Trumping Hate on Twitter?

Online Hate in the 2016 US Election and its Aftermath[*]

Alexandra Siegel[†] Evgenii Nikitin[‡] Pablo Barberá[§] Joanna Sterling[¶] Bethany Pullen[‖]
Richard Bonneau[**] Jonathan Nagler[††] and Joshua A. Tucker[‡‡]

March 6, 2019

Abstract

To what extent did online hate speech and white nationalist rhetoric on Twitter increase over the course of Donald Trump's 2016 presidential election campaign and its aftermath? The prevailing narrative suggests that Trump's political rise—and his unexpected victory—lent legitimacy to and popularized bigoted rhetoric that was once relegated to the dark corners of the Internet. However, our analysis of over 750 million tweets related to the election, in addition to almost 400 million tweets from a random sample of American Twitter users, provides systematic evidence that hate speech did not increase on Twitter over this period. Using both machine-learning-augmented dictionary-based methods and a novel classification approach leveraging data from Reddit communities associated with the alt-right movement, we observe no persistent increase in hate speech or white nationalist language either over the course of the campaign or in the aftermath of Trump's election. While key campaign events and policy announcements produced brief spikes in hateful language, these bursts quickly dissipated. Overall we find no empirical support for the proposition that Trump's divisive campaign or election increased hate speech on Twitter.

[*]The authors gratefully acknowledge the financial support for the NYU Social Media and Political Participation (SMaPP) lab from the INSPIRE program of the National Science Foundation (Award SES-1248077), the William and Flora Hewlett Foundation, the Rita Allen Foundation, the Knight Foundation, the Bill and Melinda Gates Foundation, Craig Newmark Philanthropies, the Democracy Fund, the Intel Corporation, the New York University Global Institute for Advanced Study, and Dean Thomas Carews Research Investment Fund at New York University. We thank Sean Kates for his feedback in designing our coding scheme, NYU Undergraduate SMaPP Research Assistants for their coding work, and Yvan Scher and Leon Yin for programming support. A.S. and J.T. designed the research plan and outline for the paper. A.S. conducted the statistical analysis and wrote the first draft of the paper. A.S, J.S., and B.P. designed and implemented the dictionary-based coding method. E.N. designed and conducted the non-dictionary based analysis and wrote the corresponding section of the Supplementary Materials. P.B., R.B., and J.N. contributed to the data collection and design of the analytic tools. J.T. revised the original draft, and all of the authors contributed to the editing of the text.

[†]Stanford University, Immigration Policy Lab, and NYU SMaPP laboratory.

[‡]New York University, Department of Politics, and NYU SMaPP laboratory.

[§]London School of Economics, Methodology Department

[¶]Princeton University, Department of Psychology, and NYU SMaPP laboratory.

[‖]NYU SMaPP laboratory.

[**]New York University, Department of Biology, NYU SMaPP laboratory, and Simons Foundation.

[††]New York University, Department of Politics, and NYU SMaPP laboratory.

[‡‡]New York University, Department of Politics, and NYU SMaPP laboratory.

1 Introduction

From his calls for a "Muslim ban" to his retweets of white nationalist content and leaked tapes endorsing sexual assault, experts and casual observers alike have warned that Donald Trump's divisive 2016 campaign and election fanned the flames of bigotry across the United States (Camp 2016). Increased reports of hate crimes and vandalism targeting religious and racial minorities, as well as a notable rise in the number of active American hate groups, caused 2016 to be described as an "unprecedented year for hate"(SPLC 2017). Citing a "massive rise" in online hate speech following Trump's election, reports by civil rights organizations suggest that Trump's campaign legitimized extremist ideologies, bringing hostile messages that were once relegated to the dark corners of the Internet into the mainstream (Guynn 2016; ADL 2017*a*). Fearing that Trump's election created a new "safe space for hate," academics, journalists, policy makers, and everyday citizens have increasingly voiced concern about the consequences of Trump's actions and rhetoric both on and offline (Ott 2017; Milligan 2017). As a result, proposals for regulation to control online hate speech, both in the United States and abroad, have become increasingly common.[1] Analyzing the impact of Trump's Twitter rhetoric on public discourse, a 2017 study concluded that "Trump's simple, impulsive, and uncivil Tweets do more than merely reflect sexism, racism, homophobia, and xenophobia; they spread those ideologies like a social cancer...His Tweets teach us to see others as less-than-human and they inspire hatred and violence"(Ott 2017).

However, despite a wealth of anecdotal and small-scale empirical evidence of this "Trump effect," little is known about how the quantity of online hate speech, or the number of individuals producing it, changed over the course of Trump's campaign or in the aftermath of his election.[2] Here, we take a first step at addressing this gap in knowledge about the link between Trump's political rise and the use of hate speech online. More specifically, using Twitter collections of over 150 million tweets referencing Hillary Clinton, over 600 million tweets referencing Donald Trump, and over 400 million tweets collected from a random sample of 500,000 American Twitter users, we systematically investigate the degree to which the quantity of hateful tweets and size of the population producing them increased over the course of Trump's campaign or following his election on November 8, 2016.

We identify hate speech and white nationalist language using two methods: a machine-learning-augmented dictionary-based approach; and a non-dictionary-based method harnessing large naturally annotated corpora of text containing hate speech and white nationalist language from alt-right subreddits.[3] Using two different sources of data and these two different measurement strategies, we find—in contrast to the conventional wisdom—no persistent increase in hate speech or white nationalist language either over the course of the campaign or in the aftermath of Trump's election on Twitter, the president's preferred social media platform. While key events including terror attacks or Trump's travel ban announcement produce temporary upticks in hateful rhetoric—and these bursts of hate speech are

[1] See for example Rainie, Anderson and Albright (2017), Financial Times (2017), and Marwick (2017).

[2] Indeed, a parallel concern—about the rise of bots as a vehicle for spreading online hate—might in fact suggest that current estimates of the number of *people* producing hate speech on Twitter could be biased upwards.

[3] Subreddits are anonymous online forums dedicated to discussing specific topics on Reddit, a popular social news aggregation, web content rating, discussion forum, and social media platform. This method is explained in detail in Section 5 as well in Section A.2 in the online appendix.

not inconsequential—they are not indicative of a systematic increase of hate speech in the American Twittersphere over the course of the 2016 US Presidential election campaign and its aftermath, and appear similar to the bursts of such rhetoric regularly observed in the pre-Trump era.

Our study therefore highlights the need to move beyond conventional narratives and small-scale empirical analysis to more systematically assess trends in online discourse and behavior over time. Trump's campaign and election undoubtedly drew a great deal of attention to online hate speech and white nationalist rhetoric, but at least on Twitter, such language did not become more common or popular. The extent to which this trend may have materialized in other online and offline forums—as well as in other parts of the Twittersphere not captured by our analysis—importantly is beyond the scope of this paper and remains a subject for future research. We offer many caveats in our conclusion about what our findings do *not* imply. But we do prove an important empirical point about the frequency of hate speech in one of the major online forums frequented by a large cross-section of Americans.

Beyond our empirical findings, we also hope to make a more general contribution to the development of methods for understanding the evolution of speech in the digital era. As the role of social media platforms in fostering extremism and offline violence has come under scrutiny, online hate speech has received increased attention from academics and policy makers alike. But despite a growing body of research devoted to defining and detecting online hate speech, the existing scientific literature lacks a systematic framework for assessing how the volume and content of these harmful messages change over time. We therefore hope the research design applied here—relying on multiple sources of data and utilizing multiple methods to analyze online speech changes over time—might be used by other researchers to systematically evaluate the real-time dynamics of online hate speech and other discourse in diverse contexts.

The rest of this paper is structured as follows: in Section 2 we introduce our motivation and hypotheses; in Section 3 we describe our data and methods for our primary dictionary-based analysis; in Section 4 we outline our empirical strategy and results from the primary analysis; in Section 5 we present robustness tests of our primary findings utilizing our alternative reference-text based approach; and Section 6 offers conclusions, important caveats, and steps for future research.

