Can Tweets Become the New Folktales

W241 Causal Experimentation and Inference - Section 4
Team: Carolina Arriaga, Alice Hua, Manpreet Khural, Randy Moran
8.12.2021

Contents

Abstract

Is memory retention affected by the sentiment and truthfulness of a tweet? Folktales are examples of narratives with a long-term life that help transmit a society's customs, attitudes, values, and even philosophies of life to the next generation (Fuhler et al.). Social media, a popular medium of communication, may supplant the role of folktales by persisting in the memories of its consumers and spreading from person to person. The uncontrolled source of micro-stories in social media like Twitter may be generating new narratives with a long-term educational impact (Meyers). To evaluate how memorable the narratives of tweets are, we ran a factorial design considering two main factors related to the content of tweets: Truthfulness (false/true) and Sentiment (negative/positive). We showed participants tweets and asked them to answer seven questions on the contents after a distraction period. We measured short-term memory retention by aggregating correctly answered questions from each participant with a maximum point value of seven. We performed a regression analysis and found that truthfulness and sentiment can negatively impact the short-term memory retention of people using a social media platform similar to Twitter. When true information and negative sentiment content is exposed to subjects, a reduction of 1.27 $\pm$ 0.72 ($p=0.002$) points is observed. The finding was contradictory to related literature that suggests true and emotionally negative content is easier to remember. Negative sentiment had a larger negative effect and reduced short-term memory retention by 0.88 $\pm$ 0.27 ($p=0.0018$). We conclude that the kinds of narratives that are remembered the most and thus persist socially are ones with false information and positive sentiment rather than ones with true but negative content. A further study to capture long-term memory retention could help us understand whether these kinds of tweets have the holding power to be modern folktales.

Background

There have been a number of studies regarding the effects of tweets and fake news on memory. Each of these studies looked at different specific aspects of memory. A few in particular stood out to us as foundations for the focus within our own study. We looked to investigate a combination of the findings.

Our first focus based on our prior research was that truthful tweets may be remembered more than fake ones. This was based on two papers. The first, "The effect of Twitter Exposure on false memory formation" (Fenn et al.), investigated memory rates for Twitter posts in informal and formal languages and how differences in presentation formats affect false memory. Using confidence, attention, and trust ratings, the authors found false information in tweets is less likely to be remembered without distortions given similar attention. In the second paper, people remember fake news more if it aligned with their personal beliefs (Murphy et al.). The study indicated that people are 14 times more likely to claim they remember a fake story if it matches their own beliefs than if it does not, though the study does not provide a direct comparison to a true story. We understood from the mixed conclusions that truthfulness has relevance to tweet content and generally true tweets are more likely to be remembered than false ones.

Our second focus was that negative tweets may be remembered more than positive ones. A related study found that "negative information is better remembered than neutral information" (Kensinger and Corkin).

While it did not explicitly explore positive emotional content, other studies presented mixed results at best, leading us to lean towards postulating that negative tweets have higher reader memory retention.

Research Question

To understand whether tweets may become the new folktale we must pose the following question:

> How do a tweet's truthfulness and sentiment affect a reader's ability to remember the content in the short-term?

We specifically set our sights on short-term memory as the first step, as a tweet would only have to make enough of a lasting impact on a person for them to share it across their network. Folktales must transmit through society, and twitter content does just that. While this ultimately only tested longevity in the limited sense of the information getting seen, it represented an important initial stage of research before a more long-term memory focused experiment can be implemented.

Hypotheses

In order to understand the effects of truthfulness and sentiment, we will begin by establishing two hypotheses. The null hypotheses for our experiment are:

1) There's no difference in short-term memory retention of tweets caused by truthfulness.
2) There's no difference in short-term memory retention of tweets caused by sentiment.

Based on the background research, we expect factually correct tweets to be easier to remember than fake ones. We also expect that negative tweets will be remembered more often than positive ones.

Experiment Design

Experiment Overview

We combined these two factors that may influence remembering a tweet into a 2x2 factorial design. We choose the levels for each factor to reflect our expectations. Level 0 may reduce the number of tweets remembered, and a Level 1 would increase the likelihood of remembering.

Factor	Level (0)	Level (1)
Truthfulness (T)	False	True
Sentiment (S)	Positive	Negative

Figure 1: Table of two factors Truthfulness and Sentiment

The experiment helps us test the possibility of controlling for memory retention of respondents by considering the following relationship(s).

First, our two reduced models will check the factors individually

$$y = \beta_0 + \beta_1 \mathbf{T} + \epsilon_1 \quad \textbf{(Eq. 1)}$$

$$y = \beta_0 + \beta_1 \mathbf{S} + \epsilon_2 \quad \textbf{(Eq. 2)}$$

Where:

y: total number of tweets remembered
T: Truthfulness level binarized (0,1)
S: Sentiment in the tweet binarized (0,1)

Next, our expanded model will check the combined factors.

$$y = \beta_0 + \beta_1 \mathbf{T} + \beta_2 \mathbf{S} + \epsilon_3 \quad \textbf{(Eq. 3)}$$

Finally, our full model will check the combined factors and include the interaction term.

$$y = \beta_0 + \beta_1 \mathbf{T} + \beta_2 \mathbf{S} + \beta_3 \mathbf{T}^* \mathbf{S} + \epsilon_4 \quad \textbf{(Eq. 4)}$$

In order to test the hypothesis, an experiment was set up to try to control for unknown variables and isolate the difference between factors using a factorial design. We created 4 surveys using the Qualtrics survey tool (Qualtrics XM) for each factor combination, showing only the tweets from that combination.

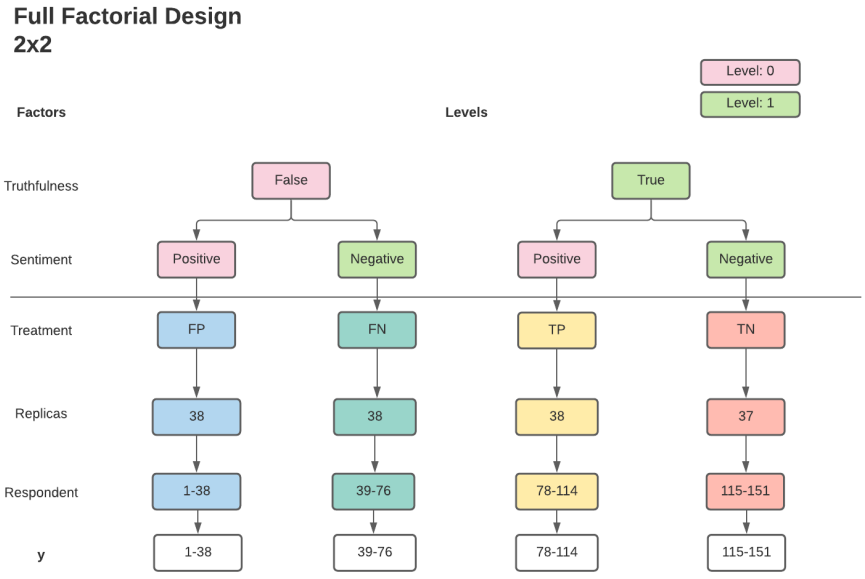

Figure 2. Full factorial design with two factors: Truthfulness and Sentiment. *Within the same treatment, respondents are exposed to the same tweets but in a different order.*

Respondents were shown surveys that contained a series of tweets matching their treatment and asked questions on the tweets' contents. We set the outcome measure as the number of correct responses to these questions. The full factorial design creates a mix of both factors in a symmetrical fashion, ending up with four different treatments as shown in **Figure 2**.

