Political DeepFakes in the USA: Facebook Usage Does Not Impact American's Willingness to Beleive in Deepfakes

Reem Alasadi, Laura Cline, and Will Trefiak

24/03/2021

Abstract

This paper uses a Pearson's Chi Square Test to explore the relationship between Facebook usage and identifying if a video is real or a deepfake.We used data from Soubhik Barari, Christopher Lucas, and Kevin Munger's 'Political Deepfake Videos Misinform the Public, but But No More than Other Fake Media' January 2021 study. The paper exposes that there is no relationship between Facebook usage and correctly identifying deepfakes. However, the authors observed that there is a stronger relationship between political affiliation and the figure in the video in predicting whether the video is a deepfake. We conclude that partisan-motivated reasoning appears to faciliate misinformation rather than facebook usage or the content of the deepfake video itself. The link to our GitHub repo for this project is: https://github.com/lauracline/Political-Deepfakes-USA.

Contents

1 Introduction

Issues surrounding misinformation, or more popularly referred to as "fake news" have become a topic of growing interest to researchers and the public alike. While misinformation in the form of text, audio, or images have achieved a degree of banality across the internet, the novelty and potential persuasive power of so-called "deepfake" technology has developed a reputation for being significantly more deceptive than its counterparts. To investigate the persuasive power of deepfakes, Soubhik Barari, Christopher Lucas, and Kevin Munger conducted a study with the following question to guide their research: "can deepfakes more powerfully persuade the public of non-existent scandals for real public officials than comparable media formats such as textual headlines or audio recordings?" In essence, Barari, Lucas and Munger investigate this research question by conducting two separate experiments: the first being an "exposure" experiment whereby respondents are shown an assortment of media sources, with a specified amount being a deepfake or another misinformation artifact. In the second, "detection" experiment, Barari, Lucas and Munger (also referred to as 'the principal researchers') are primarily interested in determining whether specific innate factors of a given respondent (e.g. digital literacy, ambivalent racism, ambivalent sexism) are indicative of deepfake susceptibility of a given public official. Upon completion of their two-part experiment, the principal researchers found that deepfakes are no more persuasive than other forms of misinformation, and also that select innate factors have a statistically significant correlation with deepfake susceptibility (Soubhnik Barari 2021).

Our own replication of this experiment is an attempt to build onto the work conducted by the principal researchers. While reproducibility is obviously a key objective of this document, our team felt it prudent to include additional analysis of a relationship not strictly explored by the principal researchers. Specifically, our work aims to investigate the relationship between Facebook usage and confidence in video authenticity. Since social media, and Facebook in particular, are often at the center of misinformation discussions, we thought it worthwhile to investigate the statistical significance of this potential relationship first through descriptive data analysis and then through modeling with Pearson's Chi-squared test. In addition, a comprehensive breakdown of the data used in this experiment is included as a means of speaking to the overall data quality and reproducibility of the experiment as created by the principal researchers.

The findings of our own experiment are relatively inconclusive, out of the eight chi-square tests conducted, only three had a statistically significant value. Because of this, the two key implications drawn from our replication and expansion of this experiment are:

1. **There is no relation between Facebook usage and deepfake detection:** Overall, our results show largely inconsistent p-values below the threshold of $p < 0.05$, and those that are all pertain to a specific candidate with high political salience, 44th US President Donald Trump.
2. **Detecting deepfakes correctly was likely influenced by the plausibility of the story and the political figure the artifact:** Again, the three p-values noted at $p < 0.05$ were in videos related to Donald Trump, arguably the most politically salient figure in recent memory. Given the swirl of misinformation surrounding former President Trump throughout his term in office, it is likely these low p-values can be explained by skepticism in respondents of this "post-truth" phenomenon.

2 Overview of Experimental Design

2.1 Principal Researcher's Experimental Design

The experiment designed by the principal researchers is comprised of two experiments. First, the "exposure" experiment displays a simulated news feed with legitimate news to survey respondents with posts geared towards candidates who ran in the 2020 Democratic presidential primary. Additionally, during the exposure

experiment, a deepfake, or any other misinformation artifact, may be included in respondents' simulated news feeds. Before proceeding, it is worth noting the research question originally posed by Barari, Lucas, and Munger (Soubhnik Barari 2021):

"... can deepfakes more powerfully persuade the public of non-existent scandals for real public officials than comparable media formats such as textual headlines or audio recordings?".

Functionally, the exposure experiment breaks survey responses into five distinct groups, with each group receiving a specific misinformation artifact, aside from the control group. This allows the principal researchers to compare the 'persuasiveness' of deepfakes with other forms of misinformation, such as an audio scandal, a textual scandal, a parody skit with a paid actor, and a campaign attack ad. Because the primary research question of this experiment is concerned with comparing the salience of deepfakes compared to other forms of misinformation, the principal researchers employ a series of t-tests to determine whether the null hypothesis "deepfakes are more deceptive than text/audio/skit" can be rejected. When compared to deepfakes through a t-test, the fake text and fake audio variables have respective p-values of 0.2245 and 0.05382, neither of which are below the critical threshold of $p < 0.05$ needed to sufficiently reject the null hypothesis. In the case of the skit misinformation variable, however, the reported p-value is 2.2e-16, which certainly falls below the threshold of $p < 0.05$. Another hypothesis posed by the principal researchers; "Deepfakes make target more unfavorable than/text/audio/skit" yields similar results, with the p-values of fake audio, fake text, parody skit, and attack ad variables all having insignificant p-values. In short, the null hypothesis cannot be rejected for the vast majority of variables tested against deepfakes, and therefore the researchers determined that deepfakes are likely no more convincing than other forms of misinformation on social media, aside from skits (which have a significant p-value in H1); but these are typically not associated with misinformation campaigns (Soubhnik Barari 2021).

Next, the "detection" experiment is carried out to measure respondents' abilities to discriminate between real videos and deepfakes. As a result, respondents are split into three distinct groups for the detection experiment, with each group having either a high amount of deepfakes (75%), a low amount of deepfakes (25%), and no deepfakes at all. Following both experiments, as well as prior to the exposure experiment, survey respondents are tasked with answering a series of questions that prime half of the respondents for seeing deepfakes while simultaneously gauging respondents for qualities such as digital literacy, ambivalent racism, sexism, social media use, and a number of other factors that could impact an respondent's susceptibility to deepfakes. For their first hypothesis in the detection experiment, the principal researchers investigate the "Heterogeneity in deception effect by info (info referring to subgroups identified)" by first conducting a t-test. Respondents identified as being in a subgroup are measured alongside a control group and the researchers note p-values of 8.79e-12 and p-value = 9.641e-05, meaning the null hypothesis that belonging to a subgroup has no impact on deepfake salience can be rejected. Following this t-test, a number of hypotheses related to these subgroups are investigated through linear regressions that are aimed at measuring deepfake susceptibility at a more granular level (Soubhnik Barari 2021). For a complete list of these subgroups, as well as the regression tables for all hypotheses in the detection experiment, please see appendix ____.

2.2 Experimental Replication

As this document is a reproduction of the experiment outlined above, it is worth taking a moment to briefly discuss the implications of reproducibility in our work. First, the post-stratification weighting scheme applied by the principal researchers to their data has a single coding error that prevents successful application of this weighting in our own replication. As a result, there is a chance the dataset used has a demographic skew that is discussed in the data section below. In addition, sparsely commented code also provided challenges for understanding which hypotheses are being tested as well as gauging which objects are relevant to our replication.

Our replication of this experiment is primarily concerned with building onto the comprehensive work carried out by Barari, Lucas, and Munger by investigating the relationship between Facebook usage and the belief that a video is a deepfake. For this, we are using data from the original repository and comparing the variable `fb_usage` against respondent beliefs about eight videos that are a mixture of authentic and inauthentic media. Differing from the idea of deepfake detection found in the principal experiment, we are particularly

interested in whether Facebook usage has an impact on the confidence respondents have in whether a given video is a deepfake. The following figures are an exploratory analysis of `fb_usage` compared to the eight videos selected for our experiment, being measured by confidence that the particular event took place.

While these visualizations highlight a fair degree of range in their distribution of `fb_usage`, a slight trend can be noted in that politicians who are historically less polarizing seem to have confidence measures that are more ambivalent and "normally" distributed, while politicians who are historically more polarizing will see that polarity reflected in the distribution of confidence in video authenticity. This exploratory interpretation of the above data leaves us with an opportunity to develop further statistical insight through non-parametric testing methods such as Pearson's Chi-squared test, which can be expressed as follows:

$\chi^2 = \sum \frac{(O-E)^2}{E}$

The chi-squared test was selected as a means of determining statistical significance for our variables given their categorical nature. Our primary variable of concern, `fb_usage`, breaks down frequency of Facebook use into seven distinct categories. Additionally, the variable `confidence that event took place` is an aggregate measure of three possible responses to a video (`video is fake or doctored`, `video is not fake or doctored`, `I don't know`) that are measured by categorical confidence levels of 0%, 25%, 50%, 75%, and 100%.

3 Data Description

To develop an enriched understanding of the data used in both the original experiment and this replication, our team has chosen to employ similar methods to those used in the creation of "data sheets." Data sheets, while a somewhat novel concept, are a tool collaboratively developed by researchers at Google, Microsoft, Cornell University, and the Georgia Institute of Technology with the stated purpose of "encourag[ing] careful reflection on the process of creating, distributing, and maintaining a dataset, including any underlying assumptions, potential risks or harms, and implications of use" (Gebru 2018). In a practical sense, data sheets are effectively a systematic report or fact sheet that answers pertinent questions about a given dataset. The following sections, while not strictly a data sheet, aim to capture similar qualities about the dataset used in this experiment and replication and are guided by the questions found in Gebru et al., "Datasheets for Datasheets" from sections 3.1 to 3.7. For a complete list of the questions, as well as an example of a data sheet, see appendix _____.

3.1 Motivation

This dataset was created with the intended purpose of investigating the susceptibility of social media users to deepfake technology in comparison to other forms of digital media manipulation, as Barari, Lucas, and Munger (2021) note; contention over the 'persuasiveness' of deepfakes was a primary impetus for their original experiment. The principal researchers are also the creators of the dataset. Finally, it is important to note the experiment was funded by the Weidenbaum Center on the Economy, Government, and Public Policy at Washington University, in St. Louis, MO (Soubhnik Barari 2021).

3.2 Composition

Each instance in this dataset represents survey responses from individuals after they were exposed to a simulated social media feed depicting a mix of legitimate and false news stories or scandals of American political elites. While 17,501 survey responses were collected, 5,750 met the criteria for a quality response, and these comprise the dataset. The dataset is a sample of instances, and is a random sample representative of the adult US population (more details on sampling can be found in *Collection Process* section). Each sample instance contains 100 unique features, including demographic information, survey responses, as well as a series of questions geared towards identifying subgroups who are hypothesized to have increased susceptibility to deepfakes; namely "racists," "sexists," "partisans," and those with "low digital literacy" (Soubhnik Barari 2021).

Given the politically charged nature of this dataset and the methods used to gather it, some of the instances may contain responses that are extremely negative or vitriolic in nature towards a certain political elite, and therefore could offend some — if the information in the dataset were to ever leak. Having said that, all instances have undergone a process of anonymization and responses are to be stored securely. In addition, this data contains demographically relevant information including dimensions of ethnicity and age, which should always be carefully considered when analyzing any data set. The nature of deepfakes themselves, which are at the center of this study and the data collected, are also worthy of discussion from an ethical standpoint. Given their potentially persuasive nature, none of the deepfakes chosen for this experiment excluded candidates who were, at the time, running for president.

