The Varied Impact of Advertisement Negativity on Political Advertisement Engagement: A Computational Case Study of Russian-Linked Facebook and Instagram Content

Vargo, C., & Hopp, T. (2020)

A pre-print of the article that appeared in *Journalism & Mass Communication Quarterly*. 97(3), 743-761. https://doi.org/10.1177/1077699020911884

This study examined associations between uncivil content and engagement with political advertisements placed on Facebook and Instagram by the Internet Research Agency (IRA). Using U.S House of Representatives data and Google's Perspective API, 2,515 IRA-linked advertisements were coded for: identity-based insults, inflammatory language, obscene language, and threatening language. Theory suggests a positive link between negative commentary and political engagement. We hypothesized that advertisements with uncivil text would also be associated with higher clickthrough rates. Clickthrough rates improved across time.
Still, support was offered for inflammatory, obscene, and threatening advertisements. Contrary to expectations, we observed a negative relationship between the use of identity-based insults and advertisement engagement.

The Varied Impact of Advertisement Negativity on Political Advertisement Engagement: A Computational Case Study of Russian-Linked Facebook and Instagram Content

In terms of advertising, the 2016 U.S. presidential election was novel for many reasons. While digital advertising has been ubiquitous for recent election cycles (e.g., Kreiss, 2016), this was the first documented time foreign actors became advertisers. Perhaps the most notable instance occurred when the Internet Research Agency (IRA), located in Russia, created and promoted social media content on Facebook and Instagram. The Democrats' Intelligence Committee subsequently found that the IRA did so "to sow discord in the U.S. political system" as a means of interfering with the 2016 presidential election (Boland, 2018).

Beyond foreign interference, the 2016 election shone a light on the fact that, while there are no current legal and self-regulatory rules to protect consumers from malicious television content, broadcasters tend not to air content that is openly racist or blatantly vulgar (Wang, 2018). Moreover, it is required for television advertisers to disclose themselves. Given the scale and scope of online advertising, disclosure is not required and most advertisements are not manually reviewed before they appear in front of consumers. Instead, social media platforms tend to rely on automated methods to screen ads for approval (e.g., Facebook Newsroom, 2018). As a result, the advertisements that ran leading up to and through the 2016 elections were subject to limited amounts of scrutiny from the social media platforms on which they were placed.

Given the lack of regulation and oversight, advertisements that were run by foreign actors were negative in new and often extreme ways (Popken & Cobiella, 2017). For the first time, advertisements exhibited hate, insults, and threats geared towards identities and groups. Unlike most traditional negative political advertisements, most Facebook and Instagram advertisements weren't created with the sole aim of electing a candidate, but instead to divide Americans

(Penzenstadler, Heath, & Guynn, 2018). Indeed, these advertisements commonly provoked individuals to engage in "emotional political issues" including LGBT rights, police brutality, gun regulation, and immigration (Edgerton & Frier, 2018). The thinking among disinformation analysts is that the ultimate goal of these appeals was to further polarize Americans on divisive political issues (e.g., Zadrozny, Popken, Collins, & Lehren, 2018).

To shine further light on these new foreign advertisements, the Democratic minority of the U.S House of Representatives' Permanent Select Committee on Intelligence released a total of 3,525 Facebook and Instagram advertisements placed from 2015 to 2017. The advertisements were "… one of the clearest demonstrations of Russia's financial investment in disrupting American politics" (Edgerton & Frier, 2018, para. 2). The promoted ads were seen by an estimated 11 million Americans (Ha, 2018) and drew clicks from upwards of 4 million Facebook and Instagram users (Boland, 2018).

While most digital advertising on Facebook is designed to deliver some type of engagement in return for advertising spend, we show here that user engagement with IRA-placed Facebook and Instagram advertisements was substantially higher than rates commonly observed on Facebook and Instagram (Nanji, 2017). Beyond this, we draw upon prior theory pertaining to negative political advertising effects and propose that IRA-placed advertisements that invoke negative and/or uncivil language will receive even more engagement. To test this hypothesis, we employ advertising data made public by the United States Government and Google's Perspective API to detect the presence of identity-based insults, inflammatory language, obscene language, and threatening language. Using metadata associated with each advisement, we subsequently assessed the degree to which these toxic language components were associated with clickthrough rate, a commonly employed metric of digital user engagement that describes the percentage of impressions that are converted into clicks.

Literature Review

Negative Political Advertisements

A brief history of negative political ads. While many things were unprecedented in the 2016 presidential election, negative political advertisements are not new. The first commonly accepted direct attack one candidate made against another was in 1952 when Estes Kefauver attacked then presidential candidate Eisenhower (Diamond & Bates, 1988). In this way, the basic facet of the majority of political ads have remained the same: advertisements are made by, or on behalf of, a political candidate to disparage and discredit a competing candidate for a political office (Jamieson, 1993). Negative political ads can be used to persuade the public on important policy issues, such as the Affordable Care Act (Gollust, Barry, Niederdeppe, Baum, & Fowler, 2014). These ads, while rising in popularity on social media and digital platforms, appear to be a minority compared to candidate-based ads.

While salacious and extreme candidate attacks have been printed by other candidates in papers for centuries, many have attributed the surge of negative political ads to the rise of television (Johnson-Cartee & Copeland, 2013). Television opened up direct access to the consumer in immersive ways that allowed for greater campaign discourse (Jamieson, 1993). Cable television only furthered campaigns reliance on TV ads because of its ability to directly target specific geographic locations (Fowler & Ridout, 2012). For the first time, battleground states and hotly contested districts could be targeted directly. With these affordances comes added cost. Television advertising is not cheap. In 2016, Donald Trump and Hillary Clinton spent some $595 million on TV and radio (Dumenco, 2016). Scholars believe that the sheer cost of television advertising has driven it to be more negative in an effort to maximize the effects of the advertising (Lau, Sigelman, Heldman, & Babbit, 1999). During the 2012 presidential election, 85% of ads sponsored by groups were directed solely at attacking a candidate (Dowling & Wichowsky, 2015). The majority of ads that attacked Obama didn't even mention Romney, and vice versa.

The effects of negative political advertising. Some debate still exists in the literature as to the true effects of negative advertising on individuals. While the majority of scholarship seems to find power in negative advertising, some disagree as to the degree to which negative, or attack advertising influences specific audience behaviors such as political polarization and voting behavior (e.g., Ansolabehere & Iyengar, 1995; Lau et al., 1999). The dissenting group of scholars feel that negative advertisements have contributed significantly to the increasingly negative political landscape. The thinking is, as political discourse becomes more negative, Americans are more likely to become dissatisfied with the state of politics. This dissatisfaction results in a withdrawal of involvement in the political process (Ansolabehere & Iyengar, 1995). This perspective holds that negative political advertising *prohibits political engagement* through disenfranchisement.

A newer batch of research has found that as negative, attack-oriented advertisements increase, so does citizen engagement with the political process (e.g., Lau et al., 1999). A number of mechanisms may explain this effect. Negative advertisements signal "that something important is at stake in the outcome of the election" (Goldstein & Freeman, 2002, p. 735). This heightened signal of importance can stimulate and motivate audiences to act (Martin, 2004).

Negative ads may communicate to the voter that their participation truly matters. Martin (2004) also suggests that negative campaign advertising invokes a sense of duty. His thinking includes the idea that American citizens share some concern over the future of the country; if that concern is activated, it encourages participation. For instance, attack advertisements express concern over the status quo, that everything would not be fine if a certain candidate was elected. That concern, if activated, could motivate individuals to take action in an effort to remedy the situation. This thinking is not unique. Research has long shown that citizens act when they feel the status quo is unacceptable (Berelson, Lazarsfeld, & McPhee, 1954).