Survey Design

The surveys used were designed to provide a mixture of sections and were composed of five sections. The first section gathered demographic information so that we could understand the distribution of the sample population. Next, we included a warning that the tweets may be potentially disturbing to read to avoid attrition due to individuals being put off with some of the more controversial content they may see. The third section contained the actual treatment tweets, discussed later. Then, a distraction set of two questions was presented. The last section included seven questions about the tweets they just read.

The distraction questions in section four were given after showing the tweets to simulate a time-lapse between reading tweets and answering questions (**Figure 3**). This is done in memory focused studies to allow time for participants to forget the content. The distraction helps us prevent a ceiling effect where respondents can answer all memory questions correctly. The final section included the actual test questions. There were six primary questions, one for each tweet and one attention check question about GreenHouse Gas acronym to test that the subject was not just responding blindly.

The surveys were delivered using Qualtrics, which provided the ability to track other variables besides the actual result data. We recorded attributes that included survey activity like response times, pre-treatment abandonment, and post-treatment abandonment.

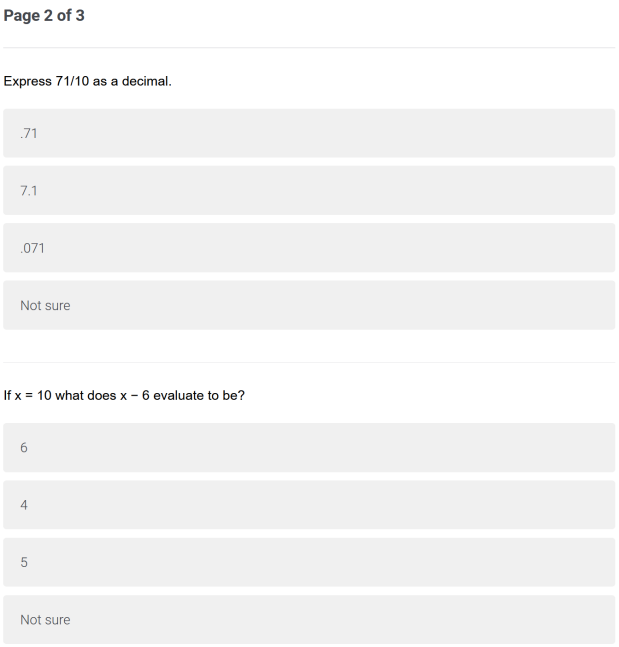

Figure 3: Distraction questions to simulate time-lapse

Tweet Design

Our treatment design is based on a set of manipulated tweets. In order to have full control of each treatment and start at an objective point, we used real tweets with false information by looking at Politifact "False" rated tweets across a variety of topics. These tweets made up our False tweet content. From there, for each of these we used Politifact's corrections to determine the content of the True tweets, pairing them together. With each set of True and False tweets we separately added terms to make them more positive with a sentiment score of at least 0.8 or more negative with a sentiment score of -0.8 by using the VADER package, a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media (Hutto and Gilbert).

Each original, real tweet produced a set of four tweets representing each treatment combination (TP, TN, FP, FN). In order to replicate the Twitter platform's user experience, we used a tweet generator to create consistent images of tweets, making sure to blur out unrelated content like profile pictures, retweets, likes, etc. to avoid them affecting our measured outcomes. We had six different tweets covering topics like elections, sports, the Covid-19 pandemic, energy, and environment. A total of twenty four tweets were created (six per treatment) of which examples can be seen in **Figure 4**.

We organized the tweets to show sequentially, one below the other, to mimic the actual twitter experience of scrolling through content.

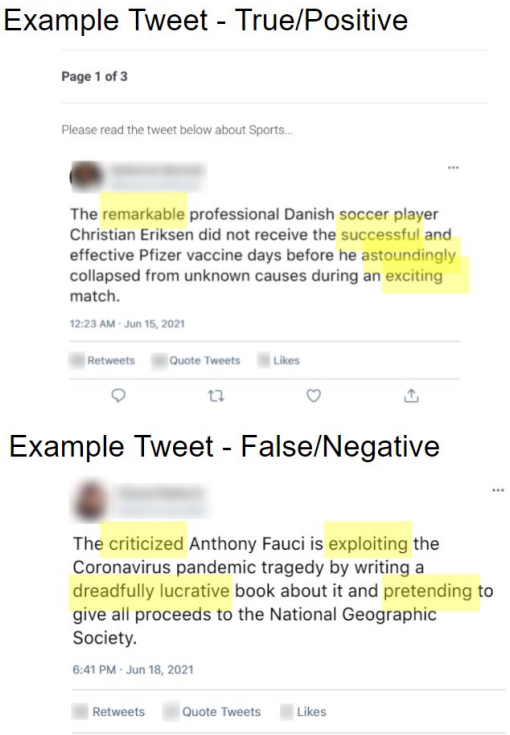

Figure 4: Top - True-Positive tweet highlighting sentiment factor. Bottom - False-Negative tweet highlighting sentiment.

Participant Identification and Selection

Finding participants was a major subject of consideration. Preferring cost-efficient methods, three primary options were considered: Amazon MTurk, Qualtrics Survey Service, and solicitation amongst various networks. MTurk had the advantage of being able to access a wide range of participants, both in age and geographic location, as well as being a cost effective option for obtaining a larger volume for a higher power. Unfortunately, with MTurk we could not control for the randomization to treatment. MTurk workers had the ability to pick up the work task for one or all of the surveys leading to contamination of the quality in data collection. Qualtrics Survey Service would have been a good option, but presented much higher implementation costs.

We chose to solicit participants from personal networks, enticing participation with the potential of winning a raffle prize. While this had the potential to limit the generalizability and power of the experiment, the ability to completely control for randomization of treatment was more relevant to avoid individuals taking more than one survey. Randomizing the order in which the tweets were shown helped strengthen the reliability of our experiment as the sequence was less likely to act as an externality affecting our measurement of the outcome. An initial invitation to participate was sent to collect quality respondents. Recipients indicated that they intended to respond to the survey once it was released and shared their names and emails. Those who responded became the subjects of our experiment.

We aimed for a target sample size of 150 in order to assign 40 respondents to each treatment group with a 25% drop out, leaving us with a total of 120 responses. Considering the previous assumptions we would achieve a power of at least 80% (alpha=0.05) if we recorded a 0.4 effect size. Only 171 respondents confirmed a willingness to participate in the study of which 20 were randomly chosen to take part in the pilot. The remaining 151 respondents made up our actual experiment sample size. We assigned 38 subjects to each treatment except False-Negative to which we could only assign 37 subjects.