2 Motivation and Expectations

Beginning in the early days of his campaign, Donald Trump's rise as a mainstream political candidate was marked by unusually blunt bigoted and racist statements. From his assertion in July 2015 that Mexican immigrants were criminals and rapists (Walker 2015) to his repeated derogatory remarks against Muslims including phrases like "Islam hates us," his campaign rhetoric directed at minority groups and immigrants was far from subtle (Filimon 2016). His courting of support from anti-LGBT groups stoked fears in the LGBT community (Stack 2016), and his use of misogynistic rhetoric, including repeated blaming of sexual assault victims and a leaked tape in which he advocated groping women against their will, highlighted his lack of respect for women (Cohen 2016). His refusal to denounce the support of former Ku Klux Klan leader David Duke, not to mention his post-election appointment of Steve Bannon, the self-proclaimed creator of a "platform for the alt right," lent further credence to the view that the Trump administration tolerated, if not condoned, extremist white nationalist ideology (Huber 2016).

During the campaign, round-the-clock traditional and social media coverage of Donald Trump were littered with examples of his prejudicial rhetoric and behavior. Trump's prolific Twitter use far surpassed that of all other candidates and became "a tool of political promotion, distraction, score-settling and attack"(Barbaro 2015). Many of Trump's most inflammatory statements during the campaign were 140 character statements disseminated through his Twitter account. These were then repeatedly amplified and defended by his growing base of online supporters (Wells et al. 2016). This endless stream of social media content meant that Trump was regularly trending on Twitter, providing constant fodder for journalists covering the campaign. As Wells et al. (2016) describe, Trump attracted a great deal of attention by taking advantage of this hybrid media system. He received media attention through conventional channels including rallies, press conferences, and interviews—not to mention uninvited call-ins to radio and television programs. But he also unleashed "tweetstorms" that galvanized his supporters and helped catapult him to unmatched media coverage in the 2016 campaign.

As Trump's campaign gained momentum, reports began to emerge of increased hate speech, bias incidents, and hate crimes—with some perpetrators explicitly claiming that Trump had motivated their actions. Attracting more mainstream coverage with the release of Univision anchor Jorge Ramos' widely publicized documentary, "Hate Rising," stories of how the Trump campaign was emboldening hate groups and giving rise to a new wave of anti-minority hostility proliferated (Montagne 2016). Graphic evidence of anti-Semitic harassment of Jewish journalists on Twitter by Trump supporters gained widespread attention as well, frequently accompanied by death threats and Holocaust imagery (Rapaport 2016).

These reports continued and intensified in the aftermath of Trump's election. Articles like the New Yorker's "Hate Is on the Rise After Trump's Election," The Guardian's "Trump's Election led to Barrage of Hate," and Vox's "The Wave of Post-Election Hate Reportedly Sweeping the Nation, Explained," became increasingly widespread. James King's year-in-review column, "The Year in Hate: From Donald Trump to the Rise of the Alt-Right," Salon's "A Short History of Hate" which tracks the alt-right's 2016 ascendance, and the New York Times hate-speech aggregator, "This Week in Hate," are just a few examples of this trend (Duncan 2017).[4]

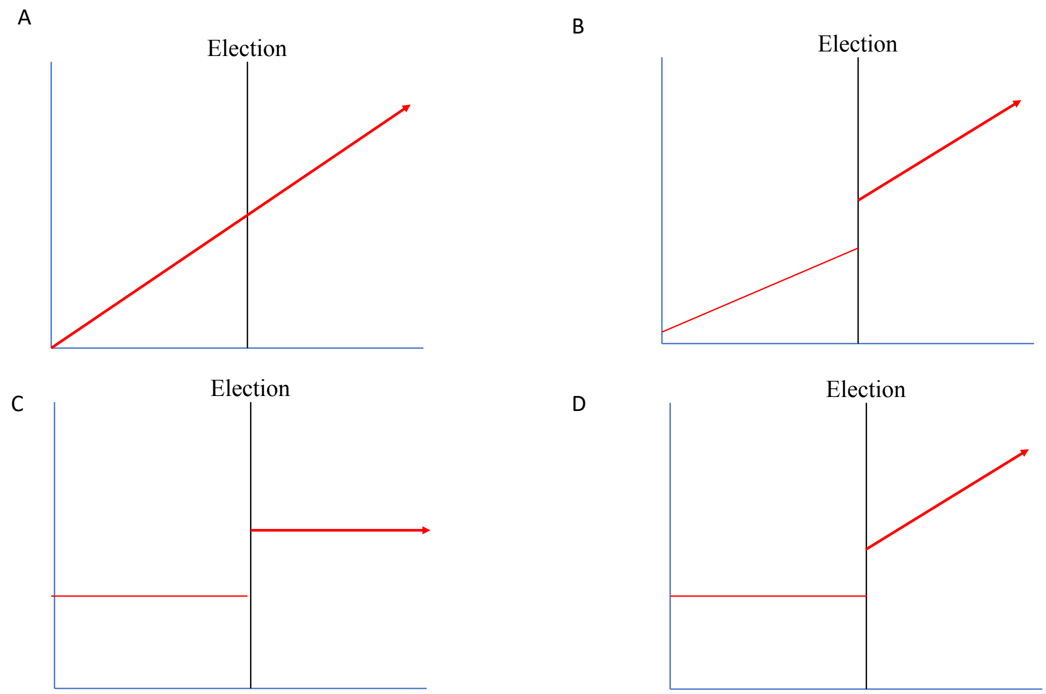
This anecdotal evidence of increased hate crimes and hate speech suggest that Trump's rise may have played a role in legitimizing and mainstreaming extremist rhetoric. It is also

[4] Links to the aforementioned news stories are listed here: Desmond-Harris, Jenna. The Wave of Post-Election Hate Reportedly Sweeping the Nation, Explained. Vox 17 Nov. 2016. http://www.vox.com/2016/11/17/13639138/trump-hate-crimes-attacks-racism-xenophobia-islamophobia-schools; King, James. This Year in Hate. Vocativ 12 Dec. 2016. http://www.vocativ.com/383234/hate-crime-donald-trump-alt-right-2016/; Okeowo, Alexis. Hate on the Rise after Trumps Election. New Yorker 17 Nov. 2016. http://www.newyorker.com/news/news-desk/hate-on-the-rise-after-trumps-election; Sidahmed, Mazin. Trump's Election Led to 'Barrage of Hate', Report Finds. The Guardian 29 Nov. 2016. https://www.theguardian.com/society/2016/nov/29/trump-related-hate-crimes-report-southern-poverty-law-center;Weisberg, Jacob. The Alt-Right and a Deluge of Hate. Slate 1 Nov. 2016. http://www.slate.com/articles/podcasts/trumpcast/2016/11/how_the_alt_right_harassed_david_french_on_twitter_and_at_home.html.

consistent with a long-standing political science literature on the effects of elite cuing on mass attitudes and behaviors. The bulk of this literature has tended to focus on the effects of elite cuing on public opinion formation and policy preferences (Zaller 1992, 1994; Brader, Tucker and Duell 2013). Perhaps of greater relevance to understanding the purported "Trump effect" on rising hate speech, the social movements literature has demonstrated that elite cues can create discursive opportunity structures that make it easier for far right movements or ideologies to gain traction (Giugni et al. 2005; Koopmans and Olzak 2004; Koopmans and Muis 2009). Exploring the effect of elite cues on the tenor of mass rhetoric, political scientists have demonstrated that elites play an important role in changing social norms and spreading racist and intolerant discourse in a variety of cultural contexts (Van Dijk 1992). This finding is also reflected in the communications literature on the spiral of silence, which suggests that people are alerted to social norms by elites, particularly the media. While this theory was first formulated to explain social desirability biases in survey responses (Noelle-Neumann 1974; Glynn, Hayes and Shanahan 1997; Scheufle and Moy 2000), it has also been used to explain self-expression on social networking sites (Fox and Warber 2014; Pang et al. 2016; Lee and Chun 2016; Sherrick and Hoewe 2018). Moreover, recent work in the Arab context demonstrates that elites play an important role in instigating and spreading hate speech and intolerant rhetoric in the online sphere (Siegel 2015; Siegel et al. 2018*a*,*b*). Academics have also explicitly theorized about Trump's role in legitimizing fringe groups and ideologies (Ott 2017; Barkun 2017). These findings from diverse bodies of literature all suggest mechanisms by which the spread of online hate speech during the 2016 election campaign and its aftermath may have been exacerbated by Trump's divisive campaign and election.