Moreover, scholars have shown that people are more attentive to negative information over positive and neutral information. This is thought to be involuntary in nature. As illustrated by Pratto and John (1991), “Events that may negatively affect the individual are typically of greater time urgency than are events that lead to desirable consequences” (p. 380). Thus, attack ads may garner more attention than non-attack advertisements (Marcus, 2000), resulting in increased engagement and amplified levels of voter turnout (Martin, 2004). When negative, arousing emotional states are activated in audiences through advertising, these emotions encourage individuals to take action. High arousal emotions such as outrage and disgust can increase voter turnout (Martin, 2004).

Turning to negative political advertising on social media, Settle and colleagues (2015) indicated that Facebook users who resided in contested “battleground” states were more likely to post election-relevant content than those living in less competitive “blackout” states. Although advertising effects were not directly measured, their results described the competitive nature of campaigns, all of which were driven by heavy use of attack advertising. The authors concluded that advertising likely drove engagement online. In a more direct test, Hopp and Vargo (2017) looked at the 2012 presidential election and tested the relationship between negative political advertising and negative political engagement on Twitter. Using advertising television expenditure data and messages from 140,000 Twitter users across the U.S., the results showed that as more negative candidate advertising was served in an area, Twitter users engaged more and generated more social media content in those areas.

The IRA Facebook and Instagram ads. The advertisements released by Democrats on the U.S. House Intelligence Committee were unlike any negative political advertisement aforementioned in the literature (Edgerton & Frier, 2018). During the 2016 election, no U.S. regulatory body provided oversight on what political advertisements ran on social media platforms (Popken & Cobiella, 2017). At the same time, it appeared that Facebook and others were neither prepared nor willing to screen or protect consumers from political advertisements that were divisive in nature (Facebook Newsroom, 2018). As a result, advertisements ran largely unregulated and unchecked. Unlike the majority of traditional negative political advertisements, most Facebook and Instagram ads weren't created with the sole aim of electing a candidate, but instead to divide Americans. Only "about 100" of the ads mentioned Donald Trump or Hillary Clinton (Penzenstadler et al., 2018). Instead, the majority of ads provoked individuals to engage in "emotional political issues" (Zadrozny et al., 2018).

In the below sections, we provide an overview of typical forms of affectively negative or otherwise inflammatory language featured in many of the IRA-linked advertisements. It is our general contention, based upon the general political advertising literature discussed above, that the presence of these communicatory attributes is positively linked to user engagement. Four negative communication attributes are identified: identity-based insults, inflammatory language, obscene language, and threatening language.

Identity-based insults. Commentators have observed that IRA-linked Facebook advertisements often contained “hateful, xenophobic rhetoric” (Penzenstadler et al., 2018). Indeed, close examination of the advertisements shows that some copy text exhibited hate towards specific groups of people based on their identities. For example, one advertisement noted that “We simply can't allow Muslims to wear burka, otherwise everybody who wants to commit a crime or terror attack would wear this ugly rug and hide his or hers (sic) identity behind it.” Another advertisement stated that “America is deceased. Islamic terror has penetrated our homeland and now spreads” (Boland, 2018).

Inflammatory language. IRA-linked advertisements often contained language that was inflammatory in nature (Fandos, Kang & Isaac, 2017). Broadly construed, inflammatory language refers to low-quality commentary that is not rooted in substantive criticism but instead employs unnecessary insults or is unnecessarily aggressive (Hmielowski, Hutchens, & Cicchirillo, 2014). Importantly, inflammatory language extends beyond “bad words” and can instead be understood as insulting language that isn’t linked to racial, ethnic, or similar identity attributes. In the instant corpus of advertisements, there are numerous examples of copy text employing pejorative descriptors such as “sissy”, “psychopath”, and “mentally ill idiot.” Other ads use descriptions such as “deplorable” and “libtard” (Boland, 2018).

Obscene language. Social media advertisements placed by the IRA often employed obscene or profane language. Prior work on political discourse has classified obscene language as vulgar, socially restricted language (e.g., swear words that do not appear on television due to Federal Communication Commission restrictions; Sobieraj & Berry, 2011). In this study, we generally conceptualized obscenities as undirected in nature (i.e., can be contrasted with inflammatory language on the basis of a lack of clear target identification). Obscene language found in ad body copy included all commonly regarded swear words and several fewer common obscenities (Boland, 2018).

Threats. Finally, the Facebook and Instagram IRA advertisements often described potential threats facing individuals and urged them to act. Just as Martin (2004) observed of traditional negative political advertisements, the IRA often encouraged groups to reject the status quo and act to change the situation. Calls to action such as "They killed an UNARMED guy again!!!We MUST make the cops stop thinking that they are above the law. That is in our power to do?" These advertisements were almost always targeted to the individuals that the threat was of consequence to. Advertisements encouraged audiences to attend rallies and speak out against issues, particularly around race. These ads were created to "stok[e] fear to get a negative action directed at a targeted population" (Penzenstadler et al., 2018, para. 25).

Examples of the IRA advertisements are provided in Figure 1.

Hypotheses

As demonstrated above, IRA-generated Facebook and Instagram advertisements were digital and negative in extreme ways. Given Settle et al.'s (2015) and Hopp and Vargo's (2017) findings that negativity in ads fosters increased engagement on social media, coupled with the broader research that explicates the stimulation effect negative advertising has on citizens (e.g., Goldstein & Freeman, 2002; Lau et al., 1999), we posit that these new, extremely negative advertisements will receive increased engagement. In other words, much like negative television advertising may stimulate citizens to pay attention to a political campaign (e.g., Brians & Wattenberg, 1996) or mobilize voters (e.g., Martin, 2004), we suggest that negative ads placed on Facebook and Instagram will garner increased attention and engagement, as evidenced in the form of advertising clickthrough rate. Thus, the following hypotheses are posited:

H1: *IRA advertisements exhibiting identity-based insults will be associated with higher clickthrough rates.*

H2: *IRA advertisements exhibiting inflammatory language will be associated with higher clickthrough rates.*

H3: *IRA advertisements exhibiting obscene language will be associated with higher clickthrough rates.*

H4: *IRA advertisements exhibiting threatening language will be associated with higher clickthrough rates.*

Finally, the extent to which the observed clickthrough rates found in this data exceed commonly held benchmarks in the advertising industry is unknown. Given the aforementioned unusual, unprecedented nature of these advertisements, it remains worthwhile to assess and compare advertising effectiveness. To achieve this, this paper compares observed benchmarks on Facebook and Instagram to historical industry gold standards to ultimately assess the degree to which engagement of these advertisements deviated from the norm by asking:

RQ1: *Do IRA-sponsored ads receive higher clickthrough rates than typically observed for digital advertisements?*

Finally, it would stand to reason that as the IRA continued to launch paid Facebook and Instagram campaigns, they would in time improve in their ability to do so effectively. Given that clickthrough rate is a commonly tracked benchmark, it also stands to reason that it was a key performance indicator of campaign success and one that was likely monitored. For the reasons above, an analysis of clickthrough rate across time is necessary:

RQ2: *Do the clickthrough rates of IRA-sponsored ads improve across time?*

Method

The IRA Facebook and Instagram dataset was downloaded directly from the source. To download and view the raw data visit: https://democrats-intelligence.house.gov/social-media-content/. The data itself was roughly one gigabyte of Portable Document Formats (PDFs). Each file contained one advertisement.