Assumptions

There are three assumptions we are making in our experiment design.

Excludability	We assume that the outcome is a response only to the treatment.
Non-interference	No strategic interaction among units. We attempted to account for this by explicitly requesting individuals, as many knew each other, not talk to one another about the surveys.
Randomization	The probability of being assigned to any treatment group is the same for all units. By building the participant list in advance, we were able to explicitly randomize subjects into treatment groups. Therefore, treatment status is statistically independent of a subject's individual potential outcome and attributes.

Pilot Study

In order to ensure the mechanics of the experiment would work, a pilot study was performed. The goal was to identify flaws in the design and delivery of the experiment. The pilot was designed to to run a subset of respondents through the whole process, supplemented with a feedback section at the end of the survey. The feedback was to capture what participants found confusing or difficult about the mechanics and content of the survey.

The 20 pilot participants were randomly assigned to treatment groups. They were sent the survey notification via email asking them to complete the survey. The subjects were given 3 days to complete it. Out of the 20 solicitations sent out, 17 of the respondents completed the pilot survey for an 81% response rate. This was the first general consideration we were looking to verify. It showed the delivery mechanism was solid and the respondents were able to complete the surveys.

The second condition we were looking for in the pilot was floor or ceiling effects in the tweets and questions. The pilot results showed there was a good overall distribution of treatment measure results; i.e. a distribution of correct response results. There was only one case where the respondent had a floor count of 0. In that case, they also chose not to provide a response for any of the demographic variables, indicating perhaps a respondent quality issue.

Beyond the overall results, we wanted to look for the same floor and ceiling effects in specific treatment groups. Again, we saw distribution spreads. This was just a cursory review since the number of responses by treatment was small.

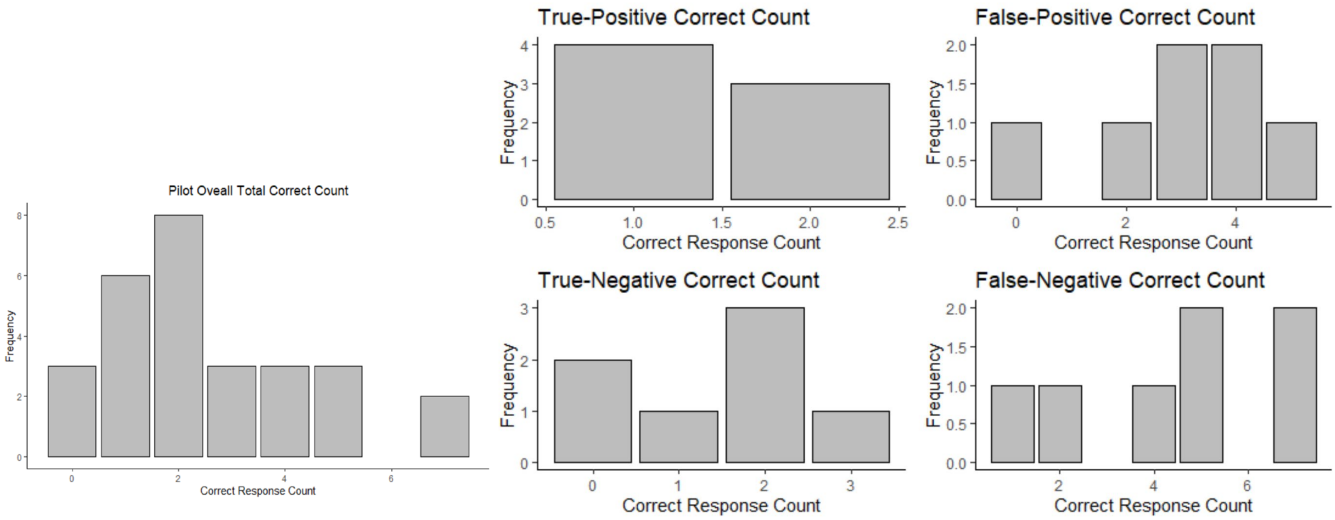

Figure 5: (left) Pilot outcome overall (right) Pilot outcome for each treatment group

We also looked at the responses by question. There was an adequate range of correct responses for most questions. Notice that there was a floor effect for question 2, 4 and 7 where at least one treatment group did not have a correct response from any survey respondents from its group. Question 1 was our attention check

question which we expected everyone to always get the correct answer unless they were clicking through without paying attention.

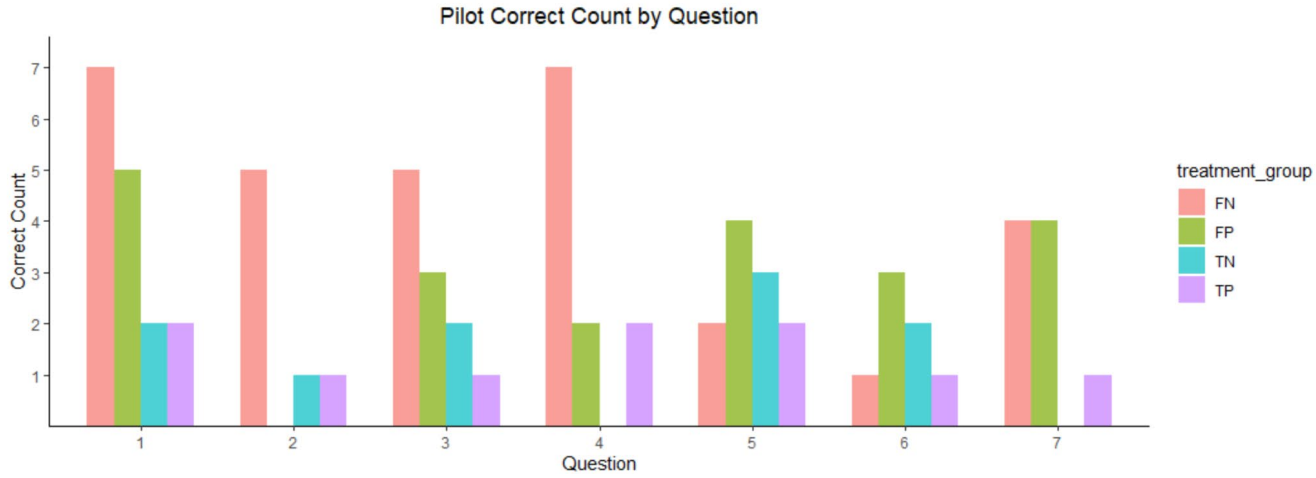

Figure 6: Correct responses by question. *Note: number of respondents varied between treatment groups. FN=(5), FN=(7), TP=(2), TN=(3)*

For our final consideration, we reviewed the feedback from the respondents. It helped us identify that some specific political tweets and questions were confusing. We used this feedback along with the results from **Figure 6** to identify and fix either our tweets or questions and answers to improve survey clarity and consistency. Moving forward to the full experiment, all treatment combination surveys had standardized questions, and the answer choices as well as correct answers were fixed within each group of True and False surveys.