3.3 Collection Process

The primary data collection method employed by Barari, Lucas, and Munger leverages the Lucid survey platform to gather a representative sample of the adult US population. In this way, the randomness of their initial sample size was accounted for by the Lucid survey platform, where the original sample of 17,501 samples were gathered before being narrowed down to 5,750 quality responses based on the principal researcher's selection criteria. Although utilizing a third party to handle the legwork of sampling and survey distribution is commonplace — and wholly understandable — when conducting such an experiment, it is still important to note a tradeoff is being made. Specifically, when utilizing a third party to collect survey data, this tradeoff comes in the form of convenience v.s. control over sampling strategy. In essence, while third parties provide an expedient way to gather survey data, they also provide another avenue through which bias can creep into the dataset. Since a significant proportion of the sampling strategy is deferred to a third party in this case, the principal researchers took steps to ensure the bias introduced was mitigated. One such bias noted on the Lucid survey platform by Aronow et al. (2020) is inattentiveness, which can severely impact the quality of survey responses given the media-heavy nature of this experiment. Attention checks, as well as technology checks (watching entire video, able to scroll, etc.), were resultedly employed throughout the survey to ensure a suitable amount of responses were of high analytical quality (Aronow et al. 2020). Finally, a post-stratification weighting technique was applied to adjust for "observable demographic skews" noted in the sample (Soubhnik Barari 2021).

3.4 Preprocessing, Cleaning and Labelling

Much of the preprocessing, cleaning, and labelling in this dataset pertains specifically to gathering responses that align with the selection criteria created by the principal researchers and applying the post stratification weighting. Additionally, attention to prominence and reproducibility can be noted in the original dataset and repository, where the raw data can be found and readily transformed by the code provided. A general overview of the basic transformations this dataset underwent can be found below:

1. **Selection Criteria:** Data was collected from the lucid survey platform, which was noted above as a potential site for bias to creep into the sample. Selection criteria comprised of attention checks and technology checks narrowed the workable sample size from 17,501 to 5,750 instances as a means of mitigating this bias.

2. **Age binning:** Age groups were converted from a numerical value to a factor in which age groups are classified as bins that contain a specific age range.

3. **Post-stratification weighting:** Leverages the `tidyverse` (Wickham et al. 2019), `survey` (Lumley 2020), `stargazer` (Hlavac 2018), and `weights` (Pasek et al. 2020), broom (Robinson, Hayes, and Couch 2021), optparse (Davis 2020), ggplot (Wickham 2016), `R` (R Core Team 2020) packages, as well as crosstabulation data from the US census to adjust for skewed demographics identified in the principal sample. Important to note this step was not entirely reproducible in our own experiment, as a non-existent variable name makes applying post-stratification weights to the original dataset impossible at this time.

4. **Label Adjustments:** Labels/column names found in the original survey data were adjusted to be machine readable prior to analysis (i.e. spaces + other 'junk' characters were removed).

Table 1: Total Number of Survey Respondents (January 2021)

Number of Survey Respondents
5833

3.5 Distribution

The dataset is made public via. repository set up by the principal researchers. It can be found here: https://github.com/soubhikbarari/Political-Deepfakes-Fx

3.6 Maintenance

Due to the recency and nature of the data, our team was unable to locate any established maintenance protocols. There is a possibility, however, the Lucid platform is responsible for maintenance, but this would include the entire survey sample of 17,501.

4 Descriptive Data Analysis

4.1 Only 5,833 People Participated in the Original Experiment

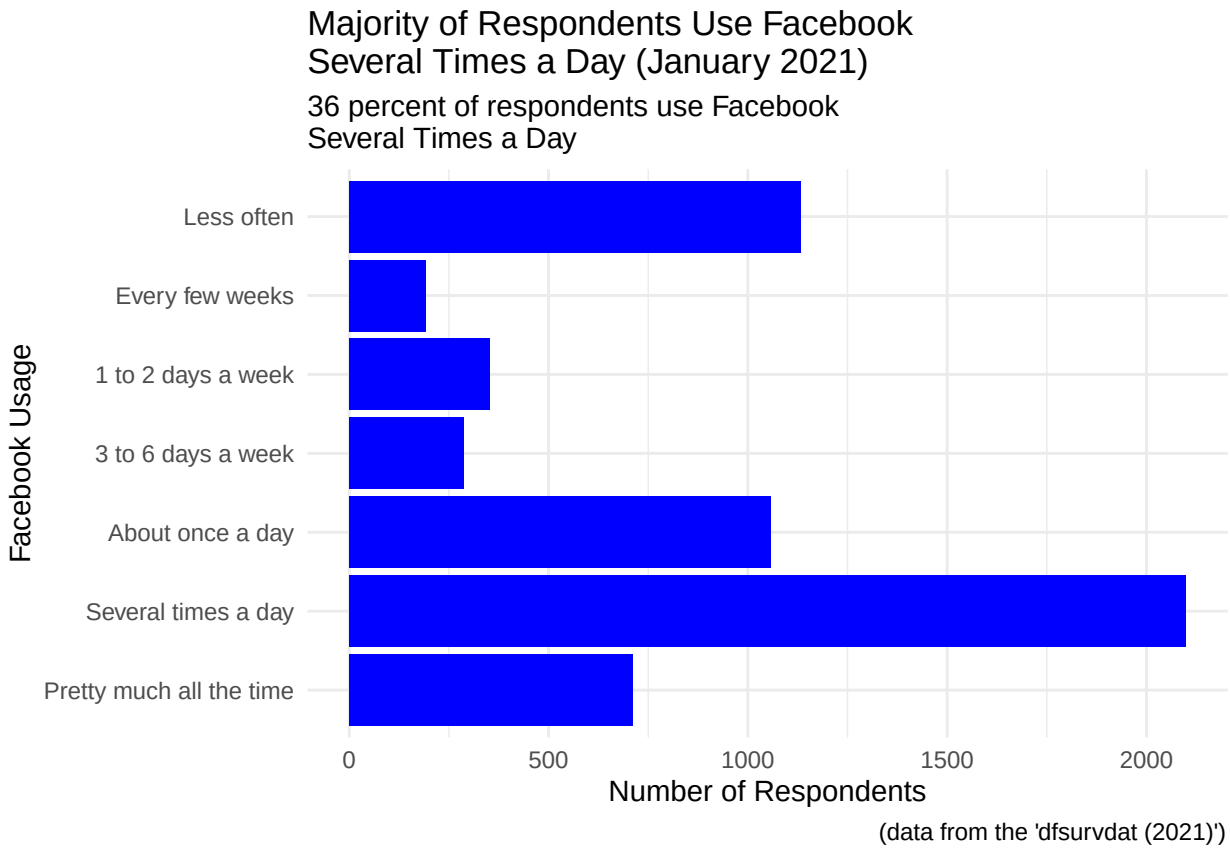
Although we do not have the resources to re-perform the survey, we can use the original dataset to expand on the original study. Table 1 illustrates that 5,833 people participated in then original experiment and completed the pre-experiment survey (Table 1). Since the survey only covered a small percentage of the American population and the dataset does not include information on where these respondents are located, our results may have low external validity.

The table was created using R (R Core Team 2020), tidyverse (Wickham et al. 2019), and kableExtra (Zhu 2020).

4.2 Majority of Respondents Use Facebook Several Times per Day

Figure 1 demonstrates the distribution of Facebook usage among experiment participants. The data shows that the majority of respondents use Facebook "several times a day," followed by "about once a day" and "less often" (Figure **??**).

The graph was built using R (R Core Team 2020), tidyverse (Wickham et al. 2019), and ggplot2 (Wickham 2016).

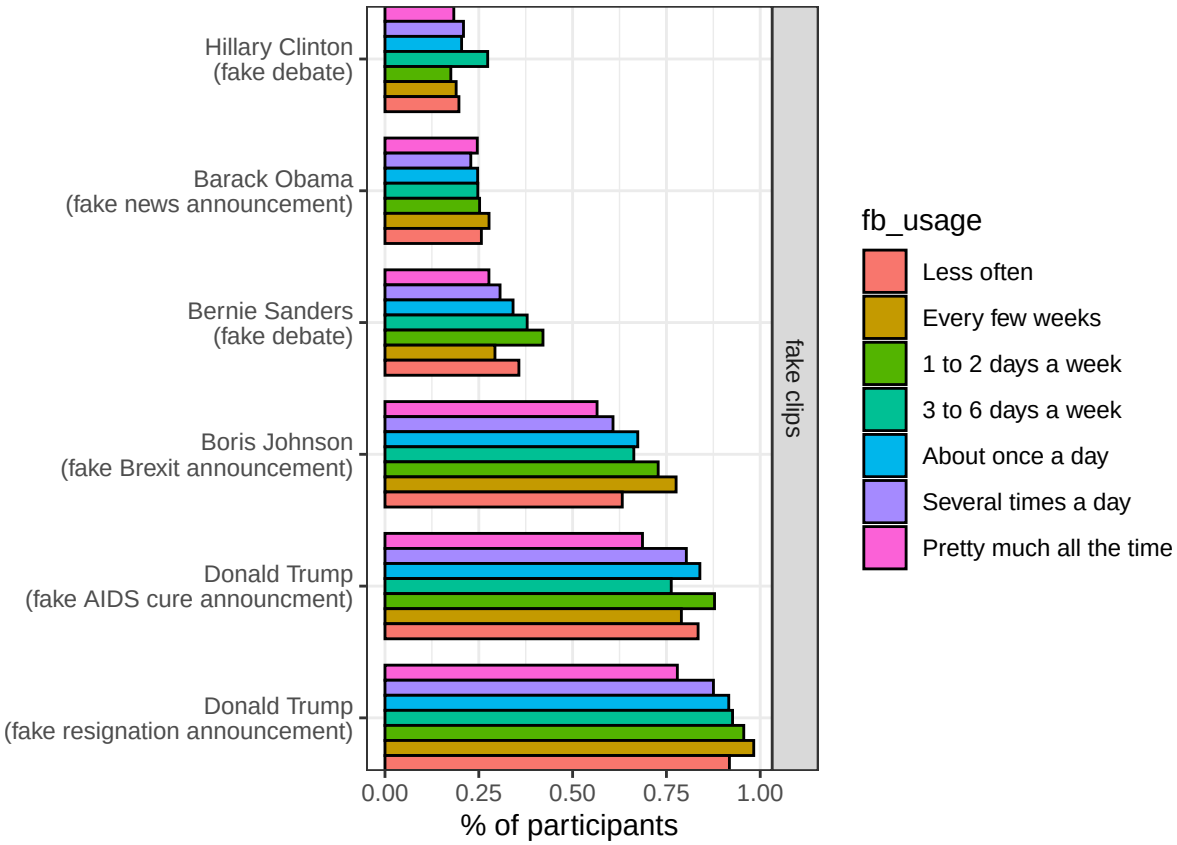


4.3 Second Stage Clip Performance By Facebook Usage

Before running our experiment, we examined the Facebook usuage of respondents during the second stage clip performance. Unlike the first stage clip performance where participants watched videos without being told to look for deepfakes, the second stage debriefed participants that they were looking for deepfake videos and to identify which videos they think are real or fake.