Substantial data preparation efforts were necessary. As a first step, the data needed to be converted to a text format. The PDF files were prepared in a way that made text extraction difficult. For instance, if one were to highlight the text from an advertisement, copy it to the clipboard, and paste it to a text editor, the text would be incomplete and riddled with errors (e.g., missing spaces, text appearing out of order). To address this, the Python package *textract* (textract, 2018) was used to extract all text from the PDF files. String and regular expression matching were then used to extract specific fields out the ads, including the body copy (i.e., text) of an advertisement and its metadata. If an advertisement contained an image that had text inside of it, that text was included in the body copy section of the advertisement's metadata. This means that the text of images were included in the analysis, but the actual images were not.

Parsed data needed to be cleaned extensively. Commas were removed, dollar signs were removed, and all numerical values in the data needed to be inspected for non-alphanumeric characters (e.g., a phantom letter or period appearing in a numerical value due to text extraction errors). All other data was also converted to a tabular format and saved as a .csv file.

In addition to text extraction, we also extracted the following metadata for each ad: start date of the ad campaign, end date of the ad campaign, the landing page the advertisement led to if clicked, the number of impressions an ad received, the number of clicks an ad received, the amount spent on the ad, and if the campaign was run on Facebook or Instagram. All dollar amounts were converted to dollars. Ad spend was listed in Russian rubles for less than 1% of campaigns. For these cases, the currency exchange rate was used for the day in which the campaign was launched. In addition, targeting parameters were included for each ad and included interests, locations, "people who match" and so on. For a review of the various targeting parameters used see Bump (2018). In all, 1,438 unique targeting parameters were found in the data. Each parameter was given a column and was coded as "1" if that parameter was present in a given advertisement.

Notably, while a corpus *n* of 3,517 ads is commonly referred to in press releases (e.g., Penzenstadler et al., 2018), the researchers only found 2,552 files that had an associated spend amount greater than $0.00 (i.e., were actually *advertisements that ran*). Of those files, 36 cases did not have text, and thus could not be coded. Finally, in the case of one advertisement, we observed a higher number of clicks than impressions. Because this was likely due to an encoding error, we deleted the case in question when conducting the analyses related to research. In all, these processes left us with an analytic *n* of 2,515 advertisements.

Measures

Annotation of the advertisements. The body copy and text found in images for each advertisement were annotated for the presence of the aforementioned four types of negativity frequently found in the IRA ads. Google's Perspective Application Programming Interface (API) was used. For API documentation visit https://github.com/conversationai/perspectiveapi. The API serves deep neural network-based machine learning models developed with Google's TensorFlow, a set of libraries that support the development of advanced machine learning algorithms (Abadi, et al., 2016). The Perspective API is comprised of a series of algorithms designed to identify individual behaviors conceptually linked to toxic online conversation. The algorithms were themselves developed using the judgements of thousands of crowd workers on hundreds of thousands of comments. The algorithms have been tested across multiple domains, including the comments section of *The New York Times*, *Wikipedia's* 'Talk Pages', and most recently, on political talk from Twitter and Facebook (Hopp, Vargo, Dixon, & Thain, 2019). Here we used four specific Perspective API measures: identity-based insults, inflammatory language, obscene language, and threatening language. Identity-based insults referred to the degree to which anger, disgust, hatred, other negative emotions against a person or group based on identity attributes were featured in the ad. Inflammatory language referred to the use of language with an inferred intent to provoke or inflame. Obscene language referred to the use of crude and/or vulgar language (including, but not limited to cursing). Threatening language referred to the degree to which the ad employed verbal intention to inflict pain, injury, or violence against an individual or group. For each advertisement, the algorithm returned a value ranging from 0 to 1 for each attribute of interest. Values near 0 were indicative of a low probability that the advertisement contained the attribute of interest while values close to 1 indicated a high probability that the advertisement contained the attribute of interest.

Content Analysis. Each post was assigned four probability scores (ranging from 0-1) that described the probability, or likelihood, that the post contained the four incivility attributes of primary interest. To perform an external validity the toxicity API on our present data, a random sample of tweets was selected by probability. The random sample was stratified by social media platform (Instagram or Facebook), by binned probability range (50%, 75% and 90%) and by negative behavior type (identity attacks, inflammatory language, obscenities, and threats). If there were more than 20 ads for a particular platform, probability range and behavior type, a random sample of 20 was chosen. If less than 20 ads existed, all ads were coded. In all, 304 ads were annotated for their presence of the corresponding behavior. Two coders coded the sample independently according to the codebook found in Appendix A.

Agreement was assessed via the use of pairwise agreement and Gwet's AC1, a chance-adjusted measure where values closer to 1 suggest higher levels of agreement among the coders. Prior research shows that Gwet's AC1 outperforms more traditional chance-adjusted measures when feature prevalence is very low (Wongpakaran, Wongpakaran, Wedding, & Gwet, 2013). First, we compared the degree to which the human coders agreed with one another. Of the 500 decisions, the human coders disagreed XX times (pairwise agreement=XX.X%; Gwet's AC1=.XX). Disagreements were solved via a deliberation process and in all cases were resolved by a coder acknowledging that a incident was missed. Next, we assessed conformity between the human and computer-derived annotations. To a certain extent, these analyses were complicated by the fact that the computationally-derived annotations for each post were continuous values while the human coded outcomes were binary. In light of this challenge, we examined conformity between the human and computer coded samples using several different cut points for the continuous data. To start, we classified any post with a probability score >.50 on any of the four attributes as uncivil in nature. Here, pairwise agreement between the human and computer annotations was XX.X% (Gwet's AC1=.XX). Next, we classified any post with a computer-derived any attribute probability score >.75 as uncivil. Pairwise agreement between human and computer was XX.X% (Gwet's AC1=.XX). Finally, we categorized any post with averaged cross-attribute probability scores >.50 as uncivil in nature. Pairwise agreement between the human and computer-coded outcomes was XX.X% (Gwet's AC1 =.XX).

Control variables. Using the meta-data associated with each advertisement, we generated a number of control variables. First, using the raw dollar amount ($), we addressed the amount spent on placing the advertisement. Next, using the date the advertisement was posted, we created a binary variable that described whether or not the advertisement was posted before or after the 2016 presidential election. Third, we generated a binary variable that described if the advertisement appeared on Facebook or Instagram. Finally, we summed the number of targeting parameters that were placed on the advertising campaign. Targeting parameters can be thought of as the ways in which the audience of the ad was tailored. For instance, popular targeting parameters were geographic area (by both regions and states), interests (e.g., guns, southern pride) and demographics (e.g., age, sex, gender). The more targeting parameters, the smaller and more tailored an audience generally was.

Descriptive statistics for the variables of interest are provided in Table 1. Figure 2 shows a density plot for the clickthrough rate variable.

TABLE 1 ABOUT HERE

FIGURE 2 ABOUT HERE

Facebook and Instagram CTR benchmarks. To assess the degree to which Facebook and Instagram advertising campaigns are engaged with relative to the norm, or competition, it is common practice to benchmark key performance indicators (KPIs) of digital advertising campaigns (Burton, 2009). By comparing the average performance of an advertising platform against a specific campaign, marketers can evaluate success. eMarketer, a leading digital marketing market research company, publishes industry reports, comparative estimates, and charts related to digital advertising performance (eMarketer, n.d.). The most commonly reported KPI included in the IRA dataset when it comes to engagement was clicks.

Clicks are the only type of engagement presented in this Facebook dataset. Clicks illustrate an action on behalf of the consumer. Clicks include, but are not limited to times that audiences: click on a profile, click on an image (Facebook only), click to reveal comments, click to share a post, click to bookmark a post, click to like a post, click to react to a post (Facebook only), clicks to URLs linked in posts (Facebook only), and click to reveal tagged people on the post. As broadly construed as this variable is, clicks in this study can be thought of as encapsulating the majority ways in which a post can be engaged with.