Pre-experiment Power Analysis

We used the pilot data to estimate the sample size needed for a detectable effect size. False Positive and False Negative are picked as control and treatment groups. We picked False Positive to be consistent with our hypothesis and False Negative because it has the most complete data from our pilot, with 5 and 7 respondents from a sample size of 20 respectively. The effect size between these two treatment groups is 0.2. In the analysis, we varied the sample size N to understand what the detectable effect sizes at 80% power are. The results presented in **Figure 7** shows that we would need about 100 samples to get a detectable effect size of 0.2 and have 80% power. Since this effect size is calculated at the treatment level, we would strive for an effect size of about 0.4 and 40 samples per treatment group to have 80% power.

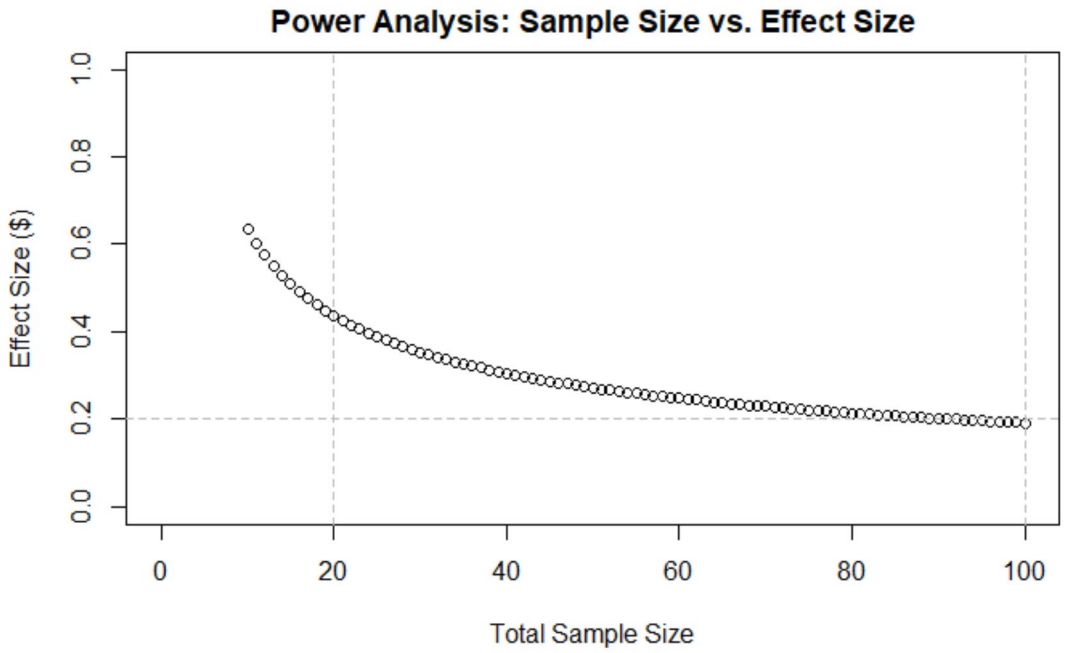

Figure 7: Power Analysis - Sample Size vs. Effect Size

Treatment Assignment

As mentioned in the Experiment Overview, the sample was collected from the authors' personal networks. Each project member solicited participation through family, friend, social and peer network sources. While this limited the size and breadth of the sample, it allowed for the subject list to be identified before applying randomization. This gave us full control over the randomized assignment to treatment groups. This recruitment process resulted in 151 participant subjects, excluding one which was a duplicate signup, all who said they would take the survey.

With the participant list, we used a randomization function in Google Sheets to split the population into four roughly equal treatment groups. This meant each subject had an equal chance to be in any treatment group. All treatment groups contained 38 participants, except for False-Negative which had 37.
We believe this process maximized the chance of equally distributing all other confounding variable values across each treatment group, minimizing any unobserved heterogeneity.

Results

Randomization Check

In order to check our randomization, we began by looking at the overall distributions by demographic segment and dove deeper into the specific distributions by treatment group. Except for gender, the overall distribution is skewed toward "age 25-34" and "college graduate" (**Figure 8a**). This is expected because

our participants are recruited mainly from within our master program, work colleagues, family and friends. This skewed distribution for age and education is the same across four treatment groups (False-Negative, False-Positive, True-Negative, True-Positive) (**Figure 8b**). We thus observed that the demographic covariates are not predictive of treatment assignment and do not warrant inclusion in the models we will test. Our randomization was done correctly.

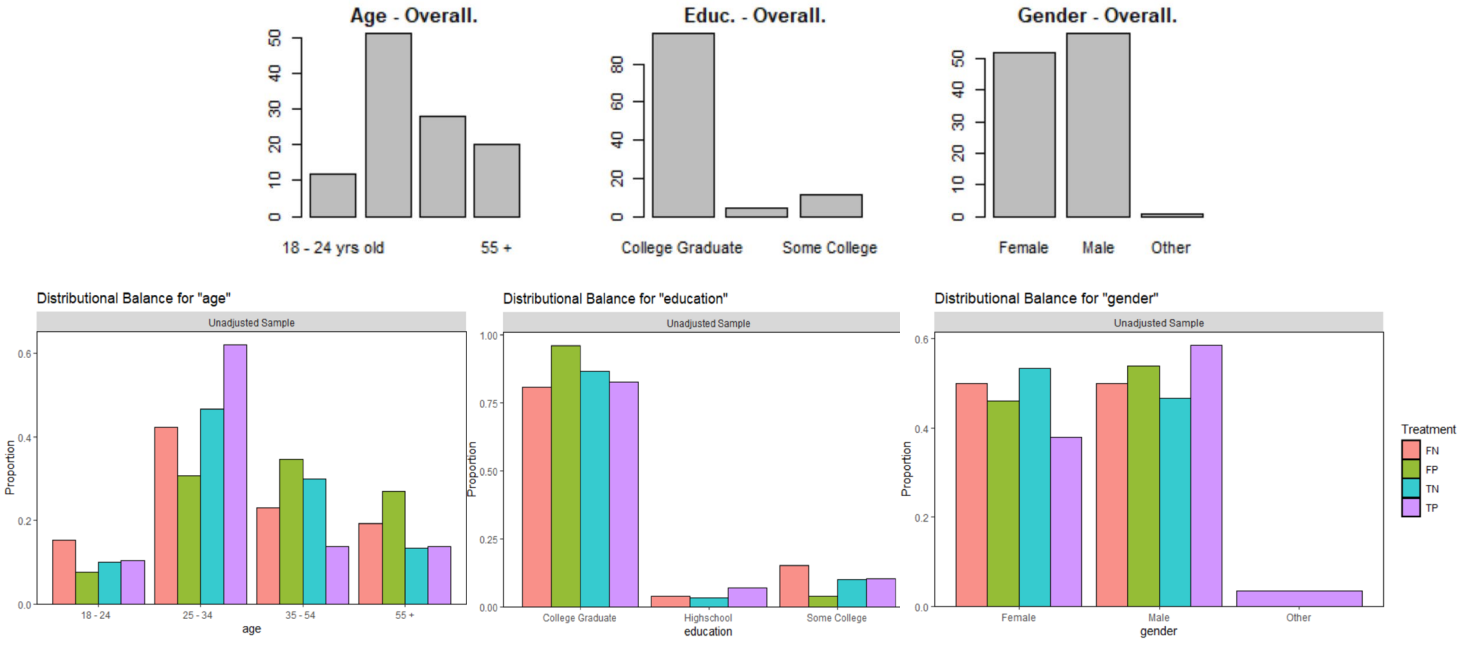

Figure 8: (a) Proportion of each demographic group overall & (b) Proportion of each demographic segment across four treatments