While both social science theory and popular narratives suggest that Trump's campaign and election may have increased the popularity of hate speech online, existing evidence is largely anecdotal. Here we provide the first large-scale empirical test of this relationship. More specifically, we test the extent to which hate speech and white nationalist rhetoric on Twitter increased over the course of the campaign and/or following the election. What exactly would such empirical evidence look like in practice? On the one hand we might expect that Trump's political rise throughout the 2016 campaign, and the period following his election, were accompanied by a steady increase in hate speech (a positive relationship between the time since Trump declared his candidacy and the prevalence of hate speech). On the other hand we might expect that Trump's unexpected election led to a sudden increase in hate speech (a discontinuity surrounding Trump's victory). We can therefore imagine several different patterns that might be evidence of a lasting effect of the campaign and/or election on the popularity of online hate speech. These representations are displayed in Figure 1.

Figure 1: Hypothesized Effects of 2016 Election Campaign and Election on Hate Speech



Here the Y axis is the prevalence of hate speech and the X axis represents time before and after November 8, 2016, which is marked by the vertical "Election" line.

If Trump's rise as a mainstream candidate and his divisive campaign increased the popularity of hateful language, then we might expect to see a positive upward slope in hate speech over the course of the campaign and through the election period (A).[5] If the election itself—Trump's unexpected victory—impacted the popularity of hate speech, then we should see a discontinuity or jump after the election (C) and/or an upward trend in the use of hate speech following Trump's victory (D), relative to the pre-election period. We might also observe a combination of increasingly hateful rhetoric over the course of the campaign as well as a jump following the election (B). Other combinations, such as a positive increase over the course of the campaign (as in A or B in the pre-election period), followed a sustained (flat) level of online hate following the election (as in C), with or without a discontinuity at the election, would also be evidence of this relationship.

3 Data and Measurement

In this section, we first describe the primary data sources used in our analyses. We then present our machine-learning-augmented dictionary-based approach to measuring hate

[5] As we will explain in much greater detail in the following section, there are a variety of different ways to measure the prevalence of hate speech. We rely on four such measures for our dictionary-based analysis: the counts of tweets containing hate speech; the count of users producing hate speech; the proportion of tweets containing hate speech; and the proportion of users producing hate speech. For the non-dictionary based methods, we rely on changes in the semantic similarity of daily tweets over time to naturally annotated corpora of online hate speech.

speech and white nationalist language. To analyze the extent to which Trump's campaign and election were associated with an increase in hate speech and white nationalist rhetoric on Twitter, we rely on two sources of data. Our Political Twitter datasets contain all tweets referencing Donald Trump and Hillary Clinton produced between June 17, 2015 and June 15, 2017. These include tweets directly mentioning the candidates using their Twitter handles, such as tweets that include @realDonaldTrump or @HillaryClinton, and any tweets that contain the names of the candidates.[6] We began collecting tweets mentioning Donald Trump on June 17, 2015, the day after he declared his candidacy. Our analysis therefore covers the entire period of his campaign as well as more than half a year after his election. This Political Twitter dataset contains over 150 million tweets related to Hillary Clinton and over 600 million tweets related to Donald Trump. This gives us a very large snapshot of political discourse throughout the 2016 campaign and election period.

However, in order to test the degree to which hate speech and white nationalist language became more mainstream over the course of the 2016 presidential election campaign and its aftermath, we also move beyond the political Twittersphere to assess these trends in the American Twittersphere more broadly. To do this, we use a collection of tweets sent by a random sample of 500,000 American Twitter users, sampled by generating random user IDs and then checking that those random numbers correspond to accounts that were active and located in the United States.[7] In the period under study, from June 17, 2015 to June 15, 2017, this collection contained approximately 400 million tweets. Together, these datasets enable us to test whether there was a relationship between Trump's political rise and the prevalence of online hate speech—in any of the ways identified in the previous section and illustrated in Figure 1—in either political or general discourse on Twitter.

Systematically measuring shifts in the popularity of hate speech and white nationalist language over the course of the 2016 campaign and election first requires defining and operationalizing these terms. There is no agreed upon definition of hate speech, and the topic has been hotly debated by academics and policymakers alike. In general, there are two primary tendencies in the literature. At one end of the spectrum are broad more comprehensive definitions that are designed to identify hate speech in a variety of incarnations. At the other end are narrower definitions, which characterize hate speech as "dangerous speech" that is explicitly intended to incite violence or to call for threatening action against an outgroup.[8]

Attempting to gain a more systematic understanding of the use of online hate speech in the 2016 election period, we define hate speech as *bias-motivated, hostile and malicious language targeted at a person or group because of their actual or perceived innate characteristics, especially when the group or individual are unnecessarily labeled* (Cohen-Almagor 2011). By this definition, endorsements of groups associated with hate crimes or bias-motivated behavior, such as the Ku Klux Klan, or statements showing excessive pride in one's own race or

[6]These tweets were collected using Twitter's streaming API and capture all tweets sent in this period mentioning the candidates, with the exception of infrequent missing data caused by Twitter's rate limits.

[7]The random selection was achieved by sampling users based on their numeric ID: first, we generated random numbers between 1 and the highest numeric ID assigned at the time of data collection (3.3 billion); then, for each number we checked whether the user existed, and whether the 'time_zone' field in their profile was one of the time zones in the United States or whether their 'location' field mentioned the full name or abbreviation of a U.S. state or one of the top 1,000 most populated cities; if the user met one of these conditions, they were included in the sample.

[8]See Gagliardone, Patel and Pohjonen (2014), Kennedy et al. (2018), and Siegel (2019) for a detailed overview of the literature on defining hate speech on and offline.

group do not constitute hate speech. While such content is often offensive and frequently occurs alongside hateful language, following previous studies we define hate speech as requiring a disparagement of others (Warner and Hirschberg 2012).[9]

However, because we are also interested in the extent to which white nationalist or extreme right wing rhetoric increased over the course of Trump's campaign or following his election, we explicitly measure white nationalist rhetoric as well. Following the political science literature on white nationalism, we define this language as content praising or associated with "white nationalist" groups also known as "racist right-wing," "extreme right," "far right," or "hate" groups.[10] What distinguishes white nationalist ideology is its concern with protecting a white racial identity under siege. White nationalist groups ascribe to the belief that people of white European backgrounds are a separate community based on myths of common ancestry and culture that transcend national boundaries (Kaplan 2000). In particular, they oppose anything that they believe will dilute the purity of their exceptionalist white culture including immigration, interracial marriage, globalization, and multiculturalism (Fording 2014). For the purposes of this paper, we define white nationalist rhetoric—in contrast to hate speech—as *any rhetoric or content that praises known white-nationalist groups, shows excessive pride in the white race, espouses white supremacist or white separatist ideologies, or focuses on the alleged inferiority of nonwhites.*

