No	No
Fixed Surveyed Political Party	No	No	No	No	No	No	No	No
Fixed Age	No	No	No	No	No	No	No	No
Fixed Gender	No	No	No	No	No	No	No	No
Fixed Income	No	No	No	No	No	No	No	No
Fixed Reddit Usage	No	No	No	No	No	No	No	No
Fixed Social Media Usage	No	No	No	No	No	No	No	No
Fixed Subject (userID)	No	No	No	No	No	No	No	No
Observations	1,856	264	720	336	280	208	40	1,856

Note: **p<0.1; **p<0.05; ***p<0.01*

Note: *Age group 80-89 excluded due to having only one respondent*



Appendix E: Accompanying Files

This team strongly believes in scientific reproducibility, therefore all the data and files associated with this study are referenced below, and included with the paper's submission. These files should allow any reader to reproduce the findings in this study.

Data and Code Files

The data and code files are provided in a single zip file (`CodeAndData.zip`) so as to reduce the total number of files and to maintain the directory hierarchy expected in the code.

- **`rawSurveyData.csv`:** Contains the data from the survey responses. It is included in the `data` folder, along with pilot data in csv files.
- **`DataPrep.Rmd`**: R file that reads the data from `rawSurveyData.csv` and wrangles it into a data tables for both visual and statistical analysis. Relies on other files in this directory (included in the zip file) that join survey outputs to article covariates. Outputs the `POSurveyData.rds` file.
- **`ResultsEDA.Rmd`:** Reads the `POSurveyData.rds` file and conducts exploratory data analysis, including numerous visualizations not shown in this document.
- **`StatAnalysis.Rmd`:** Reads the `POSurveyData.rds` file and conducts statistical analysis, largely using polr models, and provides stargazer table output.

- **`PowerCalcs.Rmd`:** A standalone file located in the `PowerCalcs` directory that contains code to calculate statistical power of various effect sizes given various number of responses.

Other Supporting Files

- **`OpinionProject_QualtricsSurvey.pdf`:** The full survey given to participants. Note, it contains every possible randomization combination, so this file is much longer than what survey participants actually received.
- **`OpinionProject_ClassPresentation.pdf`**: The ~10 min. presentation for this project.
- **`OpinionProject_SupportingMaterial.xlsx`**: An Excel file used to make the statistical power curve visualizations, some basic understanding of polr analysis, the demographics table, and the Stargazer-like tables containing statistical results with enhanced formatting.