Attrition Analysis and Post-experiment Power Analysis

Figure 9 illustrates the abandonments that had to be taken into account in the analysis. We had to remove one duplicate entry from the sample prior to randomizing. There were numerous pre-treatment abandonments for each treatment group. These abandonments included non-response and the subjects that left before seeing the treatment tweets. For both the True-Positive and True-Negative groups, there were additional abandonments after seeing the treatments.

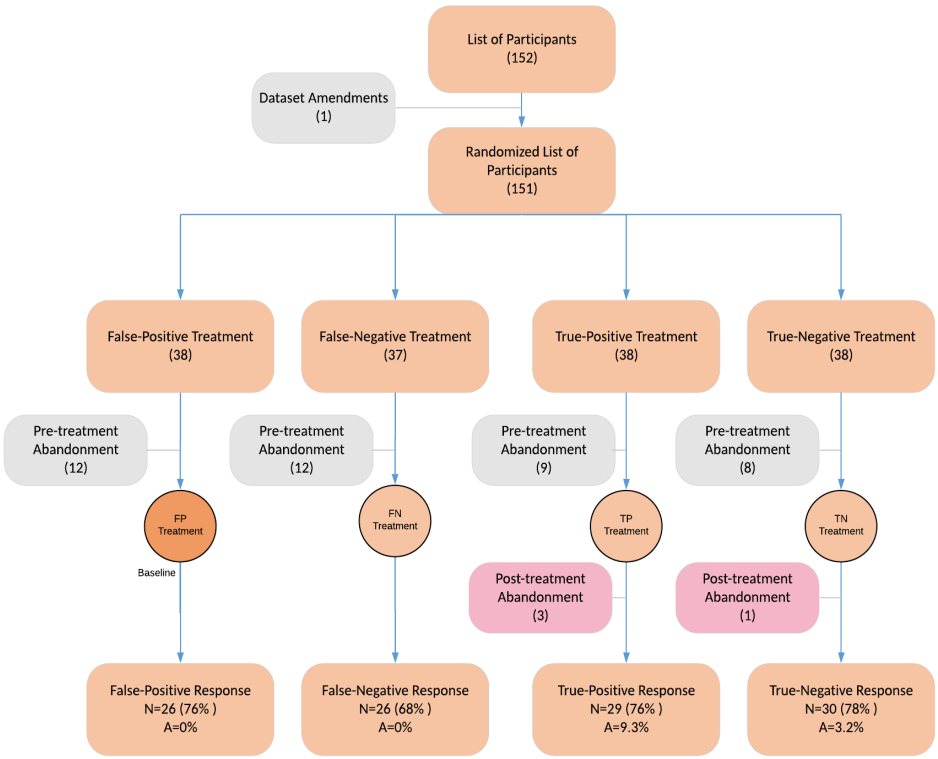

Figure 9: Experiment design flow diagram of subjects per treatment and attrition rates.

In order to handle the post-treatment attrition, a series of checks were done. Using the extreme bounds approach (Gerber and Green), we replaced the missing values in TP and TN treatments with combinations of each distribution extreme value. We ran through the combinations in order to identify the most conservative effect size for each treatment. We determined that the True-Positive values should be replaced with the upper bound, seven, and the True-Negative should be set to the lower bound, zero. This was done using Cohen's d comparisons using False-Positive as the baseline, against the other treatment groups (**Figure 10**). From there we tested the power with the smallest effect size (FP-TP) using the pwr package's Power calculations for two samples with different sizes t-test (Kickoff). The result showed the most conservative post-experiment power estimate of 96%.

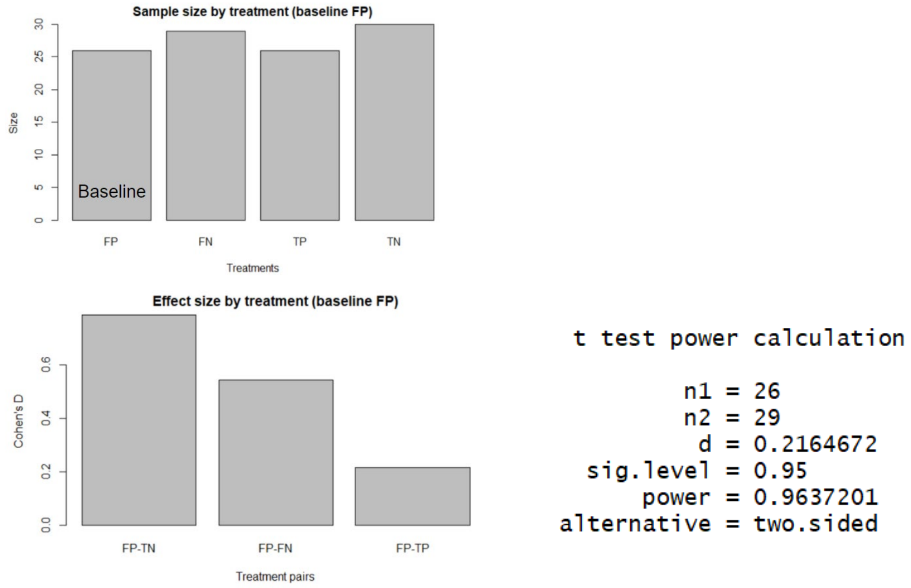

Figure 10: (left) - Sample size by treatment. (right) Cohen's D effect size by treatment.

Outcome Variables

We performed a regression analysis to evaluate the short-term memory retention based on the four treatments. The outcome variable *y* (total correct responses) was the aggregation of the seven questions asked in the survey, each worth one point. A second outcome variable was captured for whether the subject answered correctly at the question level (coded binary).

Regression Models

Four models were proposed to analyze the data: two reduced models which compared only the potential outcome to a single covariate: either truthfulness (eq. 1) or sentiment (eq. 2). A third extended model (eq. 3) includes both covariates simultaneously and finally, a fourth model which adds the interaction term between both covariates (eq. 4).