Past studies of online hate speech have frequently relied on dictionary-based methods, which require knowing words associated with hate speech in advance (Silva et al. 2016; Tuckwood 2014). Other studies have incorporated sentiment analysis, natural language processing, neural networks, and other supervised and unsupervised machine learning approaches to classifying hate speech (Fortuna and Nunes 2018; Kennedy et al. 2018; Olteanu et al. 2018; Davidson et al. 2017; Waseem and Hovy 2016; Gitari et al. 2015; Siegel 2019). We begin with a dictionary-based method, which enables us to classify tweets that contain the types of hate speech and white nationalist language that have received a great deal of coverage in reporting on how Trump's rise has increased the popularity of online hate speech. In particular, we develop dictionaries of anti-Asian, anti-Black, anti-Immigrant, anti-Muslim, anti-Semitic, homophobic, and misogynistic slurs,[11] as well as a dictionary of white nationalist rhetoric using terms collected in pre-existing databases of hate speech. These include Hatebase and the Racial Slur Database, comprehensive online repositories of global hate speech (Tuckwood 2017; Belnik 2017), and the Anti-Defamation League's database of slogans, terms, and symbols used by white-nationalist groups (ADL 2017*b*). A list of the terms used in our analysis can be found in the online appendix Section A4.[12] It is worth noting

[9]As will be discussed in more detail later, this "definitional" discussion guides our approach to dictionary based methods for identifying hate speech. Our non-dictionary based approach measure the extent to which speech on Twitter "resembles" (i.e., share semantic similarity with) hate-speech as utilized in practice online. See Section 5 below.

[10]The specific criteria used to identify hate groups are debated in the literature (for a review of the literature, see Hainsworth (2000)). See also George and Wilcox (1996) who focus on political style and tactics over ideological dimensions. For an overview of defining the American Right and "extreme right", see Eatwell and O'Sullivan (1989).

[11]While these categories are not exhaustive, they are the most common types of hate speech documented in the US context in pre-existing databases of online hate speech such as Hatebase and the Racial Slur Database described below.

[12]We began with these pre-existing databases of hate speech and white nationalist terms. We excluded common English terms from these lists that are rarely used as hate speech and appear frequently in non-hateful contexts on Twitter. The removed terms are marked in the list in the online appendix, Section A4. We then looked at random samples of tweets from our Twitter datasets containing these terms and, where relevant, supplemented these

that our hate speech categories are not mutually exclusive. One tweet can direct hate speech at several groups or individuals, as the following examples of hate speech tweets from our political Twitter dataset illustrate:

Table 1: Hate Speech and White Nationalist Tweet Examples

Hate Speech Tweets
@WhitePriinces: Cant wait for donald trump to send all the monkey looking niggers back home to mexico
RT @Bidenshairplugz: Fucking Jews, calling Donald Trump "violent" while giving negroes and Muslims a pass on their actual violence.
@realDonaldTrump yes get rid of all the beaners, monkey looking chinks and muslims trump2016
White Nationalist Tweets:
RT @WhiteGenociders: #DonaldTrump Peaceful White nationalists protect beauty, family, and land, #AntiWhites want to destroy those things
RT @WhiteGenocideWN: Donald Trump means we don't ever have to apologize for being white ever again #NPI #AltRight @LadyAodh
@HillaryClinton What is CHASING DOWN every last White person, assimilating them with nonwhites and calling it 'Diversity?. . .#whitegenocide

However, as past research demonstrates, one of the main challenges in automatic hate-speech detection on social media is the ability to distinguish between actual use of hate speech, posts denouncing or appropriating slurs, and speech using the terms associated with hate speech but conveying a different meaning. These methods often have low precision because they identify all messages containing particular slurs as hate speech, failing to recognize other uses and meanings of these terms (Davidson et al. 2017). To evaluate the extent to which our dictionaries were accurately identifying hate speech, we used trained undergraduates and crowd-sourced coders to classify a random sample of about 25,000 tweets from our political Twitter dataset containing terms from each of our eight hate speech categories as well as white nationalist rhetoric.[13] After each tweet was coded by three coders, we found that only a fraction of the tweets actually contained hate speech or white nationalist language. Many of the terms occurred in Twitter users' Twitter handles (@angry[bitch]), as part of other words ([spic]y), or homonyms (a "[chink] in his armor"). Moreover, detailed human coding on Crowdflower[14] of a random sample of 5,400 tweets[15] from our political Twitter dataset revealed that 360 tweets, or about 7% of the 5,400 human coded tweets, were explicitly *condemning* the use of hate speech against particular groups. Examples of

dictionaries with additional terms that frequently co-occurred in our data.

[13]Our random sample included up to 3,000 tweets from each of our nine categories, for a total of approximately 25,000 tweets. After initially using trained undergraduate coders to ensure that tweets could be accurately classified, crowd-sourced coding was done using Crowdflower, a data enrichment platform that allows a researcher to launch microtasks to a "crowd" of over five million contributors. For a recent overview of how to use Crowdflower in political science research, see (Benoit et al. 2016). Tweets were each coded by three coders. Test questions for quality control ensured that the contributors coding tweets were responding to tasks truthfully and conscientiously. If a contributor answered test questions incorrectly, that contributor was removed from the job and their data was erased.

[14]Coding instructions can be found in the online appendix Section A1.

[15]600 containing terms from each of our 8 hate speech dictionary and 600 containing white nationalist terms.

such tweets are displayed in Table 2.

Table 2: Tweets Condemning Hate Speech and White Nationalism

RT @Jaaezus: # DonaldTrump won the election & white people already don't know how to act This white boy told me I'm a Nigger
Already been flicked off and called a wetback and it's only been 3 days... thanks Donald trump
RT @snowangja: Donald Trump is the type to call every east asian people as ching chong too. I'm not shocked https://t.co/zFHwPLmsr9
RT @ShaunKing: This just happened in Indiana. "Fuck you nigger bitch. Trump is going to deport you back to Africa." Day 1 of Donald
@realDonaldTrump Quite frankly Mr.Trump, you could have achieved this win without trying to get support from the # WhiteGenocide nutcases.!

This prevalence of anti-hate speech tweets, as well as tweets using slurs or white nationalist terms in an irrelevant manner, highlight the need to move beyond a purely dictionary-based approach to classify our tweets. Using the 25,000 total set of human coded tweets as a training dataset, we trained two binary Naive Bayes classifiers, one to identify hate speech tweets and one to identify tweets containing white nationalist rhetoric.[16] Our classifiers allowed us to identify which of the tweets containing terms from our dictionaries actually expressed hate speech and white nationalist sentiments, significantly improving the accuracy of our method.

With a method in hand for classifying tweets as containing hate speech or not, we can now return to the question of how to best measure the "prevalence" or "popularity" of hate speech on Twitter, in order to test the hypotheses laid out in Figure 1. To ensure that our findings are not driven by one particular measure, we develop four measures of the prevalence or popularity of hate speech in each of our collections on Twitter: 1) the number of tweets containing hate speech each day; 2) the number of unique users producing hate speech each day; 3) the proportion of tweets each day containing hate speech; and 4) the proportion of unique users producing hate speech each day.

All four of these measures are substantively interesting and capture slightly different dimensions of prevalence or popularity of online hate. We can imagine being concerned about the overall incidents hate speech, in which case we would want to measure total tweets as opposed to total users. However, we might also think that the more politically relevant outcome was whether more *people* started using hate speech online. Comparing counts to proportions, one could certainly argue that what is most politically relevant is the growth in quantity of hate speech in the public discourse. However, a documented increase in hate speech that was cited as evidence of an increase in hate speech online might be suspect if the overall size of the corpus of tweets was growing at the same point in time; comparing the proportion of tweets containing hate speech on a given day insulates us against that critique.