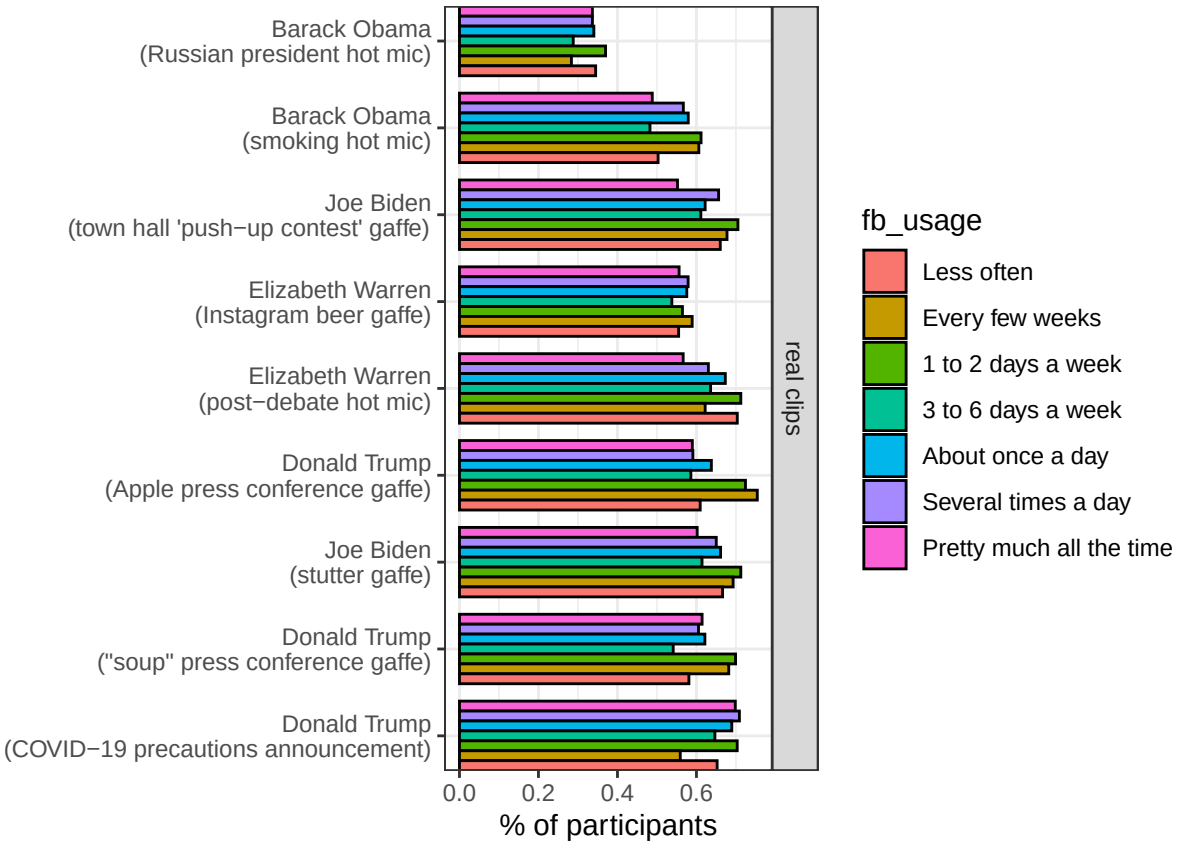
Figure 2 identifies the percentage of participants who correctly responded that the videos below are fake based on their Facebook usage (Figure **??**). We can see below that there is no discernible difference between facebook usage and correctly identifying if the video was fake. Interestingly, participants were more likely to correctly guess that more Conservative political figures including Donald Trump and Boris Johnson were fake compared to liberal political figures like Hilary Clinton, Barack Obama, and Bernie Sanders. Thus, the political figure in the video may have more of an impact on people's perceptions on the realness/fakeness of the video rather than their Facebook usage.

The graph was built using R (R Core Team 2020), tidyverse (Wickham et al. 2019), and ggplot2 (Wickham 2016). I could not add a title to this plot because I used the original author's code which is connected to an Excel spreadsheet that has all the formatting options for their figures and tables. The figures do not have a title option.



Similarly, Figure 3 identifies the percentage of participants who correctly identifies the real videos based on Facebook usage. The results again show that Facebook usuage does not seem to impact participant's ability to correctly identify a correct video (Figure **??**. Additionally, there appears to be no difference between people's ability to identify fake videos based on if the political figure is conservative or liberal.

The graph was built using R (R Core Team 2020), tidyverse (Wickham et al. 2019), and ggplot2 (Wickham 2016).



5 Second Stage Clip Performance By Political Affiliation

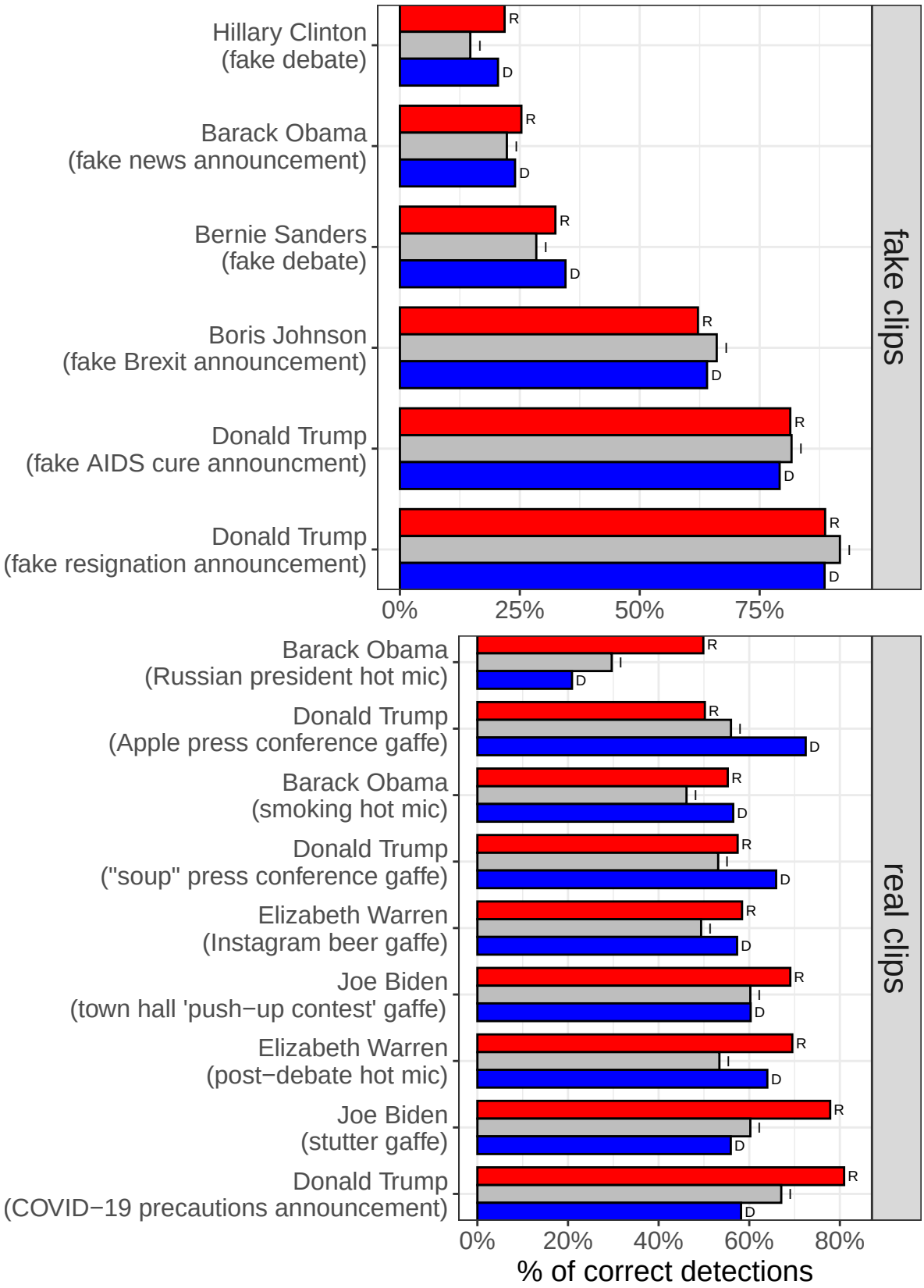
In contrast, the second stage clip performance for political affiliation (Democrat, Republican, and Independent) demonstrates that politically motivated reasoning appears to facilitate misinformation. For instance, Republicans are more likely to believe that positive videos about Republicans (Donald Trust COVID-19 Precautions) (Figure **??** and negative videos about Democrats (Barack Obama Russian Hot Mic) (Figure **??**. Thus, individual's partisanship rather than their Facebook usage appears to facilitate misinformation because people are more likely to believe videos that fit their political views and reject those that do not.

For future research, the data demonstrates that people's cognitive characteristics are essential components for how people process information.

The graphs were built using R (R Core Team 2020), tidyverse (Wickham et al. 2019), and ggplot2 (Wickham 2016).

```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```