This metric can be converted to a clickthrough rate by dividing it by the total number of advertising impressions. When reporting on Instagram and Facebook advertising clickthrough rates, eMarketer trusts Kinetic Social and the report they release on Instagram and Facebook KPIs (Kinetic, 2016). Given that Facebook and Instagram are both owned by Facebook and ads are purchased on the same demand side platform, metrics for Facebook and Instagram ads are most commonly conflated across platforms. According to Kinetic, the average clickthrough rate as of Quarter 4 of 2014 was 1.3%. The average clickthrough rate as of Quarter 4 of 2015 was 0.8%. Finally, in Quarter 4 of 2016, that number rose to 1.2%. We include these three-year measurements as they correspond to the years in which the IRA ads ran. However, because Facebook and Instagram do not officially release benchmarking data, it's hard to know the true accuracy of this data. Moreover, the tracked by Kinetic are not political. No known CTR benchmarks exist for political ads on Facebook or Instagram. Other benchmarking sites, such as MarketingProfs.com, show clickthrough rates ranging from 0.5% to 1.6% for Quarter 4 in 2016 depending on the industry, with a global average of 0.9% (Nanji, 2017). Regardless of the "true" number, no clickthrough rate found from various benchmarking sites was higher than 1.8% for the date range in question (2015 to early 2017; Nanigans, 2017).

Analytic Plan

As shown in Figure 2, the outcome variable of interest in Hypotheses 1 through 4 (clickthrough rate) was a continuous variable that featured a substantial degree of positive skew. Accordingly, to test the hypotheses, a gamma regression model with a log link function was employed. Gamma regression models are included in the family of generalized linear models and serve as an unbiased means of estimation in analytic scenarios where a criterion variable of interest is both continuous and substantially positively skewed (see Fox, 2015). The statistical model contained the four independent variables of interest (scores representing the degree to which each advertisement featured identity-based insults, inflammatory language, obscene language, and threatening language). Also included in the model were control variables representing total spend amount, whether or not the ad appeared before the 2016 presidential election, whether the advertisement appeared on Facebook or Instagram, and the total number of targeting parameters. The current data had several advertisements with an engagement rate of 0.0%. Because the gamma distribution technically requires positive integers, we added a small constant (10^{-5}, or 0.00001) to the outcome variable. Robust standard errors (HC3) were employed. For the model, we report both the logged coefficients (*b*) and the exponentiated coefficients ([*exp(b)*]). The latter coefficient is a multiplicative term that describes the degree of change in the criterion variable given a 1-unit change in the independent variable. All modeling analyses were conducted in the *R* statistical computing environment.

Time Series Analysis using Facebook's Prophet

The second research question put forward by this paper is rooted in change across time. For the clickthrough rate measure, we ask whether change across time is significant. First, a weekly average clickthrough rate was calculated by averaging all campaign clickthrough rates on a given day, and then creating an average for that week with weeks beginning on Mondays. To answer these questions we adopt Facebook Prophet, a semi-supervised modular regression model for time series analysis (Taylor & Letham, 2018). The software can be implemented as a package in Python. Prophet was adopted for its robust ability to automatically model seasonality using a residual optimization approach, and for its overall ability to consistently fit models with smaller residuals when compared to more traditional approaches such as the Autoregressive Integrated Moving Averages approach (ARIMA) (Vargo, Basilaia, and Shaw, 2015). For all models, yearly seasonality was modeled and, as a result, removed from the change parameter of the model. Removing seasonality protects from finding spurious results in general trends that could be better attributed to annual seasonality, such as election season.

Research question 1 was addressed using the three previously identified benchmarks: 0.9% (Nanji, 2017), 1.6% (Nanji, 2017), and 1.8% (Nanigans, 2017). These benchmarks represented, roughly speaking, low, moderate, and high clickthrough rates for digital advertisements. We then tested the degree to which our data varied from prior rates using a series of one-sample Wilcoxon tests. These tests, essentially, are non-parametric alternatives to one-sample t-tests and test whether the median value in an observed sample is different from fixed standard value.

Results

Hypothesis 1 suggested that advertisements with heightened levels of identity-based insults would be associated with higher clickthrough rates. Contrary to expectations, however, we observed a negative relationship between the degree to which an advertisement employed identity-based insults and clickthrough rate, b = -0.42, se_r =0.08, p < .001, 95% CI [-0.57, -0.28]; exp(b) = 0.65. Hypothesis 1 was not supported.

Hypothesis 2 predicted that inflammatory language would be positively associated with clickthrough rate. This hypothesis was supported: $b = 0.40$, $se_r = 0.08$, $p < .001$, 95% CI [0.23, 0.56]; $\exp(b) = 1.49$.

Hypothesis 3 suggested that advertisements featuring a high degree of obscene language would be subject to higher clickthrough rates. Like Hypothesis 2, Hypothesis 3 was supported as we observed a positive relationship between clickthrough rate and obscene language score, $b = 0.23$, $se_r = 0.07$, $p < .01$, 95% CI [0.10, 0.37]; $\exp(b) = 1.26$.

Finally, Hypothesis 4 predicted that threatening language would be positively associated with clickthrough rate. This contention was supported: $b = 0.17$, $se_r = 0.08$, $p < .05$, 95% CI [0.01, 0.33]; $\exp(b) = 1.19$.

A full report of the model used to assess Hypotheses 1 – 4 is provided in Table 2.

TABLE 2 ABOUT HERE

Turning next to the research question, we investigated the degree to which the observed clickthrough rate differed from general benchmarks. Summed across the sample, the advertisements collectively were associated with 40,467,355 impressions which elicited 3,723,307 clicks, resulting in an overall clickthrough rate of 9.2% for the sample as a whole. Examination of the clickthrough rates associated with each advertisement indicated mean and median clickthrough rates of 10.4% (observed range = 0.0%, 84.4%). One-sample Wilcoxon tests indicated that the median observed clickthrough rate was significantly higher than the established benchmark rates of 0.9%, 1.6%, and 1.8% (all *p*-values < .001).

Figure 3 & Figure 4 Here

In reviewing the time series, we can see that weekly average clickthrough rate appears to fluctuate from week to week. Still, a reasonable times series model can be fit. From observation to fit trend, the average weekly average error is only 2.2% falling at approximately half of one standard deviation of the clickthrough rate variable (*SD* = 4.28%, M = 9.12%). Inspecting the first plot in Figure 4, the general trend component of this model reveals a clear positive linear relationship with time and clickthrough rate. From inception to the last campaign the general trend component rises from 7.32% to 10.71% a percent change of 46.31%.

Discussion

This study analyzed the degree to which negative content in IRA-linked Facebook and Instagram ads was associated with audience engagement. The results suggested advertisements employing inflammatory language, obscene language, and threatening language had higher clickthrough rates. Contrary to expectations, we found that ads employing identity-based insults were less likely, all things considered, to be clicked on. Finally, the data indicated that the analyzed IRA ads were associated with a clickthrough rate that was significantly higher than several fairly well-known platform benchmarks. The implications of these findings are discussed below.

This work provides tentative evidence that the stimulatory effects of negative advertising (e.g., Hopp & Vargo, 2017) may be transferable to social media behavior. Specifically, the findings observed here conform to prior work which has shown that negative political advertisements motivate attention and interest in political events (e.g., Marcus, 2000; Martin, 2004; Pratto & John, 1991). In the present case, we found that advertisements featuring inflammatory language, obscene language, and threatening language saw heightened click activity. Drawing from our theoretical rationale, we surmise that these advertisements effectively invoked negative affective states, and that these affective responses, subsequently, signaled that something important was at stake, ultimately resulting in a stimulatory effect (Goldstein & Freeman, 2002; Hopp & Vargo, 2017; Lau et al, 1999).