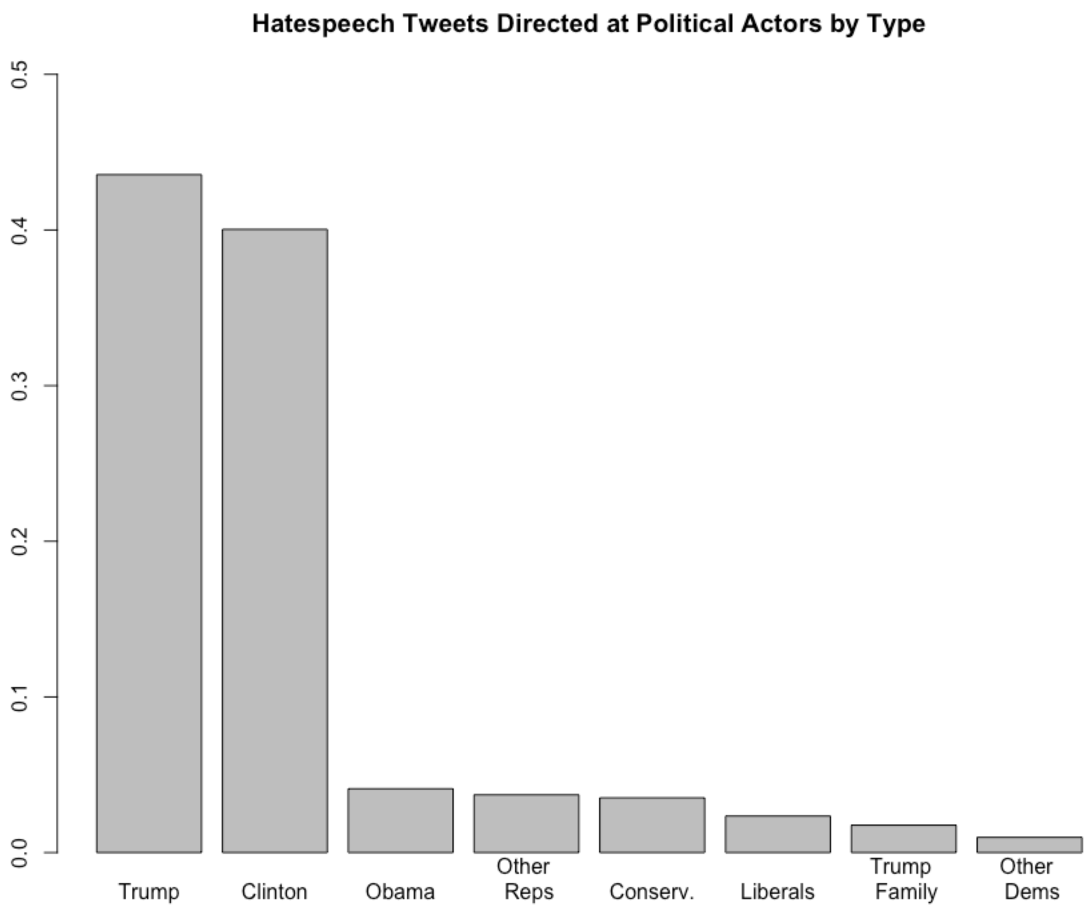
With no *a priori* reason to favor one of these measures over the others, we run our analyses on all of them. For reasons of space, we will present analyses of the proportion of

[16]This process is described in detail in the online appendix (Section A1)

tweets containing hate speech in the main text of the paper; analyses using the remaining three measures can be found in the online appendix (Section A3). As it turns out, analyses from all four measures all tell essentially the same story.

Before proceeding to our analysis, it is worth noting that almost one quarter of the tweets containing hate speech in our human-coded data[17] were directing the hate speech at political actors, especially at Hillary Clinton and Donald Trump. The breakdown of these tweets can be seen in Figure 2 below.

Figure 2: Proportion of Human Coded Tweets Directed at Political Actors



Of the 2,203 tweets classified by human coders as containing hate speech in our random sample of 5,400 tweets from the Political Twitter datasets containing dictionary terms for each type of speech, 512 (or almost one quarter of the hate speech tweets) were directed at political actors. This histogram shows the proportion of these 512 hate speech tweets directed at each type of political actor.

[17] While we asked human coders to code the first 5400 hate speech tweets along several dimensions (displayed in the online appendix) in order to get a better sense of their content, we did not train a classifier to examine targets of hate speech and therefore only conducted this exploratory analysis on human coded tweets.

The widespread use of hate speech against Donald Trump suggests that although the rise in hate speech has mostly been characterized as a phenomenon that has emboldened Trump supporters, many of Trump's opponents—on the right and left—produced online hate speech throughout the 2016 election cycle. Table 3 provides examples of these tweets containing hate speech directed at political candidates.

Table 3: Hate Speech Tweets Directed at Political Actors

Donald Trump Ahead In Internal Polls & that cunt Hillary is same as Obama nigger shit! She's a loser. https://t.co/bEYP72KADr via @YouTube
Donald trump uses self tanner... what sort of faggot shit is that. I thought we elected a man not a fairy!
@realDonaldTrump @nypost blow putin brown noser loser # dumpdrumpf # ivankarussianbride # tiffanytrumprussianslutprincess # deport them all
Fuck Donald Trump let's all impeach that pathetic neo-nazi faggotic bitch ass fat fuck whale and kick his ass back to Europe feeling lucky
@realDonaldTrump called on ALL GREAT AMERICANS to UNITE to # TrumpThatBitch -# Lockherup - # BuildtheWall -# DrainTheSwamp # MAGA!!!

4 Empirical Strategy and Dictionary-Based Results

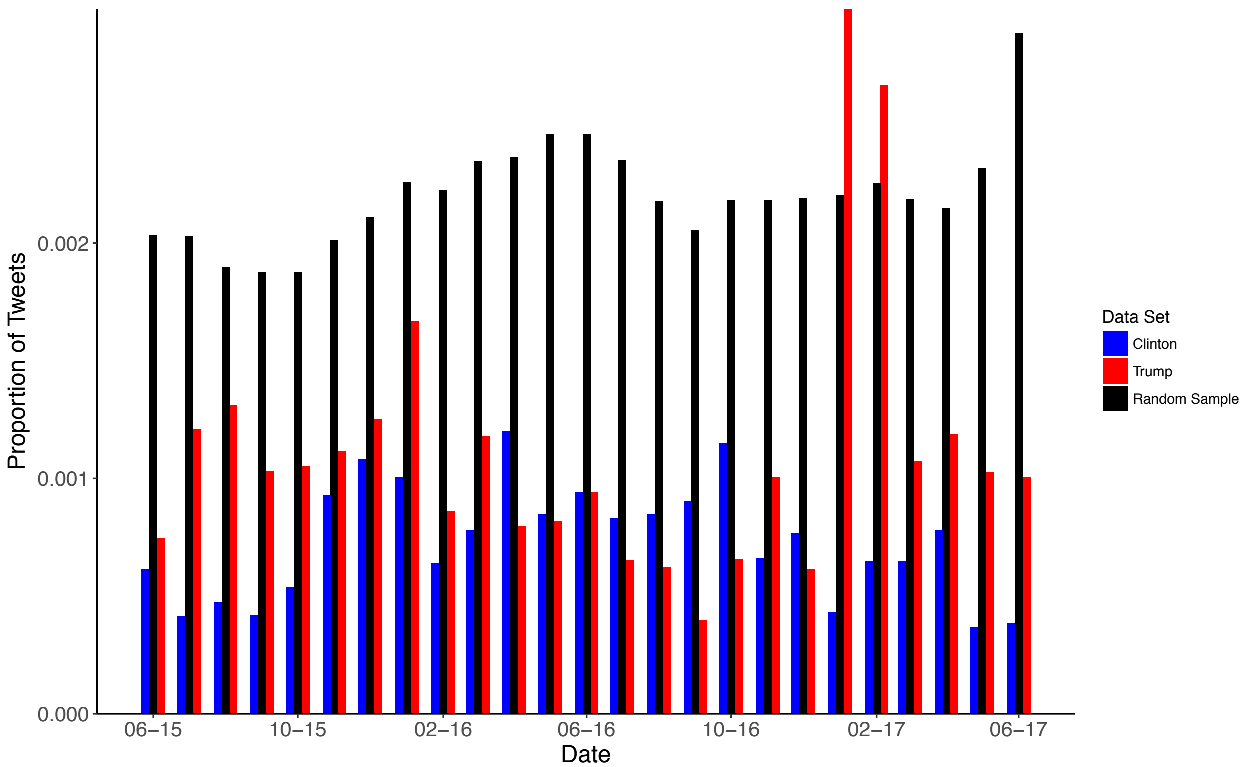
In order to assess whether there is empirical support for the any of the hypothesized relationships displayed in Figure 1, we employ Interrupted Time Series Analysis, a statistical method we describe in more detail below. Before presenting the results of these analyses, however, we start by simply displaying the raw counts of the number of tweets containing hate speech. Figure 3 shows the monthly proportion of hate speech tweets produced in the Clinton, Trump, and random sample datasets between June 17, 2015 and June 15, 2017.[18] Plots showing the raw counts of tweets as well as the number and proportion of unique users producing them look quite similar.[19] We find that on any given day between 0.001% and 0.003% of tweets contain contain hate speech, a tiny fraction of both political language and general content produced by American Twitter users. More importantly, Figure 3 also reveals that Trump's victory (in November 2016) does not appear to have increased the proportion of hate speech. The Clinton dataset contains less hate speech after the election, while the random sample data remains about the same for several months. The largest spike in monthly hate speech in the Trump dataset occurs in late January 2017. Interestingly, while still only a fraction of a percentage point, we observe a higher proportion of hate speech in the random sample data than we do in the political datasets.