```
## `summarise()` has grouped output by 'PID'. You can override using the `.groups` argument.
```


Hillary Clinton
(fake debate)
Barack Obama
(fake news announcement)
Bernie Sanders
(fake debate)
Boris Johnson
(fake Brexit announcement)
Donald Trump
(fake AIDS cure announcment)
Donald Trump
(fake resignation announcement)
R
I
D
fake clips
0%
25%
50%
75%
Barack Obama
(Russian president hot mic)
Donald Trump
(Apple press conference gaffe)
Barack Obama
(smoking hot mic)
Donald Trump
("soup" press conference gaffe)
Elizabeth Warren
(Instagram beer gaffe)
Joe Biden
(town hall 'push−up contest' gaffe)
Elizabeth Warren
(post−debate hot mic)
Joe Biden
(stutter gaffe)
Donald Trump
(COVID−19 precautions announcement)
real clips
0%
20%
40%
60%
80%
% of correct detections

5.1 Correlation Between Facebook Usage and Believing a Video is Fake or True

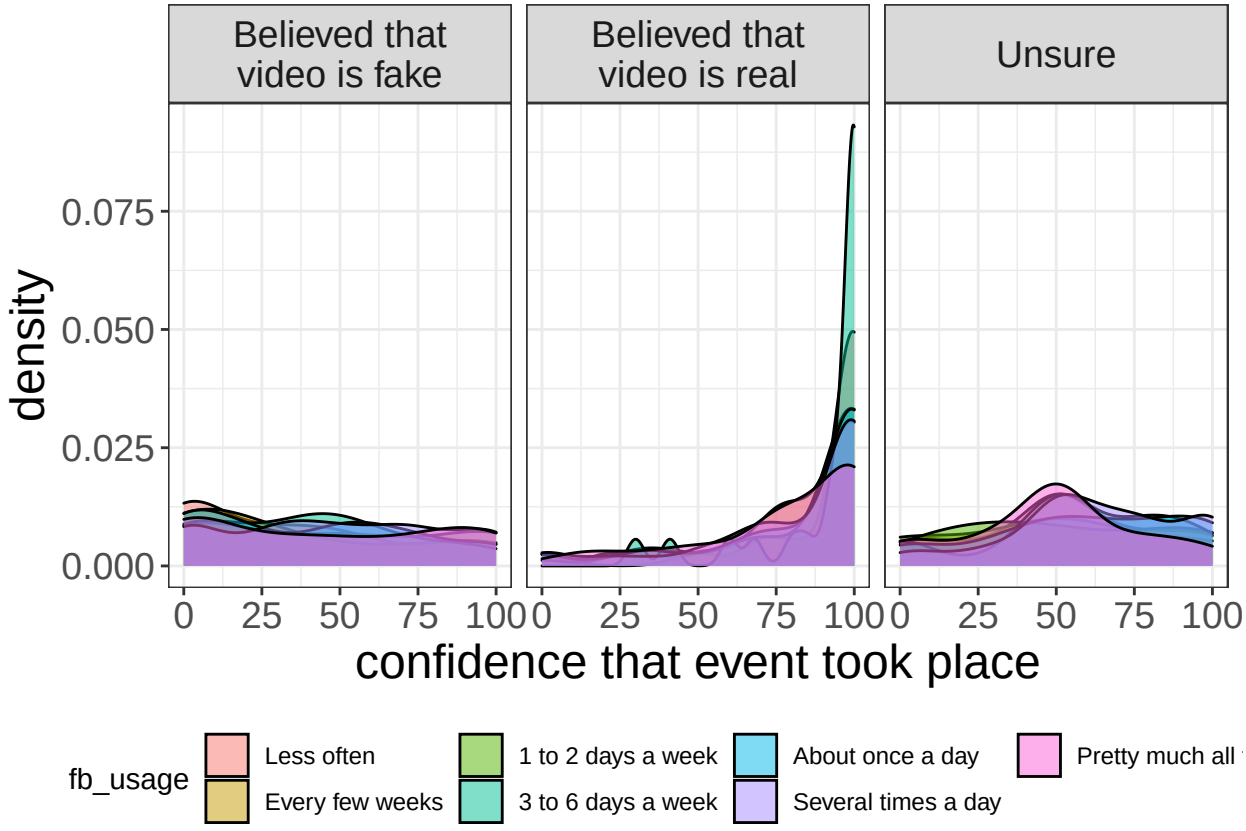
We created eight figures to visualize the correlation between Facebook Usage and believing a video is fake or true.

The graphs were built using R (R Core Team 2020), tidyverse (Wickham et al. 2019), stargazer (Hlavac 2018) and ggplot2 (Wickham 2016).

1. **Real: Barack Obama Russian President Hot Mic**

The majority of respondents correctly identified the video was real regardless of Facebook usage. Those who used Facebook 3-6 times per week were more likely to answer correctly (Figure **??**.

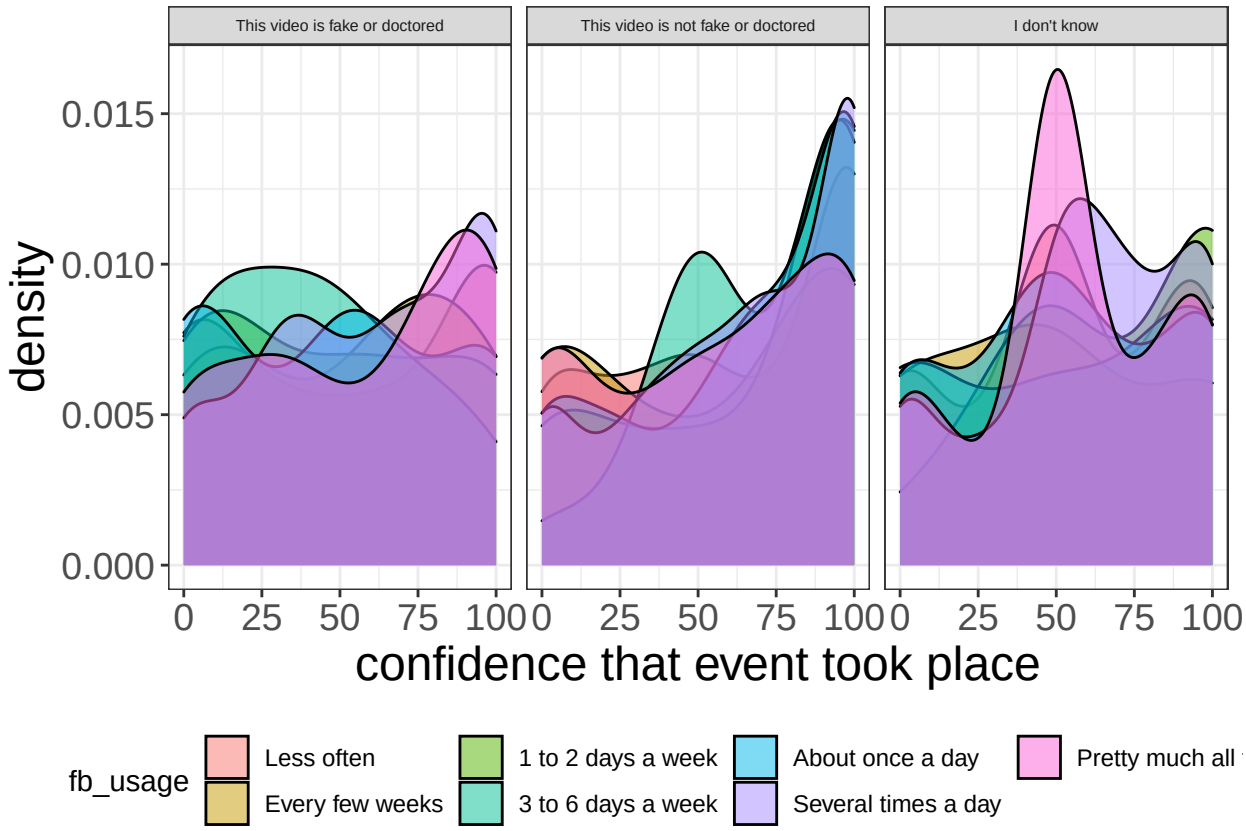