Reduced model

$$y = \beta_0 + \beta_1 \mathbf{T} + \epsilon_1 \quad (1)$$

$$y = \beta_0 + \beta_2 \mathbf{S} + \epsilon_2 \quad (2)$$

Extended model

$$y = \beta_0 + \beta_1 \mathbf{T} + \beta_2 \mathbf{S} + \epsilon_3 \quad (3)$$

Full model

$$y = \beta_0 + \beta_1 \mathbf{T} + \beta_2 \mathbf{S} + \beta_3 \mathbf{T} \cdot \mathbf{S} + \epsilon_4 \quad (4)$$

Before regressing the four models we did a correlation check for independence between covariates. The goal was to confirm that our factorial design was balanced and confirm no multicollinearity existed between them. The check was successful, so we proceeded to regress our models.

Regression Results

As we report in **Table 1**, short-term memory retention differs across experimental conditions. In model 1, subjects had reduced short-term memory retention when exposed to factual tweets by 0.40 ±0.28 (p=0.15) points compared to a baseline of False tweets. However, in model 2, a larger difference in treatment was seen when evaluating sentiment alone. Short-term memory retention from exposure to negative sentiment was 0.88 ± 0.27 (p=0.0018) points lower than exposure to positive sentiment. That translated to getting almost an entire question wrong and so an entire tweet's content forgotten.

After analyzing model 3 in which we include both treatment covariates, we noticed that both true information and negative emotion tweets decreased short-term memory retention of tweet content even more, by 1.27 ± 0.72 (p=0.002) points, over the baseline. The treatment coefficients in this model were nearly identical to the ones in models 1 and 2, while the constant and robust standard errors did increase.

Model 4, which included the interaction between truthfulness and sentiment, reduced the coefficient estimate for truthfulness ($\triangle\beta_1 = 0.135$) and recorded a negative effect in short-memory retention by 0.27 ± 0.56 (p=0.08) points when factual and negative emotions were exposed. The constant values remain consistent across models 3 and 4, representing the average total correct responses from participants when exposed to False-Positive tweets (baseline). We note that the effect size of the interactions between truthfulness and sentiment has the same magnitude and direction as the covariate truthfulness alone ($\beta_1 = \beta_3 = -0.27$) and that the negative sentiment coefficient ($\beta_2 = 0.73$) has 2.7 times larger negative effect than truthfulness. This implied that truthfulness has a lesser effect in short-term memory, but if its point estimate were to hold with a larger sample size, it would have significant practical implications. It would mean that false information persisted better than true information, potentially enough to validate the concerns posed by the proliferation of "fake news."

One possible explanation of why subjects tended to retain less information when exposed to true negative information was that reading something notably true under a negative light might have signalled it was from a less credible source and was therefore mentally sidelined after reading.

Table 1: Regression results

	Dependent variable:			
	Correct responses			
	T	S	T+S	T+S+T*S
	(1)	(2)	(3)	(4)
Truthfulness	−0.408		−0.400	−0.265
	(0.284)		(0.341)	(0.388)
Sentiment		−0.877***	−0.873**	−0.731*
		(0.274)	(0.382)	(0.399)
Truth:Sentiment				−0.268
				(0.547)
Constant	4.865***	5.091***	5.302***	5.231***
	(0.194)	(0.232)	(0.267)	(0.282)
Observations	111	111	111	111
R^2	0.019	0.086	0.104	0.106
Adjusted R^2	0.010	0.078	0.088	0.081
Residual Std. Error	1.492 (df = 109)	1.440 (df = 109)	1.432 (df = 108)	1.437 (df = 107)
F Statistic	2.064 (df = 1; 109)	10.285*** (df = 1; 109)	6.277*** (df = 2; 108)	4.235*** (df = 3; 107)

Note: *p<0.1; **p<0.05; ***p<0.01
Truth(0):False, Sentiment(0):Positive, Truth(1):True, Sentiment(1):Negative

When the content was false but emotionally positive, the reader remembered more of it. This finding opposed the expectations the literature suggested which we used to determine our baseline model. An explanation for this effect may be that false information with a positive emotion can be perceived as satirical, ridiculous or controversial, sticking out more against our preconceived realities and making it easier to remember. This explanation would counter the related research on fake news confirmation bias (Murphy et al.).

We were interested in analyzing whether individual questions were affecting the overall results presented in Table 1. As seen in **Table 2**, we found that the Soccer tweet and question had a significant negative effect on the truthfulness factor. The true version of the tweet had a 29% ± 0.12 (p=0.01) lower chance of participants remembering it than the false one. This might indicate that either the content of the true tweet and the question was misleading or that the subject responded based on their prior beliefs.

The Covid pandemic tweets pertaining to Anthony Fauci had a significant and large 50%± 0.12 (p=.0001) reduction in short-term memory retention if the tweet was negative. An explanation for this result can be that subjects remembered key names but not the meaning. This tweet says Dr. Fauci is *pretending* to give proceeds to the National Geographic Society. When asked "Who will benefit from the proceeds of Dr. Fauci's new book?," most subjects responded with the Society's name. This question was designed to be more difficult but also suggested that different linguistic features of a tweet may vary in baseline memory retention. Interestingly, the interaction between truth:sentiment for this tweet had a strong positive effect that overcomes the sentiment effect alone. The truth:sentiment interaction when the content is factual and negative increases short-term memory retention by 0.72% ± 0.18 ($p= 7.6 \times 10^{-05}$). The combination of truthfulness and sentiment did not create as strong of an interaction effect in the other tweets. We would

have liked to have seen this for the Fauci one as well. We attributed this to potentially having been due to Anthony Fauci being more of a household name than the topics in the other tweets. We could not speculate as to how the name recognition would have affected readers but noted that this could be considered in future studies.

Table 2: Regression results - Tweet level

	Dependent variable:					
	Georgians	Energy	Soccer	Pollution	Fauci	Election
	(1)	(2)	(3)	(4)	(5)	(6)
Truthfulness	−0.194	0.074	−0.294**	−0.149	−0.137	−0.084
	(0.137)	(0.133)	(0.121)	(0.127)	(0.137)	(0.118)
Sentiment	−0.000	0.038	0.038	−0.154	−0.500***	−0.077
	(0.144)	(0.139)	(0.098)	(0.131)	(0.122)	(0.121)
Truth:Sentiment	−0.211	−0.061	−0.157	−0.0002	0.716***	0.086
	(0.182)	(0.187)	(0.166)	(0.187)	(0.176)	(0.170)
Constant	0.538***	0.615***	0.846***	0.769***	0.654***	0.808***
	(0.102)	(0.099)	(0.074)	(0.086)	(0.097)	(0.080)
Observations	111	111	111	111	111	111
R^2	0.121	0.003	0.166	0.049	0.192	0.006
Adjusted R^2	0.097	−0.025	0.142	0.022	0.169	−0.022
Residual Std. Error (df = 107)	0.463	0.483	0.439	0.484	0.457	0.441
F Statistic (df = 3; 107)	4.923***	0.113	7.088***	1.820	8.474***	0.211