Analysis of the Trump data by hate speech type reveals that this spike is largely explained by a substantial uptick in misogynistic hate speech following the announcement of Trump's "travel ban" executive order. This increased misogynistic language appears to be a reaction to Clinton's decision to break her post-election silence to criticize the ban, as well as language

[18] For our statistical analysis, as well as our non-dictionary based robustness tests in the following section, we will rely on the day as the primary unit of analysis for time. As Figure 3 contains data from all three collections, we elected to pool the data by month in order to make the visualization feasible.

[19] See Figures A10-A14 in the online appendix for plots displaying these other measures.

Figure 3: Monthly Proportion of Classified Hate Speech Tweets in the Clinton, Trump, and Random Sample Datasets

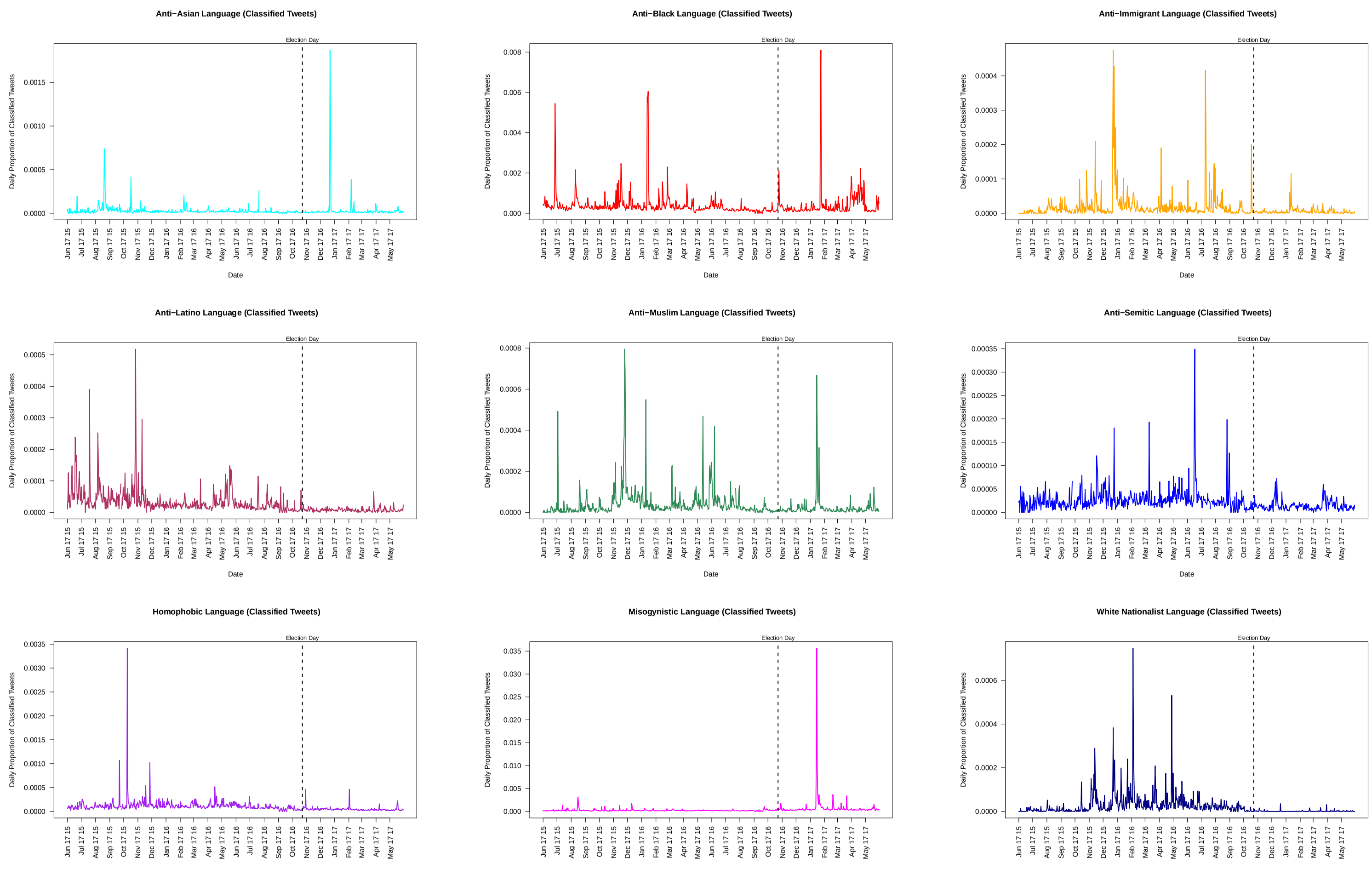


This figure shows the monthly proportion of classified hate-speech tweets in the Clinton, Trump, and random sample datasets. We classified tweets as hate speech (or not) using a Naive Bayes Classifier to remove false positives from tweets containing hate speech dictionary terms. Similar plots for white nationalist language, and plots displaying raw counts of the data rather than proportions are available in the supplementary materials (Figures A10-A14).

directed at then acting Attorney General Sally Yates, who declined to defend the ban and was subsequently fired by Trump. There are also spikes in anti-Asian, anti-Muslim, and anti-Black language in this period, though their volume is much lower. These patterns disaggregated by hate speech type are displayed in Figure 4.[20] These plots suggest that although the data is quite bursty—for example, anti-Muslim tweets spike around terror attacks and anti-Semitic tweets spike following Trump's retweet of anti-Semitic content—there is no evidence of a persistent increase in hate speech over the course of the campaign or in the aftermath of Trump's election for any type of hate speech.

[20]Comparable plots for the Clinton and Random Sample datasets, broken down by different measures of popularity, are provided in the online appendix Figures A1-A9. We also provide tables of descriptive statistics including tweets broken down by dictionary type, classification, unique users, and retweets, which generally follow very similar patterns (See Tables A1-A3).

Figure 4: Daily Proportion of Tweets Containing Hatespeech and White Nationalist Language (Trump Dataset)



This plot shows the daily proportion of tweets containing hatespeech and white nationalist rhetoric in a dataset of over 600 million tweets mentioning Donald Trump collected using Twitter's Streaming API between June 17, 2015 and June 15, 2017. Hatespeech tweets were identified both using dictionaries of slurs and a Naive Bayes Classifier trained to remove false positives from our data.