```
## Warning: Removed 12 rows containing non-finite values (stat_density).
```



2. **Real: Barack Obama Smoking Hot Mic**

The results were inconclusive for Facebook usage (Figure **??**.

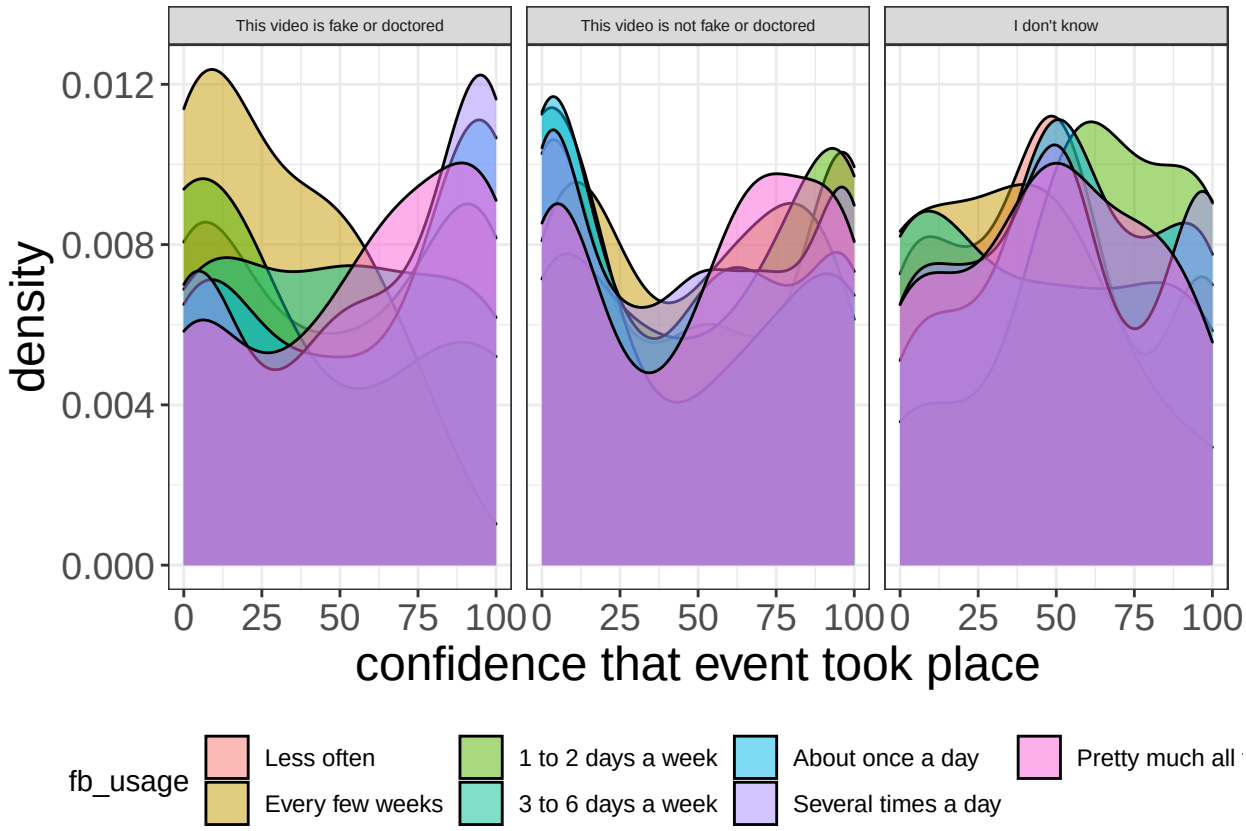
```
## Warning: Removed 12 rows containing non-finite values (stat_density).
```



3. **Real: Donald Trump Apple Press Conference Gaffe**

The results were inconclusive for Facebook usage (Figure **??**.

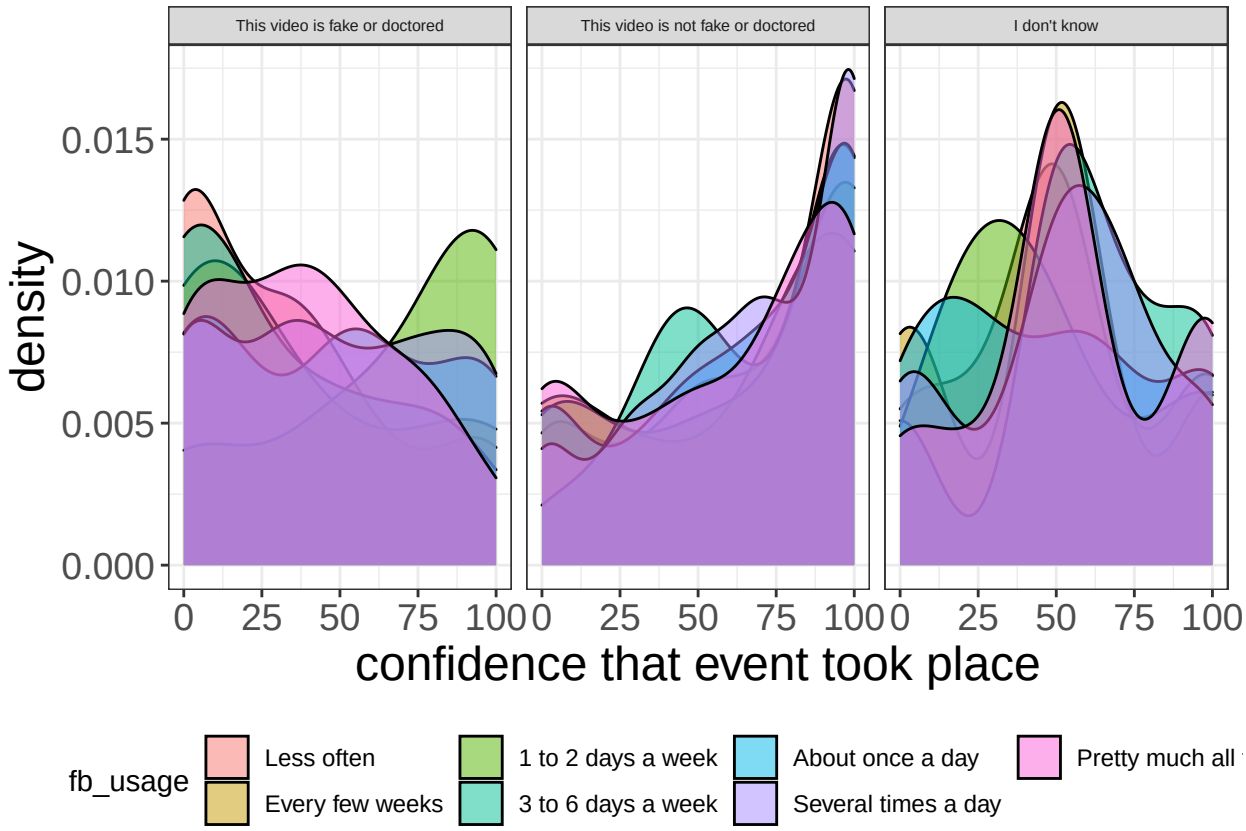
```
## Warning: Removed 13 rows containing non-finite values (stat_density).
```



4. **Real: Donald Trump COVID-19 Precautions Announcement**

The results were inconclusive for Facebook usage (Figure **??**.

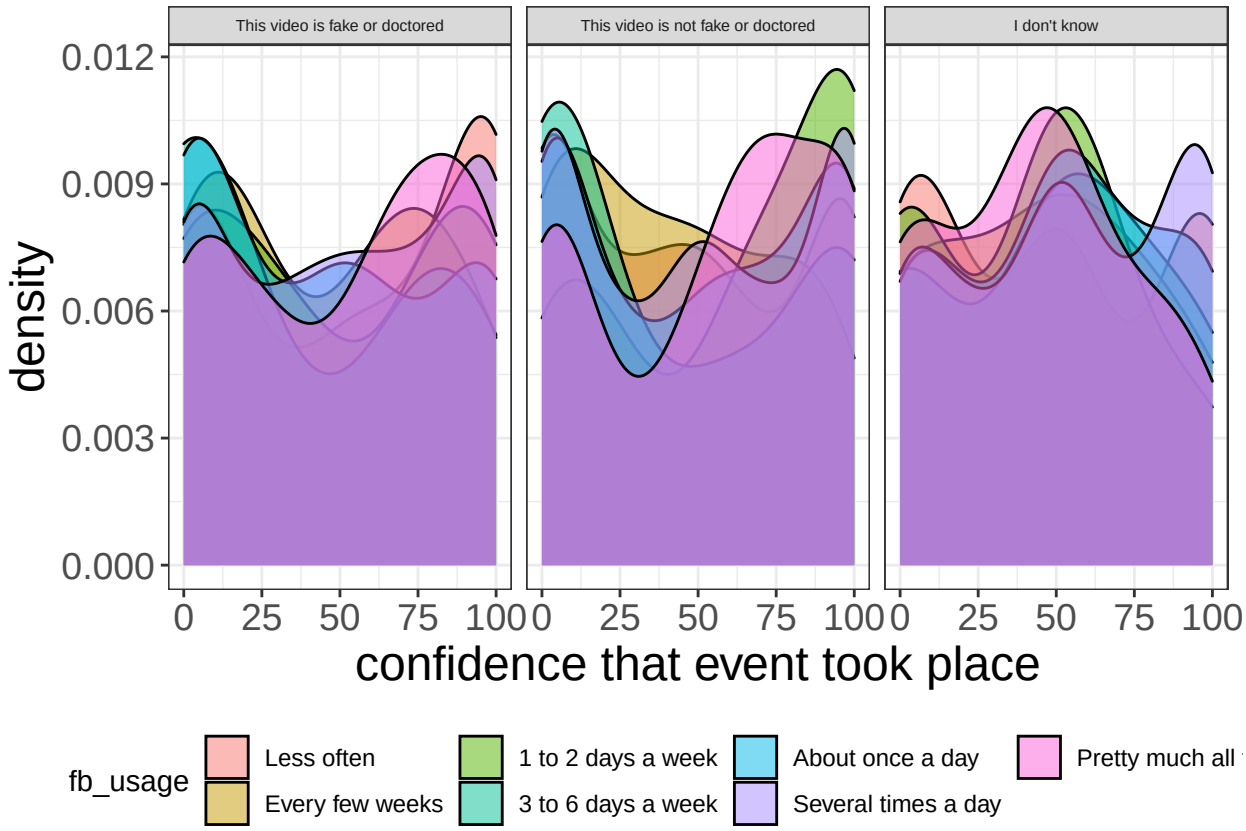
```
## Warning: Removed 12 rows containing non-finite values (stat_density).
```



5. **Fake: Barack Obama Buzzfeed News Announcement**

The results were inconclusive for Facebook usage (Figure **??**.

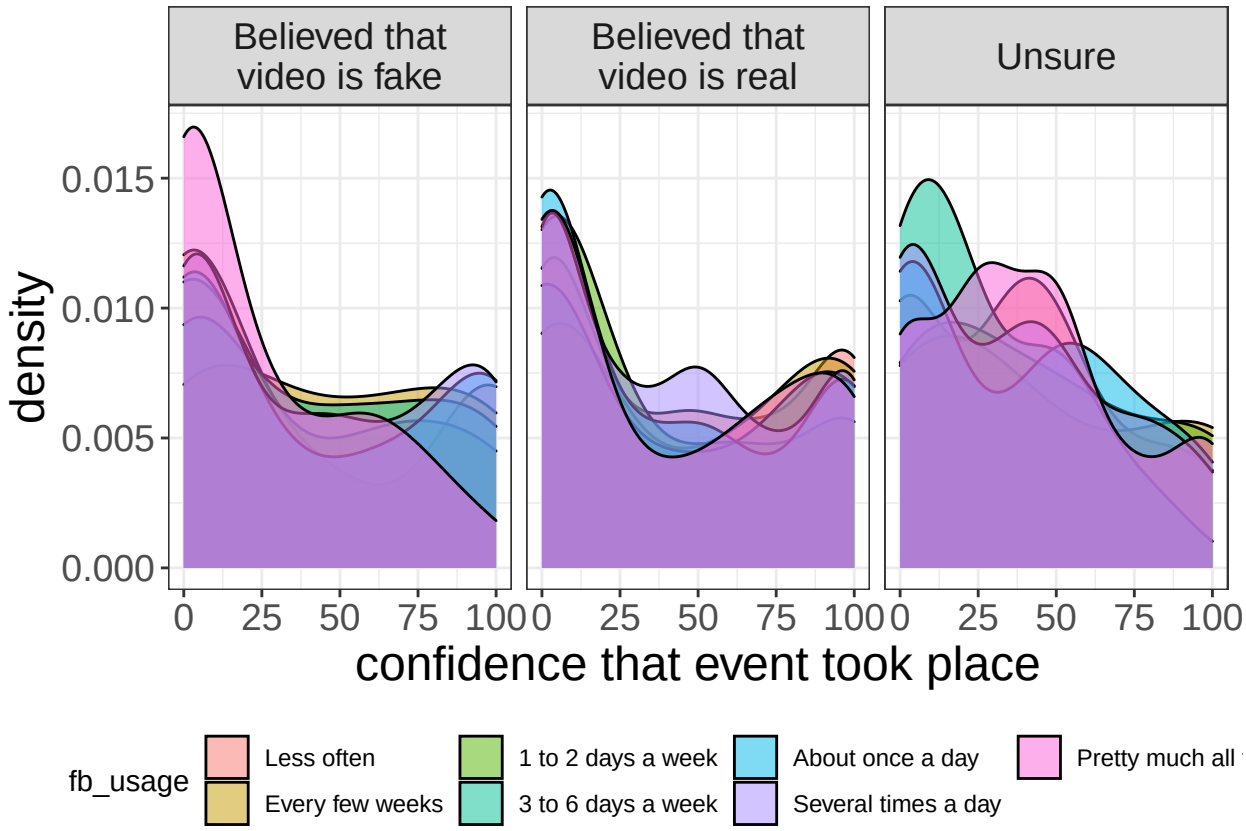
```
## Warning: Removed 13 rows containing non-finite values (stat_density).
```



6. **Fake: Hillary Clinton Debate**

The results were inconclusive for Facebook usage (Figure **??**.

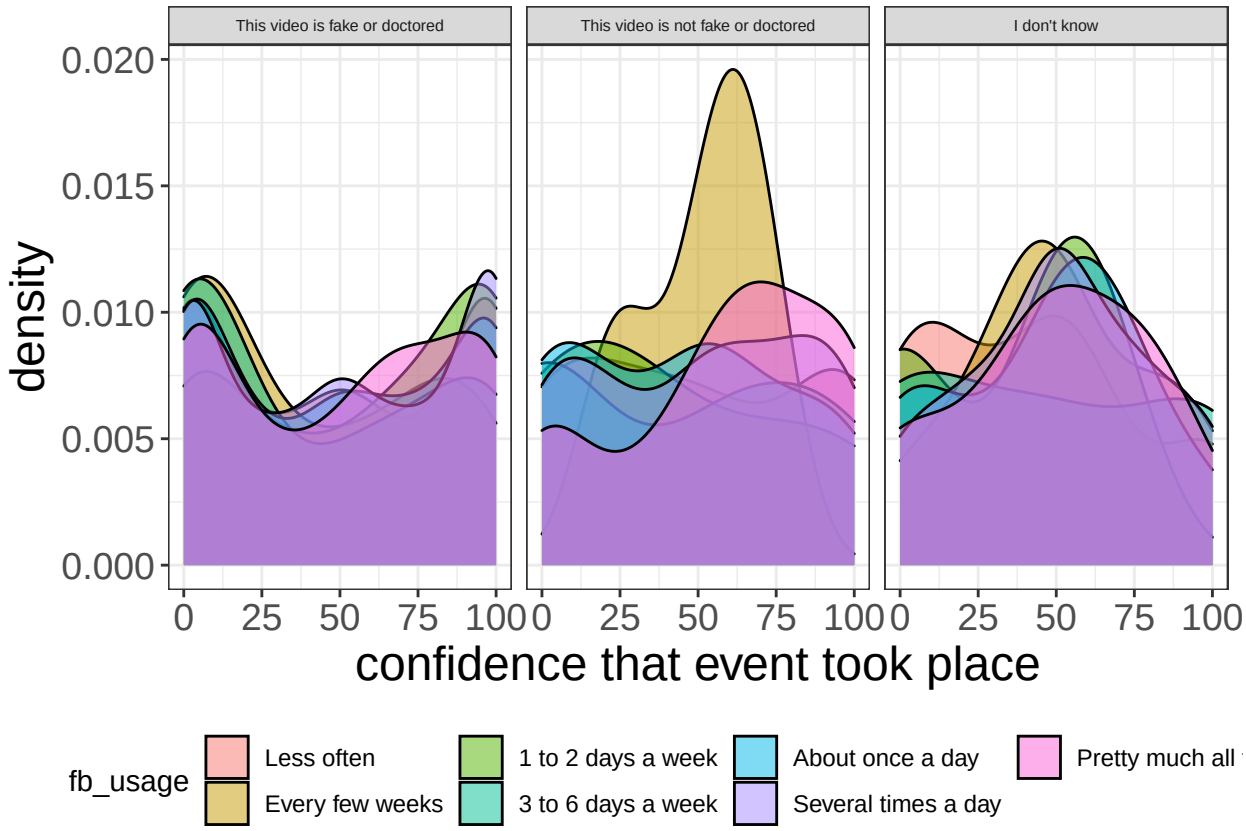
```
## Warning: Removed 29 rows containing non-finite values (stat_density).
```



7. **Fake: Donald Trump AIDS Cure Announcement**

The results were inconclusive for Facebook usage (Figure **??**.