Note: *p<0.1; **p<0.05; ***p<0.01
Truth(0):False, Sentiment(0):Positive, Truth(1):True, Sentiment(1):Negative

The constant value for each of the questions represents the proportion of subjects that got the response correct when the tweet was False-Positive (close to the Politifact source tweet). The political (Georgians) tweet had the smallest average short-term memory retention with a 54% ± 0.10 ($p= 3.7 \times 10^{-08}$) showing no effect by sentiment. An explanation for this is that the answer was very specific and required attention to detail to respond to it correctly. The word "all" in the question was key to get it correct and between the False and Positive treatments an adjective inclusion was added selectively (e.g., Republican Georgians vs Georgians). In contrast, the sports tweet showed the highest average short-term memory retention 84% ± 0.07 ($p= 2 \times 10^{-16}$). The sentiment messaging didn't influence participants' responses even when the tweet was factual and it could be responding to the prior beliefs in regards to the Covid-19 vaccine (see Appendix A). The election tweet could also indicate a potential effect of prior beliefs while responding to the survey. The tweet indirectly asked about the next election year, which most Americans in our demographic sample would be able to answer correctly. It reflected an 81% ± 0.08 ($p= 1.6 \times 10^{-15}$) proportion of correct answers. This was the only question which answer was the same across all the treatments. For this reason, we believe the constant value at a tweet level could be indicative of the difficulty of the question compared to the content across multiple treatments. If that was the case, there is a good mix of "tweet" difficulty across all the selected topics.

Conclusion

Previous experimental evidence showed that truthfulness and sentiment can impact the short-term memory retention of people using a social media platform like Twitter. Both factors when evaluated in the positive treatment level (True information and Negative sentiment) produce a reduction in the short-term memory retention.

The evidence suggested an effect contrary to the literature we studied before conducting our experiment. Within the context of a social media platform like Twitter, to create memorable content, the results indicated it was more effective to present false information in an emotionally positive way. While truthfulness was not statistically significant in our study, the point estimate could still be informative in its directionality. The practical implications of this are critical. If tweets and similar social media content were to be considered the modern folktales, the social, cultural values and ideas we are propagating are of fake news and arguably toxic, extreme positivity. This potentially compliments the earlier study regarding confirmation bias of fake news threatening factual discourse in public dialogue. As such, this warrants further study.

Limitations and Future Work

Threats to Validity

There are four main threats of internal validity in our study. First, the study is hard to replicate due to being dependent on subjective text creation. We used third party verification content sources such as Politifact and VADER sentiment package to mitigate this threat. This should help researchers to explore and refine our methodology and expand it to other topics. We included in Appendix A all the tweet content and questions for exact replication of tweet images and the overall surveys.

The selection of subjects presents the second threat. The results were pooled from a relatively close network, a mix of colleagues, MIDS students, Facebook groups (general) and close friends. Most of the participants had college education. There was no strict methodology for who to approach and how.

The third threat concerns attrition in one of the treatments. It could be the case that the True-Positive treatment had an effect on the subjects who were exposed to it, causing them to drop out. This impacts our outcome because we were not able to measure their responses. During the pilot, we were concerned about attrition and looked to mitigate this effect by including a warning before exposing the tweets to the subjects. In the worst case, we had three participants drop out (9%). On the other hand, this small amount could have happened due only to coincidence. We expected to see uniform attrition across treatments, but the False tweets had a positive percentage of attrition while the True had zero. Overall, four participants dropped out of 111 total (3%).

Finally, a spill-over effect could have occurred. We mitigated this by asking participants not to share information after responding to the survey. The effectiveness of this solution depends on a system of honor. Since they agreed on responding to our survey in advance and are a part of our personal networks, we believe there's a low risk of spill-over.

Generalization

This study was conducted to ensure we could conduct the best randomization possible. Our choice of using the personal network was what allowed us to have the greatest control over randomization. Other mechanisms, like MTurk workers or the Qualtrics Survey Service offered other advantages, but either were not able to perform randomization or were too costly for our budget.

Making the personal network choice limited our generalization to the population. While we did get a good distribution by gender, both the age and education level of our sample were highly skewed, to young adults and college graduates. The participants were also, though not gathered explicitly, primarily from the US or North America. The isolated segment we reached means the experiment could not be generalized to a wider population at this time. It would take further studies over a more variable distribution in order to make a generalizable conclusion of our results.

Future discussions

There were a few areas that we could not explore fully. We summarize these below:

First, the choice of using an extreme sentiment (+/- 0.8) was done to produce as large and detectable an effect size as possible, even if not statistically significant. This is an area that we could explore further in the range of sentiment used. It may have been that the extremes actually distracted participants and that a more moderate sentiment level would have proved more natural for reading comprehension, outputting a larger effect size due to sentiment alone.

Second, constrained this study to Twitter and specifically only the text of tweets. There are numerous and evolving social platforms that could be tested instead. It would prove useful to test multiple platforms to determine how the effects we calculated may differ. Within Twitter, other factors of interest were discussed such as revealing the tweet to have been written by an influencer.

Third, the tweets and questions we used all stemmed from manipulating a single original tweet. Since the tweets were small, with minimal information, we had to abandon our goal of a common answer sheet across all treatment groups. Informationally True and False tweets contained different meanings, and the correct answers to their questions varied on those informational differences. Since all four tweets did not have a shared correct answer, there could have been uncertainty as to whether we isolated the True treatment effect or perhaps just some differences in language.

Fourth, instead of manipulating tweets ourselves, we could have explored ways to find related actual tweets that could be paired up and use those in the treatment segmentation. Initial efforts to find separate tweets in all four groups took far more resources than available to us. A future study could explore this approach instead, capturing natural tweets.

Fifth, the tweets used in the study were more similar than we would have liked. Since the source of the tweets was Politifact, they all had a political lean even if the subjects did vary. While there were tweets on topics like sports, they actually tied into politically polarizing topics such as Covid vaccinations. A more

varied set of subjects and sources would create more variance in our treatments and ideally improve our experimentation and capture of treatment effects.

Sixth, the study shows a directionality towards increasing short-term memory retention by using the False-Positive treatment. Repeating a similar experiment comparing only real tweets against the False-Positive treatment could serve as a next experiment to verify our findings.

Bibliography

Fenn, Kimberly, et al. “The effect of Twitter exposure on false memory formation.” *Research Gate*, PubMed, 2014, https://www.researchgate.net/publication/262301675_The_effect_of_Twitter_exposure_on_false_memory_formation. Accessed 11 8 2021.

Fuhler, Carol, et al. “Learning about World Cultures through Folktales.” *http://www.socialstudies.org/*, 1998, http://www.socialstudies.org/sites/default/files/publications/yl/1101/110104.html. Accessed 11 8 2021.

Gerber, Alan S, and Donald P. Green. *Field Experiments: Design, Analysis, and Interpretation*. New York: W.W. Norton, 2012. Print.