Notably, this study is relatively unique in terms of its simultaneous focus on different textual means of communicating negativity in political advertising. Most prior studies have either focused on a specific articulation of negativity (e.g., stereotypical renderings of minority groups; Matthes & Schmuck, 2017) or classified negative advertisements on a binary basis (e.g., Wang, Morey, & Srivastava, 2014). The results of this study suggest that different ways of communicating negativity may be differentially associated with advertising engagement. As shown in Table 2 (and discussed in greater depth below), identity-based insults were associated with user engagement in a negative manner. In contrast, inflammatory language, obscene language, and threatening language were all positively related to engagement. Among those attributes positively associated with engagement, there was a disproportionality of effect strength: we observed, for instance, that the relationship between inflammatory language and engagement was substantially stronger than the relationship between threatening language and clickthrough rate.

We also found that IRA-linked advertisements, were, broadly speaking, engaged with at a higher rate than typical advertising on Facebook and Instagram. Here we showcase the heightened engagement audiences have with microtargeted political, social media advertisements online. We echo the findings of others that have suggested that data-driven political Internet advertising is powerful (e.g., Kreiss, 2016). While it’s beyond the data here to know if social media advertising had a greater effect than traditional advertising, we can surmise as others have that it was more effective and efficient than traditional advertising in the case of the 2016 presidential race (e.g., Fowler, Ridout, & Franz, 2016).

However, contrary to expectation, advertisements employing identity-based insulting language received lower levels of audience engagement. Several factors may explain this finding. First, research has shown that negative, identity-based campaign appeals are subject to moderating factors. For instance, Matthes and Schmuck (2017) found that populist advertising, which often relies on negative group-based stereotypes, was generally only effective among those with lower levels of education attainment. Thus, in some cases, it could have been the case that certain types of people (e.g., those with moderate to high levels of education) were highly unlikely to engage with the advertisements, resulting in lowered average levels of engagement. Another possible explanation could be that some identity-based advertisements came off as clownish, clumsy, and, ultimately, distasteful. Figure 1 shows advertisements that openly threaten all Obama voters, something likely to be distasteful to those with bipartisan friends. Moreover, messages have improper English and often come from offbeat Instagram accounts such as "_born__black."

Notably, this study used an algorithmic approach to classify advertising content. Such an approach offers a powerful means of assessing a large quantity of data in various different ways. In all, 10,060 annotations were made using Google's Perspective API. However, such capabilities come at the cost of assessing nuance. Separately, prior research on "dog whistle politics" (e.g., López, 2017) suggests that identity-based attacks are most effective when they are coded (i.e., apparent in nuanced social ways). It may, then, be the case that the algorithm employed in this study was unable to fully account for coded applications of divisive identity-based language, and therefore was unable to accurately assess the relationship between the identity-linked insults and user advertising engagement.

What remains to be studied is exactly why this particular type of divisive political content struck a chord with audiences. In studying our data, it became quite apparent that advertisements were well-tailored to their audiences. The IRA-content was clearly microtargeted. Many examples exist in the data: advertisements that targeted African Americans who cared deeply about racial equality supported the continued fight for African American equality; advertisements that targeted Republicans in the South focused on themes of pride in heritage; advertisements aimed at gun owners called out for the preservation of the 2nd Amendment. This signals what others in the campaign have suggested: political advertising works best with niche audiences and highly-relevant customized messages (Stahl, 2017). As Trump's digital advisor himself said, "It was voters in the Rust Belt that cared about their roads being rebuilt, their highways, their bridges. They felt like the world was crumbling. So, I started making ads that would show the bridge crumbling" (Pramuk, 2017, para. 10). He added, "you know, that's microtargeting them" (Pramuk, 2017, para. 10). Our data here also show a positive relationship between the sheer number of targeting parameters used and the CTR. That is, the more tailored and specifically defined an ad was for an audience, the more engagement that advertisement received.

We present evidence here that campaign clickthrough rates improved across time. The IRA appears to have consistently improved in getting audiences to click on their content. Taken in context of an organization, the IRA appears to have been monitoring engagement, and actively altering tactics to bolster it. This furthers evidence that the Russian campaign was highly tactical, and successful at creating content that compelled audiences to respond.

Given the operationalization of clicks here, we can think of this study as showing that audiences engaged with ads in a variety of ways more oven when ads were negative. The use of this engagement metric, clickthrough rate, is an umbrella metric for engagement. Unfortunately, the only two pieces of outcome data Facebook provided was impressions and clicks. While we leverage both here, the study of specific behaviors, such as commenting behavior, may reveal more nuanced results.

Finally, the methodology presented here, specifically the process of running the body copy and image text of ads through machine learning algorithms designed to detect hate, insults and threats, worked better than expected. The advertisements found in Figure 1 were identified by simply reviewing the top Google Perspective scores for each respective category. In qualitatively looking at hundreds of ads and their corresponding scores, we conclude that the tool, while not perfect, provides a straightforward way in which advertising platforms such as Facebook can easily screen for and capture negative content *before the advertisements run on the platform.* Current methodologies rely on audience feedback (e.g., consumers reporting content), and algorithms not specifically designed obscenity and inflammatory language. Taken together, this is problematic. This paper illustrates the damage, in the form of substantially high audience engagement, that is done when these ads are allowed to run on platforms. Even when ads are live for minutes, substantial amounts of clicks can occur. If a proactive screening tool was created that flagged ads with high scores and those ads were manually reviewed by Facebook moderators, the amount of engagement for these ads could be drastically reduced while drastically reducing the sheer number of advertisements needing to be manually reviewed.

This study is limited in several regards, beyond the aforementioned limitations of Google's Perspective API. First, we only looked at Russian-linked, IRA advertisements. As such, they are not representative of all political ads found on social media. Twitter data, while included in related congressional hearings, were not analyzed here due to the fact that those messages were organic, and not paid advertisements. Finally, it is likely that the ads studied here possessed other negative characteristics that may have had further impact on engagement, such as negative imagery in the actual images used themselves. Instagram, for instance, is a particular visual medium, and while the text inside messages were included, the actual depictions themselves likely impacted engagement. It is ripe for future work to create a full typology of the various uncivil behaviors that negative social media advertisements may possess.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Zheng, X. (2016, November). TensorFlow: A system for large-scale machine learning. In K. Keeton & T. Roscoe (Eds.), *Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation* (pp. 265-283).

Ansolabehere, S., & Iyengar, S. (1996). *Going negative: How political ads shrink and polarize the electorate.* New York, NY: Free Press.

Berelson, B. B., Lazarsfeld, P. F., & McPhee, W. N. (1954). *A study of the opinion formation in a presidential campaign.* Chicago, IL: University of Chicago Press.

Boland, P. (2018, July 26). Exposing Russia's effort to sow discord online: The Internet Research Agency and Advertisements. Retrieved from https://democrats-intelligence.house.gov/social-media-content/

Brians, C. L., & Wattenberg, M. P. (1996). Campaign issue knowledge and salience: Comparing reception from TV commercials, TV news, and newspapers. *American Journal of Political Science, 40,* 172-193. https://doi.org/10.2307/2111699

Bump, P. (2018, May 10). These are the most popular stealth Russian Facebook ads from each month. Retrieved from https://www.washingtonpost.com/news/politics/wp/2018/05/10/these-are-the-most-popular-russian-facebook-ads-from-each-month/?utm_term=.55c4d017389e

Burton, J. (2009). A marketer’s guide to understanding the economics of digital compared to traditional advertising and media services. Retrieved from https://ams.aaaa.org/eweb/upload/catalog/pdfs/mg18.pdf

Diamond, E., & Bates, S. (1988). *The spot: The rise of political advertising on television* (rev. ed). Cambridge, MA: MIT Press.