```
## Warning: Removed 13 rows containing non-finite values (stat_density).
```



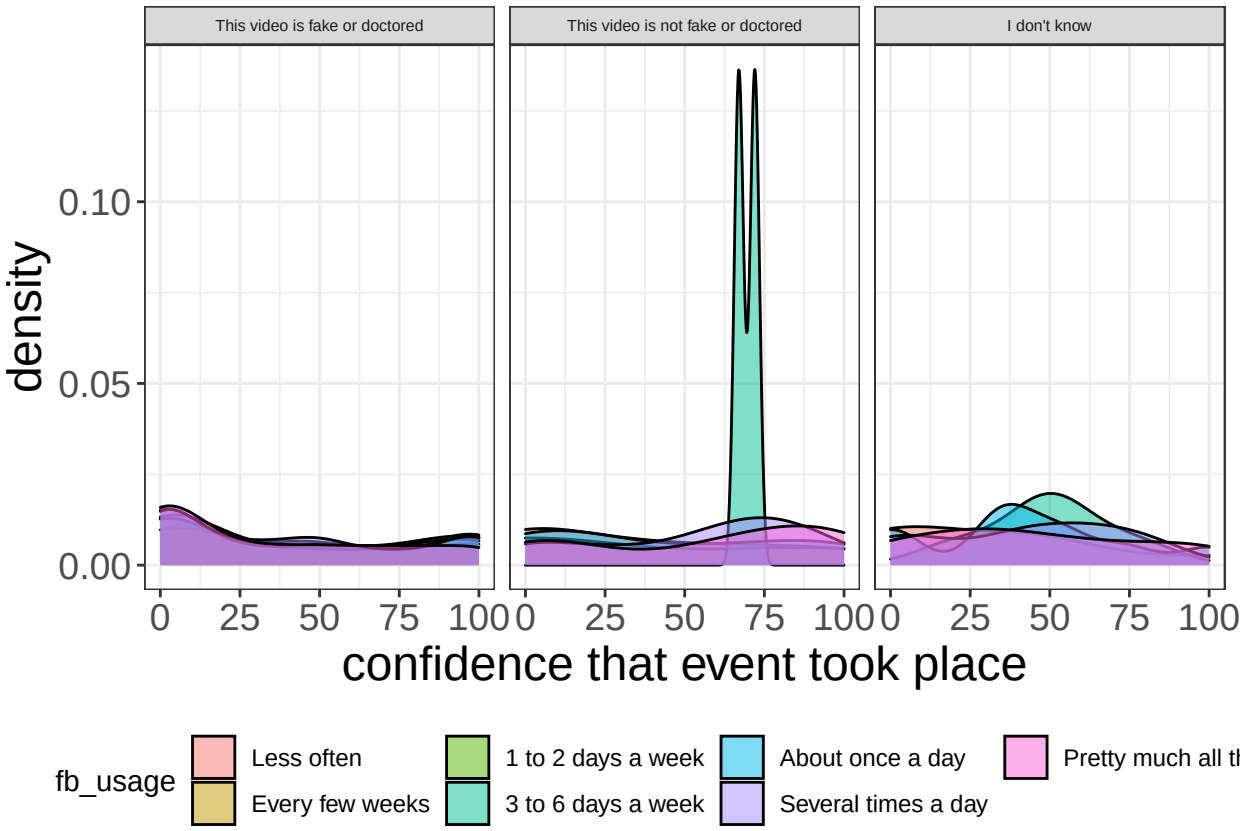
8. **Fake: Donald Trump Resigns**

The results were inconclusive for Facebook usage (Figure **??**. However, there is an odd pattern for those who use Facebook 3-6 times a week have a 60-75 percent confidence that the video is real.