Beyond these initial insights from the raw data, our daily measures of the number and proportion of tweets containing hate speech or white nationalist rhetoric and the number and proportion of unique users tweeting them across our datasets enable us to systematically test the extent to which this language became more popular both over the course of the 2016 campaign and in the aftermath of Trump's election.

In particular, we rely on Interrupted Time Series Analysis (ITSA). Interrupted time-series analysis is a powerful quasi-experimental design for assessing the longitudinal impact of an event or intervention (Bernal et al. 2013; Bernal, Cummins and Gasparrini 2016). We use ITSA to model the popularity of hate speech on Twitter over the course of Trump's campaign and the effect of his election as follows:

$$Y_t = \beta_0 + \beta_1(T) + \beta_2(X_t) + \beta_3(X_t T) \tag{1}$$

In Equation 1 above, Y_t is the proportion of tweets containing hate speech or white nationalist rhetoric at time t, T is the time since Trump announced his candidacy, X_t is a dummy variable representing Trump's 2016 election (here the pre-election period is coded as 0 and the post-election period is coded as 1), and XT_t is an interaction term. β_0 represents the baseline proportion of hate speech tweets in each dataset at $t = 0$, β_1 shows the change in the proportion of hate speech tweets associated with a one unit time increase, representing the underlying daily pre-election trend. β_2 captures the immediate effect of the election on the proportion of hate speech tweets, or an intercept change, and β_3 captures the slope change in the popularity of hate tweets following the election, relative to the pre-election trend. In other words, ITSA is a segmented regression model. Segmented regression simply refers to a model with different intercept and slope coefficients for the pre- and post-intervention time periods. Here, we use ITSA to measure the pre-election trend, the immediate changes in the proportion of hate tweets following the election, as well as the change in the slope of the daily proportion of hate tweets after the election.[21] If we observe a rise in hate speech over the course of the campaign, then we should see a positive statistically significant coefficient for the pre-election trend, β_1. If Trump's election caused a lasting increase in the use of this rhetoric, then we should either see a positive shift immediately after the election β_2 followed by a non-negative post-event slope change $\beta 3$, or a non-negative immediate effect of the election β_2 followed by a positive slope change in the relative popularity of hate speech in the post-election period β_3.

We also model this relationship including quadratic terms, in case the pre and post-election trends are not well captured by a linear model. In these models, the squared terms tell us whether the pre and post-election trends are concave (on the whole, slowly decreasing over time) in the case of a negative coefficient or convex (on the whole, slowly increasing over time).

$$Y_t = \beta_0 + \beta_1(T) + \beta_2(X_t) + \beta_3(X_t T) + \beta_4(T^2) + \beta_5({X_t}^2) \tag{2}$$

Put another way, this allows us to test all of our hypotheses laid out in Figure 1, which

[21] In order to address serial autocorrelation in our data, we use a first order autoregressive (AR1) model in our analysis instead of the standard OLS ITSA model Bernal, Cummins and Gasparrini (2016).

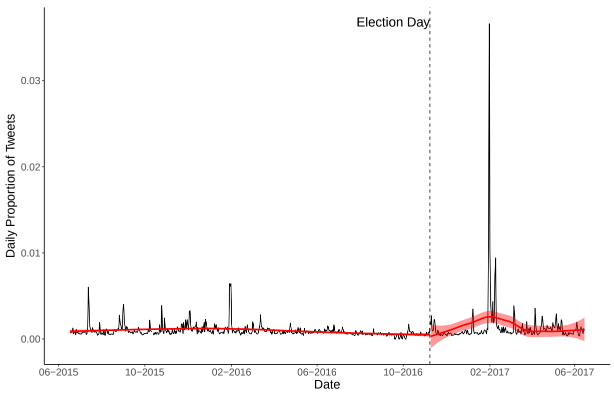
involve hate speech increasing over the course of the campaign (a positive slope over time), immediately after the election (a positive discontinuity), longer term following the election (a slope change from the pre-election to the post-election period), or some combination of those three patterns.

Conducting ITSA using our political Twitter and random sample datasets, we find no evidence of a lasting increase in hate speech or white nationalist rhetoric either over the course of the campaign or in the aftermath of Trump's election. In Figure 5 we plot the pre and post-election trends over the observed daily proportion of hate speech tweets and white nationalist language tweets in our datasets. Beginning with the Trump dataset (Panels (a) and (b))—by far the largest collection— we see no significant increase in the proportion of hate speech or white nationalist language in either period. As Panel (a) demonstrates, the largest spike in the proportion of hate speech in the Trump dataset occurred in late January and early February of 2017, in the period surrounding the aforementioned travel ban. By contrast the largest spike in white nationalist rhetoric occurs following Trump's retweet of a white supremacist account in February 2016.[22]

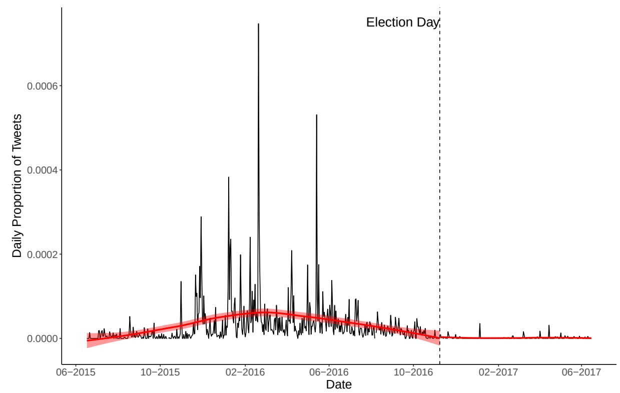
[22]Similarly, there are no persistent increases in the number of unique users producing this content and these results hold using both linear and quadratic models. Plots and regression tables displaying these results for all datasets and outcome variables are available in the online appendix (Figures A15-A32 and Tables A7-A24).

Figure 5: Effect of 2016 Election on Daily Proportion of Hate Speech and White Nationalist Language Tweets
Interrupted Time Series Analysis (Trump, Clinton, and Random Sample Datasets)