Dowling, C. M., & Wichowsky, A. (2015). Attacks without consequence? Candidates, parties, groups, and the changing face of negative advertising. *American Journal of Political Science*, *59*, 19-36. https://doi.org/10.1111/ajps.12094

Dumenco, S. (2016, October 28). State-by-state map: How the Clinton and Trump camps spent $595 million on TV, radio ads. https://adage.com/article/campaign-trail/how-clinton-trump-camps-spent-595-million-TV-radio/306496/

Edgerton, A., & Frier, S. (2018, May 10). Lawmakers release trove of more than 3,500 Russian-backed Facebook ads. Retrieved from https://www.bloomberg.com/news/articles/2018-05-10/trove-of-russian-backed-facebook-ads-show-2016-strategy-at-work

eMarketer. (n.d.). About. Retrieved from https://www.emarketer.com/about

Facebook Newsroom. (2018, May 10). Russian ads released by Congress. Retrieved from https://newsroom.fb.com/news/2018/05/russian-ads-released-by-congress/

Fandos, N., Kang, C., & Isaac, M. (2017, November 1). House intelligence committee releases incendiary Russian social media ads. *New York Times*. Retrieved from https://www.nytimes.com/2017/11/01/us/politics/russia-technology-facebook.html

Fowler, E. F., & Ridout, T. N. (2012). Negative, angry, and ubiquitous: Political advertising in 2012. *The Forum*, *10*, 51-61. https://doi.org/10.1515/forum-2013-0004

Fowler, E. F., Ridout, T. N., & Franz, M. M. (2016). Political advertising in 2016: The presidential election as outlier? *The Forum*, *14*, 445-469. https://doi.org/10.1515/for-2016-0040

Fox, J. (2015). *Applied regression analysis and generalized linear models*. Sage Publications.

Goldstein, K., & Freedman, P. (2002). Campaign advertising and voter turnout: New evidence for a stimulation effect. *Journal of Politics*, *64*(3), 721-740.

Gollust, S. E., Barry, C. L., Niederdeppe, J., Baum, L., & Fowler, E. F. (2014). First impressions: Geographic variation in media messages during the first phase of ACA implementation. *Journal of Health Politics, Policy and Law*, *39*, 1253-1262. https://doi.org/10.1215/03616878-2813756

Ha, A. (2018). House Democrats release more than 3,500 Russian Facebook ads. Retrieved from https://techcrunch.com/2018/05/10/house-democrats-facebook-russia/

Hmielowski, J. D., Hutchens, M. J., & Cicchirillo, V. J. (2014). Living in an age of online incivility: Examining the conditional indirect effects of online discussion on political flaming. *Information, Communication, & Society*, *17*, 1196-1211. https://doi.org/10.1080/1369118X.2014.899609

Hopp, T., & Vargo, C. J. (2017). Does negative campaign advertising stimulate uncivil communication on social media? Measuring audience response using big data. *Computers in Human Behavior*, *68*, 368-377. https://doi.org/10.1016/j.chb.2016.11.034

Hopp, T., Vargo, C., Dixon, L., & Thain, N. (2019). Correlating self-report and trace data measures of incivility: A proof of concept. *Social Science Computer Review*. Preprint available online. https://doi.org/10.1177/0894439318814241

Jamieson, K. H. (1993). *Dirty politics: Deception, distraction, and democracy*. Oxford, United Kingdom: Oxford University Press on Demand.

Johnson-Cartee, K. S., & Copeland, G. (2013). *Negative political advertising: Coming of age*. New York, NY: Routledge.

Kinetic Social. (2016). Q4 2016 social trends report. Retrieved from http://arbetsenhet.com/_assets/trends-report.pdf

Kreiss, D. (2016). *Prototype politics: Technology-intensive campaigning and the data of democracy*. Oxford, United Kingdom: Oxford University Press.

Lau, R. R., Sigelman, L., Heldman, C., & Babbitt, P. (1999). The effects of negative political advertisements: A meta-analytic assessment. *American Political Science Review*, *93*, 851-875. https://doi.org/10.2307/2586117

Taylor, S. J. & Letham, B. (2017, February 23). Prophet: forecasting at scale. *Facebook Research.* Retrieved from: https://research.fb.com/prophet-forecasting-at-scale/

López, I. H. (2017). *Dog whistle politics: How coded appeals have reinvented racism and wrecked the middle class.* New York, NY: Oxford University Press.

Marcus, G. E. (2000) Emotions in politics. *Annual Review of Political Science, 3,* 221-250. https://doi.org/10.1146/annurev.polisci.3.1.221

Martin, P. S. (2004). Inside the black box of negative campaign effects: Three reasons why negative campaigns mobilize. *Political Psychology, 25*, 545-562. https://doi.org/10.1111/j.1467-9221.2004.00386.x

Matthes, J., & Schmuck, D. (2017). The effects of anti-immigrant right-wing populist ads on implicit and explicit attitudes: A moderated mediation model. *Communication Research, 44,* 558-581. https://doi.org/10.1177/0093650215577859

Nanigans. (2017). Instagram advertising performance snapshot February 2017. Retrieved from https://cdn2.hubspot.net/hubfs/234606/Nanigans%20-%20Instagram%20Advertising%20Performance%20Snapshot%20-%20February%202017.pdf?utm_campaign=GA-Report-Instagram+Advertising+Performance+Snapshot+-+February+2017

Nanji, A. (2017, March 28). Facebook advertising benchmarks for 18 industries. Retrieved from https://www.marketingprofs.com/charts/2017/31801/facebook-advertising-benchmarks-for-18-industries-infographic

Penzenstadler, N., Heath, B., & Guynn, J. (2018, May 11). We read every one of the 3,517 Facebook ads bought by Russians. Here's what we found. Retrieved from https://www.usatoday.com/story/news/2018/05/11/what-we-found-facebook-ads-russians-accused-election-meddling/602319002/

Popken, B., & Cobiella, K. (2017, November 16). Russian troll describes work in the misinformation factory. Retrieved from https://www.nbcnews.com/news/all/russian-troll-describes-work-infamous-misinformation-factory-n821486

Pramuk, J. (2017, October 9). Trump's digital director explains how he used Facebook to help win the White House. Retrieved from https://www.cnbc.com/2017/10/09/brad-parscale-says-trump-campaign-used-facebook-to-beat-clinton.html

Pratto, F., & John, O. P. (1991). Automatic vigilance: The attention-grabbing power of negative social information. *Journal of Personality and Social Psychology, 61*(3), 380-391. https://doi.org/10.1037//0022-3514.61.3.380

Settle, J. E., Bond, R. M., Coviello, L., Fariss, C. J., Fowler, H. J., & Jones, J. J. (2015). From posting to voting: The effects of political competition on online political engagement. *Political Science Research and Methods, 4*, 361-378. https://doi.org/10.1017/psrm.2015.1

Sobieraj, S., & Berry, J. M. (2011). From incivility to outrage: Political discourse in blogs, talk radio, and cable news. *Political Communication, 28*, 19-41. https://doi.org/10.1080/10584609.2010.542360

Stahl, L. (2017, October 8). *How Facebook ads helped elect Trump* [Video file]. Retrieved from https://www.cbs.com/shows/60_minutes/video/Rv8I9Aw4SlKR0O3LDiW6F4kUsxEaygeF/how-facebook-ads-helped-elect-trump/

Textract. (n.d.) Read the docs. Retrieved September 24, 2018, from https://textract.readthedocs.io/en/stable/.

Vargo, C. J., Basilaia, E., & Shaw, D. L. (2015). Event versus issue: Twitter reflections of major news, a case study. In *Communication and information technologies annual* (pp. 215-239). Emerald Group Publishing Limited.