```
## Warning: Removed 29 rows containing non-finite values (stat_density).
```

```
## Warning: Groups with fewer than two data points have been dropped.
```

```
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning -
## Inf
```



5.2 Correlation Between Political Affiliation and Believing a Video is Fake or True

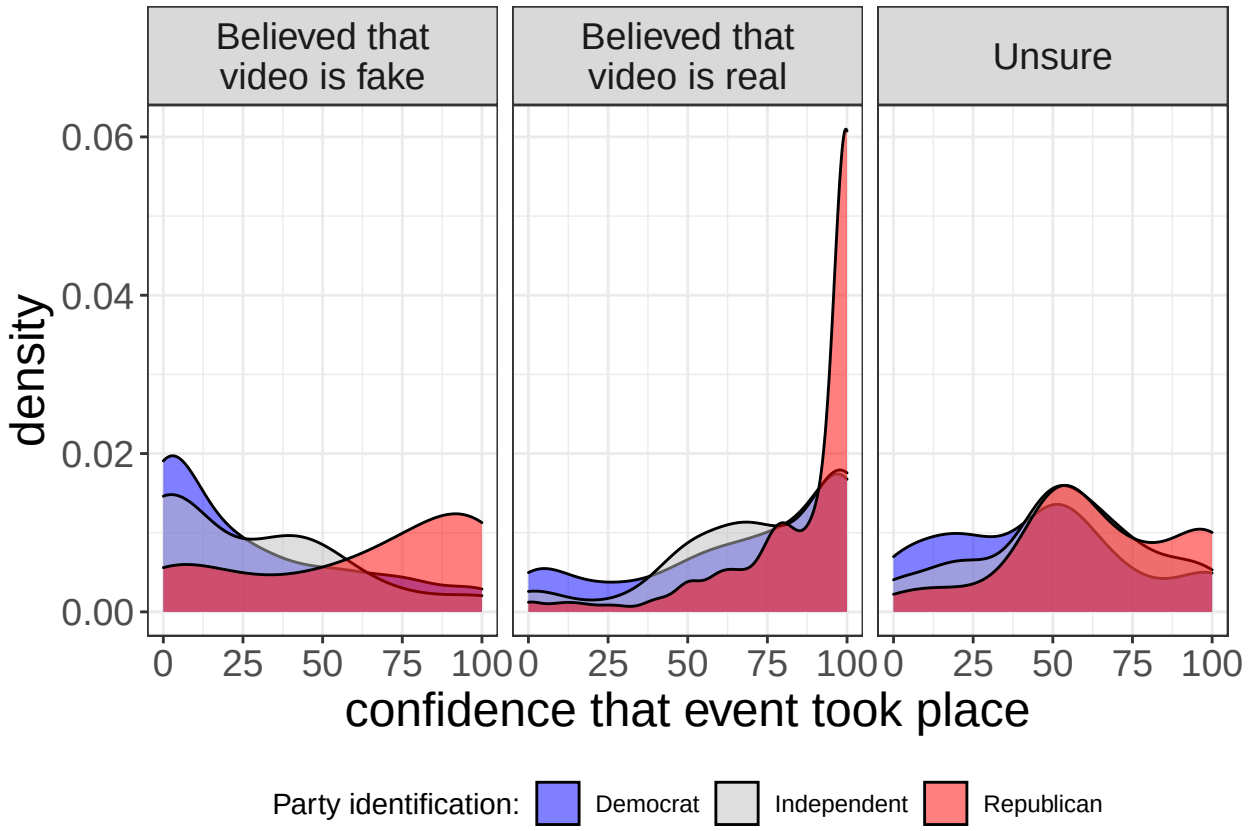
In contrast, political affiliation is a stronger predictor that a participant beleived a video was fake or real.

The graphs were built using R (R Core Team 2020), tidyverse (Wickham et al. 2019), stargazer (Hlavac 2018) and ggplot2 (Wickham 2016).

Real: Donald Trump COVID-19 Precautions Announcement

For instance, Republicans were significantly more likely to believe the video about Donald Trump's COVID-19 precautions was real compared to Democrats and Independents (Figure **??**. The data demonstrates that it does not really matter if the video is real or not because even though the Donald Trump COVID-19 video is real, Democrats and Republicans incorrectly labeled it as a "Deepfake" because the video did not conform to their political perceptions of Donald Trump.

```
## Warning: Removed 9 rows containing non-finite values (stat_density).
```



6 Results

6.1 Chi-Square Test for the Correlation Between Political Affiliation and Believing a Video is Fake or True

We used a Pearson's Chi-Square test to analyze the relationship between political affiliation and believing a video is fake or true. We are using a Chi-Square test X^2 because that was the test conducted by the principal authors. X^2 is used to determine whether there is a statistically significant result between the expected and observed frequencies in one or more categories of the contingency table. We are using a X^2 test because political affiliation and beleiving a video is real/false are both qualitative variables.

The formula for Pearson's Chi-Square test is:

$\chi^2 = \sum \frac{(O-E)^2}{E}$

Where O is the observed frequencies and E is the expected frequencies.

Our null and alternative hypothesis are:

H0: There is no relationship between facebook usage and identifying if the video is real or a deepfake.

H1: There is no relationship between facebook usage and identifying if the video is real or a deepfake.

We will set our alpha level to 0.05.

We will perform eight Chi-Square test: 2 real video and 2 fake videos for left-leaning politicians; and 2 real videos and 2 fake videos for right-leaning politicians. We used tidymodels to build these models (Kuhn and Wickham 2020).

1. **Real: Barack Obama Russian President Hot Mic**

```
##
##  Pearson's Chi-squared test
##
## data:  table(as.character(dfsurvdat$real_obama_missile[as.character(dfsurvdat$fb_usage) %in%     c("
## X-squared = 4.6801, df = 6, p-value = 0.5854
```

2. **Real: Barack Obama Smoking Hot Mic**

```
##
##  Pearson's Chi-squared test
##
## data:  table(as.character(dfsurvdat$real_obama_smoking[as.character(dfsurvdat$fb_usage) %in%     c("
## X-squared = 11.911, df = 6, p-value = 0.06398
```

3. **Real: Donald Trump Apple Press Conference**

```
##
##  Pearson's Chi-squared test
##
## data:  table(as.character(dfsurvdat$real_trump_apple[as.character(dfsurvdat$fb_usage) %in%     c("Les
## X-squared = 12.305, df = 6, p-value = 0.0555
```

4. **Real: Donald Trump COVID-19 Precautions Announcement**

```
##
##  Pearson's Chi-squared test
##
## data:  table(as.character(dfsurvdat$real_trump_covid[as.character(dfsurvdat$fb_usage) %in%     c("Les
## X-squared = 14.242, df = 6, p-value = 0.02704
```

5. **Fake: Obama Buzzfeed News Announcement**

```
##
##  Pearson's Chi-squared test
##
## data:  table(as.character(dfsurvdat$fake_obama_buzzfeed[as.character(dfsurvdat$fb_usage) %in%     c(
## X-squared = 4.3804, df = 6, p-value = 0.6253
```

6. **Fake: Hillary Clinton Debate**

```
##
##  Pearson's Chi-squared test
##
## data:  table(as.character(dfsurvdat$fake_hilary2[as.character(dfsurvdat$fb_usage) %in%     c("Less o
## X-squared = 3.3525, df = 6, p-value = 0.7635
```

7. **Fake: Donald Trump AIDS Cure Announcement**

```
##
##  Pearson's Chi-squared test
##
## data:  table(as.character(dfsurvdat$fake_trump_aids[as.character(dfsurvdat$fb_usage) %in%     c("Less
## X-squared = 56.795, df = 6, p-value = 2.01e-10
```

8. **Fake: Donald Trump Resigns**

```
## Warning in
## chisq.test(table(as.character(dfsurvdat$fake_trump_resign[as.character(dfsurvdat$fb_usage)
## %in% : Chi-squared approximation may be incorrect