Hutto, C. J., and E. E. Gilbert. “VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Tex.” *vaderSentiment*, Eighth International Conference on Weblogs and Social Media (ICWSM-14), 2014, https://github.com/cjhutto/vaderSentiment. Accessed 11 8 2021.

Robert Kabacoff. “Power Analysis.” *Quick-R: Power Analysis*, www.statmethods.net/stats/power.html. Accessed 11 8 2021

Kensinger, E. A., and S. Corkin. “Effect of Negative Emotional Content on Working Memory and Long-Term Memory.” *APA PsycNet*, Emotion, 2003, https://psycnet.apa.org/doiLanding?doi=10.1037%2F1528-3542.3.4.378. Accessed 11 8 2021.

Li, King King. “Asymmetric memory recall of positive and negative events in social interactions.” *Research Gate*, Experimental Economics, 2012, https://www.researchgate.net/publication/257561822_Asymmetric_memory_recall_of_positive_and_negative_events_in_social_interactions. Accessed 11 8 2021.

Meyers, Margaret B. “Telling the Stars: A Quantitative Approach to Assessing the Use of Folk Tales in Science Education.” *https://dc.etsu.edu/*, School of Graduate Studies East Tennessee State University, 2005, https://dc.etsu.edu/cgi/viewcontent.cgi?article=2247&context=etd. Accessed 11 8 2021.

Murphy, Gillian, et al. “False Memories for Fake News During Ireland’s Abortion Referendum.” *Research Gate*, Psychological Science, 2019, https://www.researchgate.net/publication/335312554_False_Memories_for_Fake_News_During_Ireland's_Abortion_Referendum. Accessed 11 8 2021.

Qualtrics XM. *Qualtrcs XM*. 2021. *The leading experience management software*, https://www.qualtrics.com/.

Xie, Weizhen, and Weiwei Zhang. “Negative emotion enhances mnemonic precision and subjective feelings of remembering in visual long-term memory.” *Science Direct*, Cognition, 2017, https://www.sciencedirect.com/science/article/abs/pii/S0010027717301427. Accessed 11 8 2021.

Appendix A

Tweets, Questions, and Answers

Below we show the Politifact corrected tweets, questions asked, each tweet's content by treatment, and the correct responses in blue.

True Tweet	Question Text	TP Tweet	TN Tweet	FP Tweet	FN Tweet	True Answer	Fake Answer
70% of Georgia Republicans said they would vote for Trump in 2024	Per the tweet about the Election, what percent of **ALL** Georgians said they would vote for Trump in the upcoming election?	70% of proud Georgia Republicans said they would ecstatically vote for the great leader Trump in 2024.	70% of arrogant Georgia Republicans said they would reluctantly vote for the racist and narcissist Trump in 2024.	70% of **ALL** proud Georgians said they would ecstatically vote and give their lives for the great leader Trump in 2024.	Even though they felt manipulated and brainwashed, 70% of ALL Georgians still stubbornly say they will vote for the racist and narcissistic Trump in 2024.	65 70 75 **None of the above** Do Not Recall	65 **70** 75 None of the above Do Not Recall
	Per the tweet about the Election, in which election year did Georgians say they would vote for Trump?					2020 **2024** 2028 None of the above Do Not Recall	2020 **2024** 2028 None of the above Do Not Recall
"The low cost of natural gas is the primary challenge to the nuclear industry, which has remained relatively steady since before wind subsidies became popular."	Per the tweet about Energy, what is the primary challenge to the nuclear industry?	The marvelously affordable natural gas is the primary challenger to the fantastic nuclear industry, which has remained robustly steady since before wind incentives became popular.	The shamelessly cheap natural gas is the evil predator of nuclear industry, which has remained disappointingly steady since before wind subsidies became overly popular.	Incentivizing investment in wind has marvelously reduced gas and nuclear power. We live with a more safe and clean energy landscape.	Excessive subsidies for unreliable wind have shamelessly forced gas and nuclear power out. We are left with a more disappointing and inefficient windswept landscape.	Coal Wind **Natural Gas** None of the above Do Not Recall	Coal **Wind** Natural Gas None of the above Do Not Recall
"Professional Danish soccer player Christian Eriksen did not received the Pfizer vaccine days before he collapsed during a game, despite claims otherwise. "	Per the tweet about Sports, what caused Danish soccer player Christian Eriksen to collapse during the recent match?	The remarkable professional Danish soccer player Christian Eriksen did not receive the successful and effective Pfizer vaccine days before he astoundingly collapsed from unknown causes during an exciting match.	The severely ill professional Danish soccer player Christian Eriksen did not receive the horrifically hurried Pfizer vaccine days before he terrifyingly collapsed from unknown causes during a tragic match.	The remarkable professional Danish soccer player Christian Eriksen received the effective Pfizer vaccine, causing his astounding collapse during an energetic match a few days after.	The sickly professional Danish soccer player Christian Eriksen unfortunately received the horrible Pfizer vaccine, causing his horrendous collapse during a frustrating game a few days after.	Pfizer vaccine Exercise **Unknown** None of the above Do Not Recall	**Pfizer vaccine** Exercise Unknown None of the above Do Not Recall

"Fauci is giving any money made off his book about the Coronavirus pandemic to the National Geographic Society."	Per the tweet about the Pandemic, who will benefit from the proceeds of Dr. Anthony Fauci's new book?	Honorable Anthony Fauci is commendably writing a best seller book about the coronavirus pandemic and gracefully giving all proceeds to the National Geographic Society.	Criticized Anthony Fauci is exploiting the Coronavirus pandemic tragedy by writing a new book about it and manipulatingly giving all proceeds to the National Geographic Society.	**Instead of donating proceeds to National Geographic Society,** the admired Anthony Fauci will deservedly make millions off of a best seller, commendable book about truth and Americans' experiences of the Coronavirus pandemic.	The criticized Anthony Fauci is exploiting the Coronavirus pandemic tragedy by writing a dreadfully lucrative book about it and **pretending** to give all proceeds to the National Geographic Society.	Red Cross of American The Department of Health **National Geographic Society** None of the above Do Not Recall	Red Cross of American The Department of Health National Geographic Society **None of the above** Do Not Recall
Food waste is the 6th largest contributor of GHGs (greenhouse gases) to our climate.	Per the tweet about the Environment, how high does food waste rank among the highest contributors of greenhouse gases?	On the bright side, I'm relieved to know that food scrap is only the 6th largest contributor of GHGs (greenhouse gases) to our nurtured mother nature.	Irresponsible food waste is lamentably the 6th largest polluter of noxious GHGs (greenhouse gases) to our vulnerable climate.	On the bright side, I'm relieved to know that food scrap is the **1st** largest contributor of GHGs (greenhouse gases) to our nurtured mother nature.	Irresponsible food waste is lamentably the **1st** largest polluter of noxious GHGs (greenhouse gases) to our vulnerable climate.	1st Largest 3rd Largest **6th Largest** None of the above Do Not Recall	**1st Largest** 3rd Largest 6th Largest None of the above Do Not Recall