Wang, S. (2018, April 10). Twitter follows Facebook in endorsing the Honest Ads Act. Retrieved from https://adage.com/article/digital/twitter-facebook-endorsing-honest-ads-act/313056/

Wang, Z., Morey, A. C., & Srivastava, J. (2014). Motivated selective attention during political ad processing: The dynamic interplay between emotional ad content and candidate evaluation. *Communication Research, 41*, 119-156. https://doi.org/10.1177/0093650212441793

Zadrozny, B., Popken, B., Collins, B., & Lehren, A. (2018, May 10). Congress releases 3,000 Russian ads — including some that targeted fans of Fox News. Retrieved from https://www.nbcnews.com/tech/tech-news/sean-hannity-black-lives-matter-among-targets-russian-influence-campaign-n872926

Table 1

Descriptive Statistics for Continuous Variables of Interest

Variable Name	**M**	**SD**	**Minimum Value**	**Maximum Value**
Clickthrough Rate (%)	10.4	7.1	0.00	84.4
Spend ($)	2,324.76	10,585.79	0.05	3,31675.75
Number of Targeting Parameters	9.57	4.07	3.00	53.00
Identity-Hate Toxicity Score	0.36	0.26	0.01	0.96
Inflammatory Language Toxicity Score	0.41	0.24	0.00	0.91
Obscene Language Toxicity Score	0.14	0.20	0.00	0.99
Threatening Language Toxicity Score	0.28	0.19	0.01	0.95

Note. 57.9% of advertisements had a campaign start date before the 2016 presidential election; 96.3% of campaigns were conducted on Facebook while 3.7% of campaigns were Instagram-based

Table 2

Results of a Gamma Regression Model Predicting Clickthrough Rate for IRA Advertisements

	b	**se (Robust)**	***p*-Value**	***b* LL 95%CI**	***b* UL 95%CI**	**exp(*b*)**	**exp(*b*) LL 95%CI**	**exp(*b*) UL 95%CI**	**VIF**
Spend	0.00	0.00	>.05	0.00	0.00	1.00	1.00	1.00	1.01
Before 2018 Election (1 = Yes)	-0.30	0.04	***	-0.36	-0.23	0.74	0.69	0.80	1.29
Platform (1 = Facebook)	2.61	0.16	***	2.30	2.91	13.55	10.00	18.36	1.07
No. Target Parameters	-0.03	0.01	**	-0.04	-0.01	0.97	0.96	0.99	1.18
Identity-Based Insult	-0.42	0.08	***	-0.57	-0.28	0.65	0.56	0.76	2.14
Inflammatory Language	0.40	0.08	***	0.23	0.56	1.49	1.26	1.75	2.01
Obscene Language	0.23	0.07	***	0.10	0.37	1.26	1.10	1.44	1.13
Threatening Language	0.17	0.08	*	0.01	0.33	1.18	1.01	1.39	1.35