##
```

```
##  Pearson's Chi-squared test
##
## data:  table(as.character(dfsurvdat$fake_trump_resign[as.character(dfsurvdat$fb_usage) %in%     c("L
## X-squared = 25.943, df = 6, p-value = 0.0002281
```

Out of our eight models, we received p-values that were below the alpha level for "Fake: Donald Trump AIDS Cure Announcement" and "Fake: Donald Trump Resigns."

Our p-values for Fake Donald Trump AIDS Cure Announcement and Donald Trump Resigns (2.01e-10 and 0.0002281 respectively) are well below the alpha level threshold of 0.05. Therefore, the p-value flags the data as being unusual if all the assumptions used to compute it (including the test hypothesis) were correct. A p-value less than 0.05 suggests that a discrepancy from the hypothesis prediction would be as large or larger than the observed no more than 5% of the time if only chance was creating the discrepancy. However, the low p-value may also be caused by a large random error or some assumptions other than the test hypothesis were violated. Despite the p-values being less than 0.05, the results for the other chi-square tests are all above the alpha level. Therefore, we do not feel comfortable completely rejecting our null hypothesis that there is no relationship between Facebook usuage and identifying a video as fake or real. (Sander Greenland and Altman 2016).

7 Discussion

Overall, our study concluded that there is no relationship between Facebook usage and individual's ability to detect deepfake videos. We assumed that the more our participants used Facebook, the more they would be able to detect fake news or deepfakes because Facebook uses an algorithm to detect deepfake videos and label them on the platform. Consequently, we assumed that people who used Facebook more often are more likely to be exposed to these deepfake videos in the past and thus are more likely to detect them as deepfakes. Instead, we observed that facebook usage is inconsequential to detecting deepfake videos correctly. Rather, an individual's political affiliation is a stronger predictor of identifying a video as fake or real based on their attitudes towards the political figure in the video and if the video's content matches their political values.

It is worth mentioning again that Facebook is working on detecting and eliminating deepfakes. In a statement published by the vice president of the Global Policy Management early last year, Facebook announced that it is going to remove deepfakes from their website as they "present a significant challenge for our industry and society as their use increases'' (Bickert 2020). The deepfake videos will be removed if they have been manipulated in such a way that they seem authentic and if they cannot be detectable by the average user. Additionally, the new policy "does not extend to content that is parody or satire, or video that has been edited solely to omit or change the order of words''. Which begs the question "who gets to decide what gets to stay and what should be removed?" Prior to this policy, Facebook has been criticized for refusing to remove a video of a US House of Representative speaker, Nancy Pelosi, that has been altered to make her seem to be slurring her speech (Waterson 2019). When asked about whether or not Facebook is going to remove this video under the new policy, Facebook told Reuters that it will not (Reuters 2020).

The authors of the original study took into account several factors that could influence detecting deepfakes. Factors like age, education, technical literacy, political literacy, the emotions the video stirred (funny, informative, offensive), ideologies (racism, sexism, etc.). Users were able to detect deepfakes correctly more when the news revolved around Donald Trump, the current president at the time. Fake news about finding a cure for AIDS or the president resigning yielded higher deepfake detection rates than any other news. This could be to the fact that these stories are implausible.

There are a few limitations to our analysis and proceeding experiment worth discussing, all of which circulate around a long standing tension between sample anonymity and sample accuracy. This tension is probably best laid out in work done but Lisa Austin and David Lie (2019), which speaks more directly to the balancing act that comes with making sure your data is both accurate and private. Speaking in terms of anonymization and de-anonymization, Austin and Lie engage with methods such as k-anonymity and differential privacy, measuring their respective strengths and limitations. Very briefly, k-anonymity is a process where specific data features are omitted from a sample (or for Austin and Lie, a dataset) to ensure no re-identification of

respondents is possible (Austin and Lie 2016, 592). Differential privacy is slightly more sophisticated as it requires the development of a model that measures significant "privacy loss" (Austin and Lie 2016, 583) in a sample and accounts for this metric in the proceeding statistical analysis (Austin and Lie 2016).

Another limitation was inaccessibility to study participants. Since we used thestudy was originally conducted by the principal authors, we could not redo the experiment or ask additional questions to study participants that the original authors did not consider. Rather than just focusing on Facebook usage, we could go more in-depth on the pages they visit, if they have been exposed to these videos (both real and fake before), or if they have heard information related to the videos circulating on the platform before. We could also examine their Facebook feed and see how often the participant is exposed to misinformation on the platform. With these in-depth questions, we could have conducted a larger study on how exposure to misinformation on Facebook contributes to people's willingness to believe deepfake videos.

A third limitation is that the original code was not fully reproducible. A significant challenge when designing the study and conducting the experiment was that the authors used multiple datasets, put the formatting for their graphs in a seperate spreadsheet, and the authors gave the same dataset multiple names throughout the code which made the code hard to follow. Another challenge we faced ws that the code has minimal commentary and so it was difficult to follow along with some segments and know their purpose. Thus, the original study has low internal validity.

Lastly, a weakness of the dataset was that the original author's sample collection method could potentially have selection bias. It is not clear from the original aithors how they collected the sample and how representative this sample is of the general American population they want to study. Although the same is very large at 5,833 people, this is a small number compared to the overall American population. We also don't know the hidden characteristics of these participants that were not captured in the survey which could be influencing our experiment's results. For example, the people that responded may have a large knowledge of deepfakes, they may all live in the same region, or another hidden factor not captured in the dataset. Thus, the original study has low internal and external validity.

When considering the implications of our intervention, however, there are a number of directions for further research that could all contribute unique insights into the ongoing deepfakes crisis and their impact on political behavior. For instance, the original dataset was conducted during the 2020 Presidential Election. It would be interesting to re-do the experiment a year after the study do examine if the partisan relationship still holds. In addition, further research and interventions geared towards countries outside the USA could examine if there is a relationship between people's belief that a video is real or fake based on their political affiliation and facebook usage depending on what country they are in. For example, Saudia Arabian government often uses deepfake videos as a propaganda tool. It would be interesting to reconduct the study in the country where people are more exposed to deepfake videos on a daily basis and there is greater public awareness about the technology. Would the study yield similar results had it been conducted in a different country where people are not strongly affiliated with politics or there is no democratic system? Would left-leaning democracies show different results than right-leaning democracies?

8 Appendix

Bibliography

Aronow, Peter M, Joshua Kalla, Lilla Orr, and John Ternovski. 2020. *Evidence of Rising Rates of Inattentiveness on Lucid in 2020.* http://osf.io/preprints/socarxiv/8sbe4.

Austin, L. M., and D. Lie. 2016. *Safe Sharing Sights.* https://www.nyulawreview.org/issues/volume-94-%20number-4/safe-sharing-sites/.

Bickert, Monica. 2020. *Enforcing Against Manipulated Media.* https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/.

Davis, Trevor L. 2020. *Optparse: Command Line Option Parser.* https://CRAN.R-project.org/package=optparse.

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to enable research on predicting sentiment polarity: given a piece of English text, predict whether it has a positive or negative affect—or stance—toward its topic. It was created intentionally with that task in mind, focusing on movie reviews as a place where affect/sentiment is frequently expressed.[1]

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by Bo Pang and Lillian Lee at Cornell University.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

Funding was provided though five distinct sources: the National Science Foundation, the Department of the Interior, the National Business Center, Cornell University, and the Sloan Foundation.

Any other comments?

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are movie reviews extracted from newsgroup postings, together with a sentiment rating for whether the text corresponds to a review with a rating that is either strongly positive (high number of stars) or strongly negative (low number of stars). The polarity rating is binary {positive,negative}. An example instance is shown in Figure 1.

How many instances are there in total (of each type, if appropriate)?

There are 1400 instances in total in the original (v1.x versions) and 2000 instances in total in v2.0 (from 2014).

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset is a sample of instances. It is (presumably) intended to be a random sample of instances of movie revies from newsgroup postings. No tests were run to determine representativeness.

[1]Information in this datasheet is taken from one of five sources; any errors that were introduced are our fault. http://www.cs.cornell.edu/people/pabo/movie-review-data/; http://xxx.lanl.gov/pdf/cs/0409058v1; http://www.cs.cornell.edu/people/pabo/movie-review-data/rt-polaritydata.README.1.0.txt; http://www.cs.cornell.edu/people/pabo/movie-review-data/poldata.README.2.0.txt.

> these are words that could be used to describe the emotions of john sayles' characters in his latest , limbo . but no , i use them to describe myself after sitting through his latest little exercise in indie egomania . i can forgive many things . but using some hackneyed , whacked-out , screwed-up * non * - ending on a movie is unforgivable . i walked a half-mile in the rain and sat through two hours of typical , plodding sayles melodrama to get cheated by a complete and total copout finale . does sayles think he's roger corman ?

Figure 1. An example "negative polarity" instance, taken from the file `neg/cv452_tok-18656.txt`.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)or features? In either case, please provide a description.

Each instance consists of the text associated with the review, with obvious ratings information removed from that text (some errors were found and alter fixed). The text was down-cased and HTML tags were removed. Boilerplate newsgroup header/footer text was removed. Some additional unspecified automatic filtering was done. Each instance also has an associated target value: a positive (+1) or negative (-1) rating based on the number of stars that that review gave (details on the mapping from number of stars to polarity is given below in "Data Preprocessing").

Is there a label or target associated with each instance? If so, please provide a description.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Everything is included. No data is missing.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

None explicitly, though the original newsgroup postings include poster name and email address, so some information could be extracted if needed.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The instances come with a "cross-validation tag" to enable replication of cross-validation experiments; results are measured in classification accuracy.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient con-

Figure 1: Data Sheets 1

fidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

Some movie reviews might contain moderately inappropriate or offensive language, but we do not expect this to be the norm.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

The raw form of the dataset contains names and email addresses, but these are already public on the internet newsgroup.

Any other comments?

Collection Process

Similar to Composition, this section should be read during the initial planning phase, and filled out during the collection of data. Again, these questions provide general transparency into the makeup of the data help both the dataset creator and dataset consumer uncover risks and potential harms, for example by questioning whether those whose information is contained in the dataset have control over usage of their data or the ability to remove their information from the dataset entirely.

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The data was mostly observable as raw text, except the labels were extracted by the process described below. The data was collected by downloading reviews from the IMDb archive of the `rec.arts.movies.reviews` newsgroup, at http://reviews.imdb.com/Reviews.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

Unknown.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The sample of instances collected is English movie reviews from the `rec.arts.movies.reviews` newsgroup, from which a "number of stars" rating could be extracted. The sample is limited to forty reviews per unique author in order to achieve broader coverage by authorship. Beyond that, the sample is arbitrary.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Unknown

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

Unknown

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

Unknown

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

The dataset relates to people in that the reviews themselves are authored by people. Personally identifying information (e.g., email addresses) was removed.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

The data was collected from newsgroups.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

No. The data was crawled from public web sources, and the authors of the posts presumably knew that their posts would be public, but there was no explicit informing of these authors that their posts were to be used in this way.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

No (see previous question).

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis)been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A.

Any other comments?

Figure 2: Data Sheets 2

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Instances for which an explicit rating could not be found were discarded. Also only instances with strongly-positive or strongly-negative ratings were retained. Star ratings were extracted by automatically looking for text like "`⋆⋆⋆⋆ out of ⋆⋆⋆⋆⋆`" in the review, using that as a label, and then removing the corresponding text. When the star rating was out of five stars, anything at least four was considered positive and anything at most two negative; when out of four, three and up is considered positive, and one or less is considered negative. Occasionally half stars are missed which affects the labeling of negative examples. Everything in the middle was discarded. In order to ensure that sufficiently many authors are represented, at most 20 reviews (per positive/negative label) per author are included.

In a later version of the dataset (v1.1), non-English reviews were also removed.

Some preprocessing errors were caught in later versions. The following fixes were made: (1) Some reviews had rating information in several places that was missed by the initial filters; these are removed. (2) Some reviews had unexpected/unparsed ranges and these were fixed. (3) Sometimes the boilerplate removal removed too much of the text.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

Yes. The dataset itself contains all the raw dta.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

No.

Any other comments?

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

At the time of publication, only the original paper http://xxx.lanl.gov/pdf/cs/0409058v1. Between then and 2012, a collection of papers that used this dataset was maintained at http://www.cs.cornell.edu/people/pabo/movie%2Dreview%2Ddata/otherexperiments.html.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

There is a repository, maintained by Pang/Lee through April 2012, at http://www.cs.cornell.edu/people/pabo/movie%2Dreview%2Ddata/otherexperiments.html.

What (other) tasks could the dataset be used for?

The dataset could be used for anything related to modeling or understanding movie reviews. For instance, one may induce a lexicon of words/phrases that are highly indicative of sentiment polarity, or learn to automatically generate movie reviews.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

There is minimal risk for harm: the data was already public, and in the preprocessed version, names and email addresses were removed.

Are there tasks for which the dataset should not be used? If so, please provide a description.

This data is collected solely in the movie review domain, so systems trained on it may or may not generalize to other sentiment prediction tasks. Consequently, such systems should not—without additional verification—be used to make consequential decisions about people.

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes, the dataset is publicly available on the internet.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset is distributed on Bo Pang's webpage at Cornell: http://www.cs.cornell.edu/people/pabo/movie-review-data. The dataset does not have a DOI and there is no redundant archive.

When will the dataset be distributed?

The dataset was first released in 2002.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The crawled data copyright belongs to the authors of the reviews unless otherwise stated. There is no license, but there is a request to cite the corresponding paper if the dataset is used: *Thumbs up? Sentiment classification using machine learning techniques.* Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Proceedings of EMNLP, 2002.

Figure 3: Data Sheets 3

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

Unknown

Any other comments?

Maintenance

This section should be completed once the dataset has been constructed, before it is distributed. These questions help the dataset creator think through their plans for updating, adding to, or fixing errors in the dataset, and expose these plans to dataset consumers.

Who is supporting/hosting/maintaining the dataset?

Bo Pang is supporting/maintaining the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Unknown

Is there an erratum? If so, please provide a link or other access point.

Since its initial release (v0.9) there have been three later releases (v1.0, v1.1 and v2.0). There is not an explicit erratum, but updates and known errors are specified in higher version `README` and `diff` files. There are several versions of these: v1.0: http://www.cs.cornell.edu/people/pabo/movie-review-data/README; v1.1: http://www.cs.cornell.edu/people/pabo/movie%2Dreview%2Ddata/README.1.1 and http://www.cs.cornell.edu/people/pabo/movie-review-data/diff.txt; v2.0: http://www.cs.cornell.edu/people/pabo/movie%2Dreview%2Ddata/poldata.README.2.0.txt. Updates are listed on the dataset web page. (This datasheet largely summarizes these sources.)

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

This will be posted on the dataset webpage.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

The dataset has already been updated; older versions are kept around for consistency.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

Others may do so and should contact the original authors about incorporating fixes/extensions.

Any other comments?

Figure 4: Data Sheets 4

Table F20: **Predictors of Second-Stage Detection Accuracy**

	Detection Accuracy (% Correctly Classified)					
	(1)	(2)	(3)	(4)	(5)	(6)
Digital Literacy		0.25***	0.22***	0.22***	0.20***	0.21***
		(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Accuracy Prompt	−0.002		−0.01	−0.004	−0.005	−0.002
	(0.01)		(0.01)	(0.01)	(0.01)	(0.01)
Stage 1 Debrief			0.01*	0.01*	0.01*	0.01*
			(0.01)	(0.01)	(0.01)	(0.01)
Stage 1 Info Provided			−0.01	−0.001	−0.01	−0.004
			(0.01)	(0.01)	(0.01)	(0.01)
Political Knowledge			0.18***	0.19***	0.18***	0.20***
			(0.01)	(0.01)	(0.01)	(0.01)
Internet Usage			−0.002	−0.01	0.0001	−0.004
			(0.005)	(0.005)	(0.01)	(0.01)
Low-fake Env.			0.03***	0.04***	0.03***	0.05***
			(0.01)	(0.01)	(0.01)	(0.01)
No-fake Env.			0.04***	0.04***	0.04***	0.05***
			(0.01)	(0.01)	(0.01)	(0.01)
Age 65+			0.01	0.01	0.01	0.003
			(0.01)	(0.01)	(0.01)	(0.01)
High School			0.01	0.01	0.03	0.02
			(0.03)	(0.02)	(0.03)	(0.02)
College			0.02	0.02	0.04	0.03
			(0.03)	(0.02)	(0.03)	(0.02)
Postgrad			−0.01	−0.02	0.01	−0.01
			(0.03)	(0.02)	(0.03)	(0.02)
Republican			0.06***	0.07***	0.07***	0.08***
			(0.02)	(0.02)	(0.02)	(0.02)
CRT			−0.06**	−0.06**	−0.06**	−0.06**
			(0.03)	(0.03)	(0.03)	(0.03)
Republican x CRT			0.09***	0.07***	0.09***	0.08***
			(0.01)	(0.01)	(0.01)	(0.01)
Ambivalent Sexism			0.001	−0.002	0.001	−0.0001
			(0.004)	(0.004)	(0.004)	(0.004)
Constant	0.57***	0.36***	0.16***	0.20***	0.14***	0.16***
	(0.005)	(0.02)	(0.05)	(0.04)	(0.05)	(0.05)
Weighted?				✓		✓
Low-Quality Dropped?					✓	✓
N	5,497	5,497	5,496	5,496	4,870	4,870
R^2	0.0000	0.02	0.09	0.09	0.09	0.10
Adjusted R^2	−0.0002	0.02	0.09	0.09	0.09	0.10

*p < .1; **p < .05; ***p < .01

Notes: Reference category for environment is High-fake. PID pooled for brevity.

Figure 5: Principal Experiment Linear Model 1

Table F21: **Predictors of Second-Stage False Positive Rate (FPR)**

	Detection FPR (% Real Videos Classified as Deepfakes)					
	(1)	(2)	(3)	(4)	(5)	(6)
Digital Literacy		−0.11***	−0.08***	−0.11***	−0.06**	−0.08***
		(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Accuracy Prompt	−0.01		0.004	−0.01	0.003	−0.01
	(0.01)		(0.01)	(0.01)	(0.01)	(0.01)
Stage 1 Debrief			−0.03***	−0.03***	−0.03***	−0.03***
			(0.01)	(0.01)	(0.01)	(0.01)
Stage 1 Info Provided			0.01	0.001	0.01	0.004
			(0.01)	(0.01)	(0.01)	(0.01)
Political Knowledge			−0.13***	−0.13***	−0.13***	−0.14***
			(0.02)	(0.02)	(0.02)	(0.02)
Internet Usage			−0.002	0.002	−0.002	0.001
			(0.01)	(0.01)	(0.01)	(0.01)
Low-fake Env.			0.03***	0.02***	0.02***	0.01
			(0.01)	(0.01)	(0.01)	(0.01)
No-fake Env.			0.23***	0.23***	0.22***	0.21***
			(0.01)	(0.01)	(0.01)	(0.01)
Age 65+			0.003	−0.003	0.01	0.0000
			(0.01)	(0.01)	(0.01)	(0.01)
High School			0.001	−0.003	−0.02	−0.02
			(0.03)	(0.02)	(0.04)	(0.02)
College			0.01	0.01	−0.02	−0.01
			(0.03)	(0.02)	(0.04)	(0.02)
Postgrad			0.03	0.02	0.001	0.002
			(0.04)	(0.02)	(0.04)	(0.02)
Republican			−0.06***	−0.08***	−0.07***	−0.08***
			(0.02)	(0.02)	(0.02)	(0.02)
CRT			0.03	0.03	0.04	0.03
			(0.03)	(0.03)	(0.03)	(0.03)
Republican x CRT			−0.07***	−0.07***	−0.08***	−0.07***
			(0.01)	(0.01)	(0.01)	(0.01)
Ambivalent Sexism			0.0005	0.002	−0.002	0.001
			(0.005)	(0.005)	(0.005)	(0.005)
Constant	0.28***	0.37***	0.42***	0.44***	0.42***	0.44***
	(0.01)	(0.02)	(0.06)	(0.05)	(0.06)	(0.05)
Weighted?				✓		✓
Low-Quality Dropped?					✓	✓
N	5,495	5,495	5,494	5,494	4,869	4,869
R^2	0.0002	0.003	0.16	0.16	0.16	0.16
Adjusted R^2	−0.0000	0.003	0.15	0.16	0.15	0.16

*p < .1; **p < .05; ***p < .01

Notes: Reference category for environment is High-fake. PID pooled for brevity.

Figure 6: Principal Experiment Linear Model 2

Table F22: **Predictors of Second-Stage False Negative Rate (FNR)**

	Detection FNR (% Deepfakes Classified as Real Videos)					
	(1)	(2)	(3)	(4)	(5)	(6)
Digital Literacy		−0.03	−0.04	−0.01	−0.05	−0.02
		(0.03)	(0.03)	(0.03)	(0.03)	(0.04)
Accuracy Prompt	−0.01		−0.005	−0.01	0.003	−0.003
	(0.01)		(0.01)	(0.01)	(0.01)	(0.01)
Stage 1 Debrief			0.01	0.01	0.01	0.01
			(0.01)	(0.01)	(0.01)	(0.01)
Stage 1 Info Provided			−0.002	0.003	−0.01	−0.003
			(0.01)	(0.01)	(0.01)	(0.01)
Political Knowledge			−0.0004	−0.03	−0.003	−0.03
			(0.02)	(0.02)	(0.02)	(0.02)
Internet Usage			0.02***	0.01	0.02***	0.02**
			(0.01)	(0.01)	(0.01)	(0.01)
Low-fake Env.			0.01	0.001	0.01	−0.004
			(0.01)	(0.01)	(0.01)	(0.01)
No-fake Env.			0.01	0.02	0.02*	0.02*
			(0.01)	(0.01)	(0.01)	(0.01)
Age 65+			0.01	0.02	0.01	0.01
			(0.04)	(0.03)	(0.04)	(0.03)
High School			0.03	0.04	0.03	0.03
			(0.04)	(0.03)	(0.04)	(0.03)
College			0.08*	0.13***	0.08*	0.11***
			(0.04)	(0.03)	(0.05)	(0.03)
Postgrad			−0.05**	−0.04	−0.06**	−0.04
			(0.02)	(0.03)	(0.03)	(0.03)
Republican			0.09**	0.08*	0.11***	0.10**
			(0.04)	(0.04)	(0.04)	(0.04)
CRT			0.02***	0.02***	0.01**	0.01
			(0.01)	(0.01)	(0.01)	(0.01)
Republican x CRT			−0.04***	−0.03**	−0.05***	−0.04**
			(0.01)	(0.02)	(0.02)	(0.02)
Ambivalent Sexism	0.34***	0.36***	0.16**	0.21***	0.18**	0.23***
	(0.01)	(0.03)	(0.07)	(0.06)	(0.07)	(0.06)
Weighted?				✓		✓
Low-Quality Dropped?					✓	✓
N	3,690	3,690	3,690	3,690	3,266	3,266
R^2	0.0002	0.0003	0.02	0.03	0.02	0.02
Adjusted R^2	−0.0000	0.0000	0.02	0.02	0.01	0.01

*p < .1; **p < .05; ***p < .01

Notes: Reference category for environment is High-fake. PID pooled for brevity.

Figure 7: Principal Experiment Linear Model 3

Gebru, Timit. 2018. *Datasheets for Datasets.* https://arxiv.org/abs/1803.09010.

Hlavac, Marek. 2018. *Stargazer: Well-Formatted Regression and Summary Statistics Tables.* Bratislava, Slovakia: Central European Labour Studies Institute (CELSI). https://CRAN.R-project.org/package=stargazer.

Kuhn, Max, and Hadley Wickham. 2020. *Tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles.* https://www.tidymodels.org.

Lumley, T. 2020. *Survey: Analysis of Complex Survey Samples.* https://cran.r-project.org/web/packages/survey/index.html.

Pasek, Josh, with some assistance from Alex Tahk, some code modified from R-core; Additional contributions by Gene Culter, and Marcus Schwemmle. 2020. *Weights: Weighting and Weighted Statistics.* https://CRAN.R-project.org/package=weights.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Reuters, Thomson. 2020. *Facebook to Remove Deepfake Videos in Run up to 2020 US Election.* https://www.reuters.com/article/us-facebook-deepfake/facebook-to-remove-deepfake-videos-in-run-up-to-2020-u-s-election-idUSKBN1Z60JV.

Robinson, David, Alex Hayes, and Simon Couch. 2021. *Broom: Convert Statistical Objects into Tidy Tibbles.* https://CRAN.R-project.org/package=broom.

Sander Greenland, Stephen J. Senn, Kenneth J. Rothman, and Douglas G. Altman. 2016. *Statistical Tests, p Values, Confidence Intervals and Power: A Guide to Misinterpretations.* https://www.tellingstorieswithdata.com.

Soubhnik Barari, Kevin Munger, Christopher Lucas. 2021. *Political Deepfake Videos Misinform the Public, but No More Than Other Fake Media.* https://osf.io/cdfh3/.

Waterson, Jim. 2019. *Facebook Refuses to Delete Fake Pelosi Video Spread by Trump Supporters.* https://www.theguardian.com/technology/2019/may/24/facebook-leaves-fake-nancy-pelosi-video-on-site.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Zhu, Hao. 2020. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax.* https://CRAN.R-project.org/package=kableExtra.