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

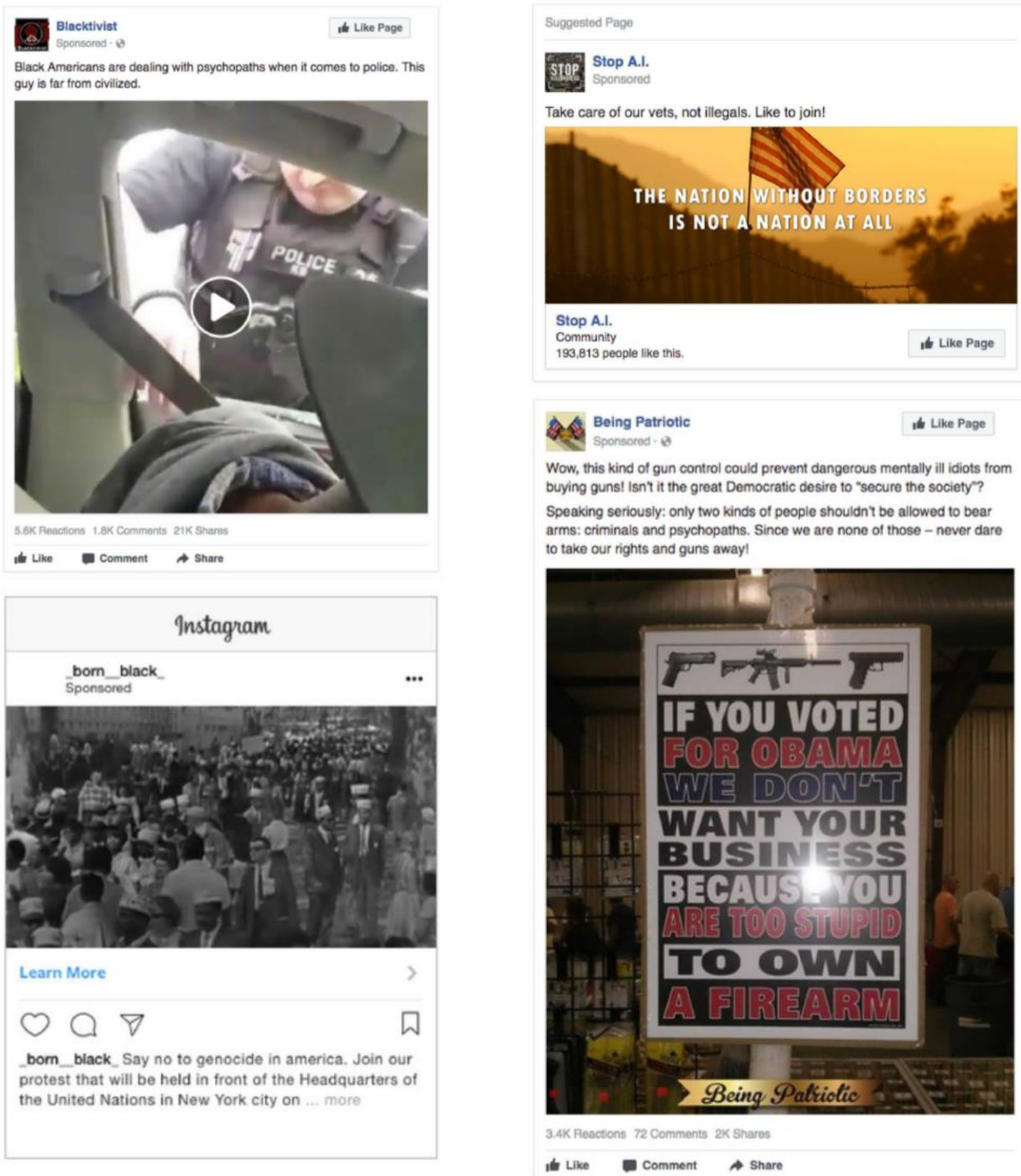


Figure 1. Example IRA Facebook and Instagram advertisements

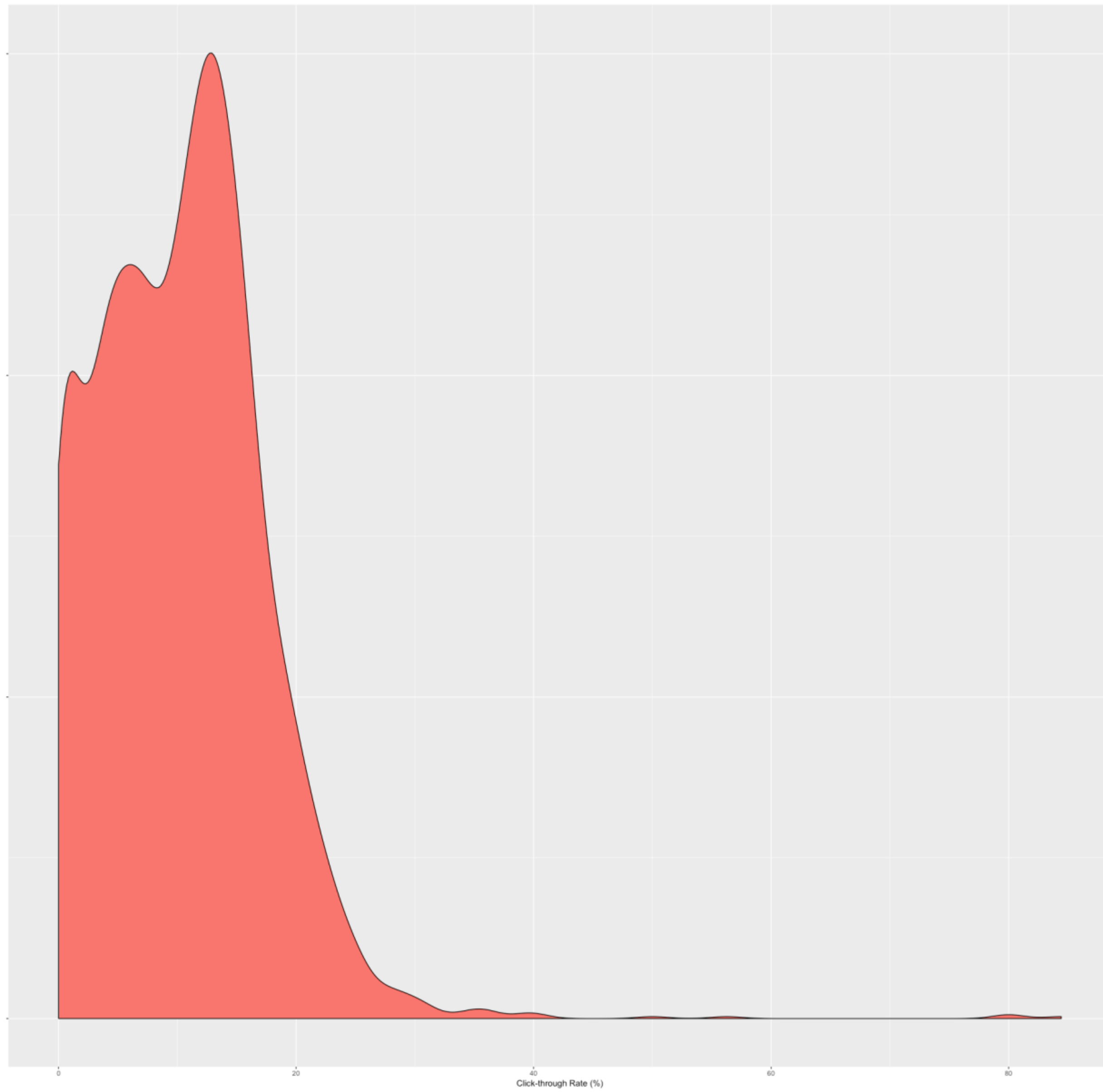


Figure 2. Density plot for advertisement clickthrough rate variable.

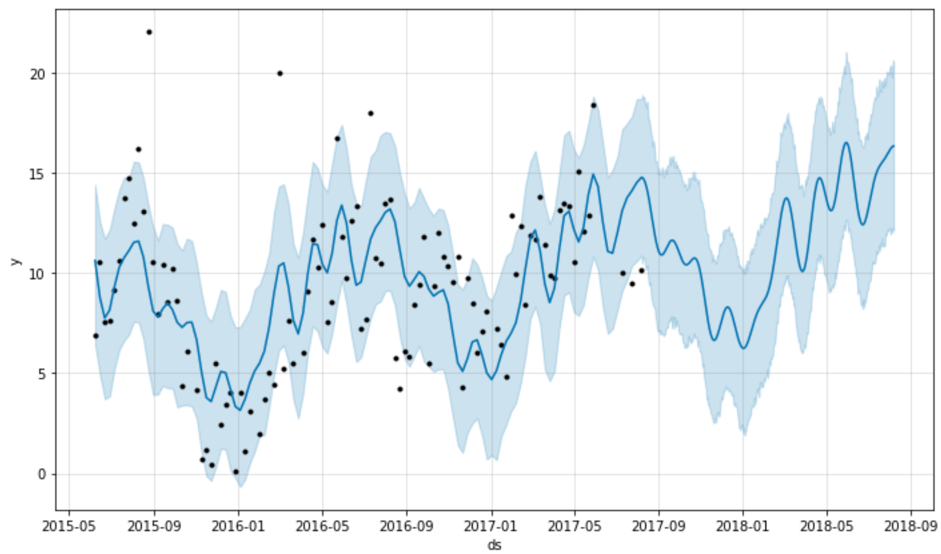
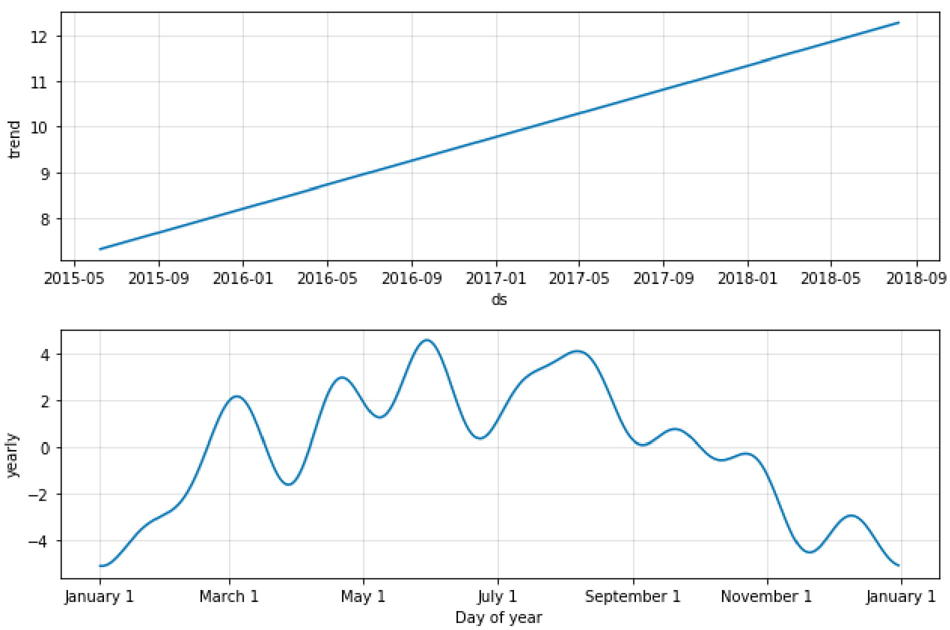


Figure 3. Time series of daily average clickthrough rate. Uncertainty in trend at 80% represented by light blue error bars

Figure 4. The trend and seasonality components of the time series.



Appendix A: Codebook and Conceptual Definitions

Code attributes as "1" if present, "0" if absent. Follow the following definitions:

Threats – If a potential threat is exposed to the audience, then this attribute is present. For instance, warnings of police violence or warnings of the risks associated with illegal immigration are all examples of threats to the audience. Remember that these threats will be suggested or implied, not directly posed to the audience.

Obscene – If a post contains any vulgar language, including any known "swear" words, no matter how mild, including "hell" or "damn," code this attribute as present. Even quoted language should be considered. Identity labels used in positive ways (e.g., gay, or Muslim) are not obscene. Attempts to sensor swear words (e.g., bullsh*t) still count as obscenity.

Inflammatory – Language geared towards inflaming audiences to be upset and/or to take action regarding a political issue. Primary devices used to inflame audiences are used of disparity and/or fear.

Identity Attack – An identity (race, sexual orientation or gender) is called out and labeled with a negative connotation. Positive mentions of identity are not considered identity attacks (e.g., celebrating Gay Pride).