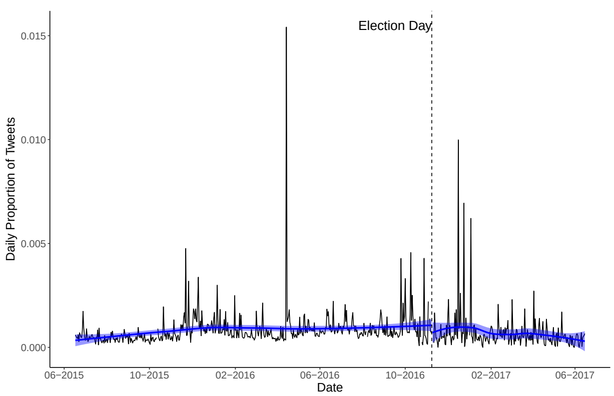
(a) Trump Data: Hate Speech



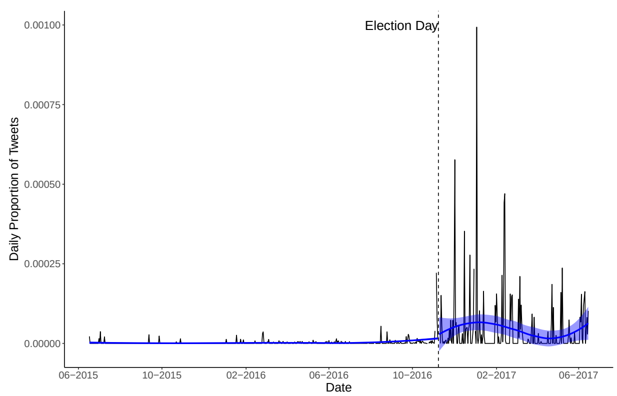
(b) Trump Data: White Nationalist



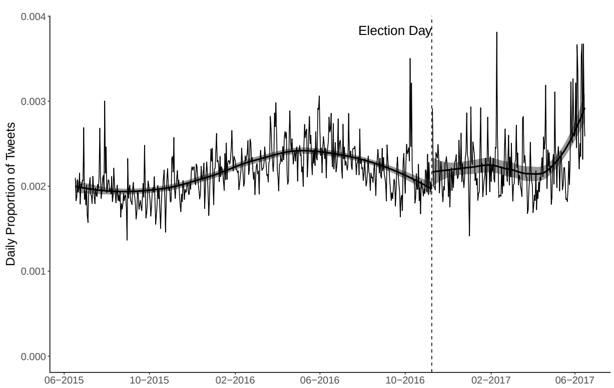
(c) Clinton Data: Hate Speech



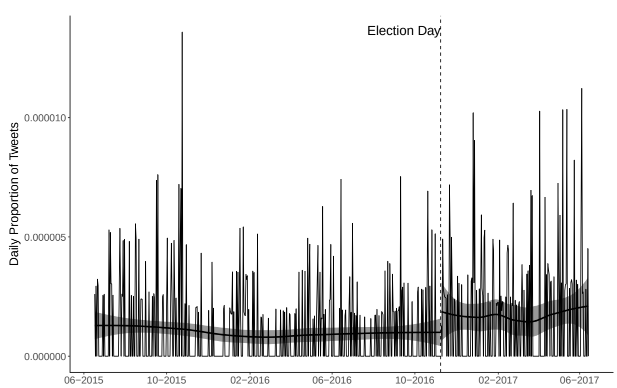
(d) Clinton Data: White Nationalist



(e) Random Sample Data: Hate Speech



(f) Random Sample Data: White Nationalist



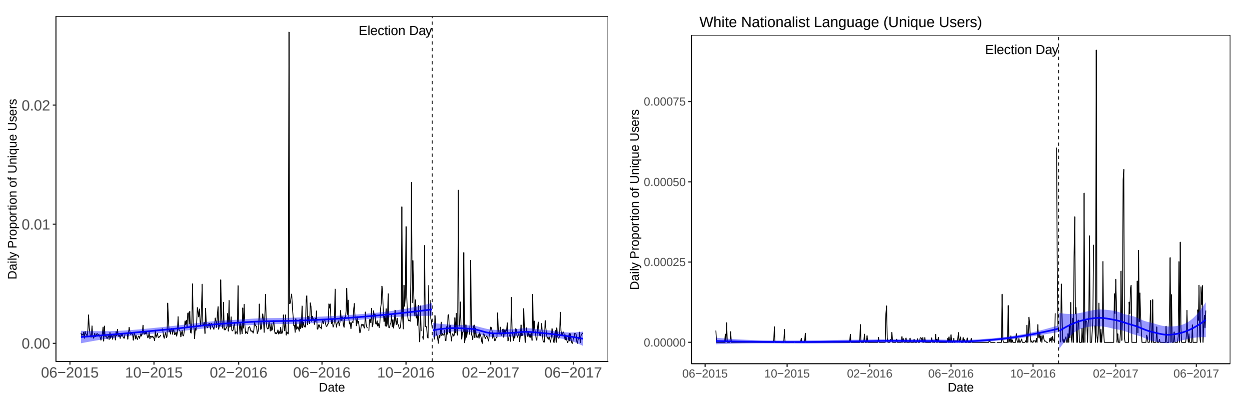
These plots show the pre and post election trends in our ITSA regression models, plotted as local regression lines with loess smoothing and 95% confidence intervals. These trend lines are plotted against the observed daily proportion of hate speech tweets and white nationalist language tweets in our datasets of over 600 million tweets referencing Donald Trump (a and b), 150 million tweets referencing Hillary Clinton (c and d), and 400 million tweets sent by a random sample of American Twitter Users collected using Twitter's Streaming API between June 17, 2015 and June 15, 2017 (e and f). Hate speech and white nationalist language tweets were identified both using dictionaries of slurs and Naive Bayes classifiers trained to remove false positives from our data.

Turning to our Clinton hate speech data (Panel (c) in Figure 5), we again observe no change in the proportion of hate speech over the course of the 2016 campaign or in the aftermath of Trump's election. In fact, we actually observe a statistically significant post-election *decrease* in the number of unique users producing hate speech, displayed in Figure 6. We do, however, observe a statistically significant increase in the proportion of unique users tweeting hate speech over the pre-election period.[23]. Figure 6 displays this trend in the daily proportion of unique users in the Clinton data. This uptick is primarily driven by the increase in misogynistic rhetoric over the course of the Clinton campaign, particularly a spike following her April 2016 debate against Bernie Sanders and a general increase as the election approached.

While we see no evidence of increasing white nationalist rhetoric over the course of the campaign in the Clinton dataset (Panel (d)), we do see an increase in the proportion of tweets containing white nationalist rhetoric—and the proportion of unique users tweeting them—after Trump's election.[24] However, this increase represents a change of only a fraction of a percentage point and—even on the most prolific days—we never observe more than a few hundred white nationalist tweets in the Clinton dataset.

Figure 6: Effect of 2016 Election on Daily Proportion of Unique Users Producing Hate Speech and White Nationalist Language Tweets
Interrupted Time Series Analysis (Clinton Dataset)

(a) Clinton Data: Hate Speech (b) Clinton Data: White Nationalist



These plots show the pre and post election trends in our ITSA regression models, plotted as local regression lines with loess smoothing and 95% confidence intervals. These trend lines are plotted against the observed daily proportion of unique users producing hate speech tweets and white nationalist language tweets in our dataset referencing Hillary Clinton between June 17, 2015 and June 15, 2017 . Hate speech and white nationalist language tweets were identified both using dictionaries of slurs and Naive Bayes classifiers trained to remove false positives from our data.

In order to address the possibility that our *political* Twitter data differ systematically from the U.S. Twittersphere as a whole, we replicate our analysis using a dataset of over 400 million tweets produced by a random sample of 500,000 American Twitter users. Consistent with our results on political Twitter, Panel (e) of Figure 5 shows no lasting increase in online

[23]The regression table displaying this result and plots of the ITSA trend lines both aggregated and disaggregated by hate speech type are provided in Figure A18 Table A12 in the appendix

[24]See Figures A15, A16, A18, A19, and Tables A7, A8, A10, and A11 in the online appendix.

hate speech over the course of Trump's campaign. Furthermore, there is no statistically significant increase in the proportion of tweets containing hate speech following the election—or the number of unique users tweeting them—although we do see a brief one-day spike in the proportion of unique users tweeting hate speech on election day. Once again, this spike in hate speech is largely driven by misogynistic language, though we see a one day spike in anti-black language as well in the random sample data.

Examining trends in white nationalist rhetoric in the random sample dataset, which we plot in Panel (f) of Figure 5, we see no increase in the proportion of tweets containing white nationalist rhetoric over the course of the campaign. While there is some evidence of a slight increase in white nationalist rhetoric after Trump's election, this effect is not statistically significant across specifications and the volume of tweets is even lower than it is in the Clinton collection. Taken together, while we do observe slight increases in white nationalist rhetoric following Trump's election in the Clinton and random sample datasets, and a one day spike in hate speech in the random sample data, these results do not provide support for the proposition that Trump's election prompted increased prevalence or popularity of online hate along any of the ways proposed in Figure 1.[25]

5 Robustness Check: Reference-Text Based Analysis

One of the potential pitfalls of relying on dictionary-based methods for identifying hate speech—no matter how sophisticated the application of these approaches—is that they force the analyst to rely on a corpus of words used in the past to code speech in the present. In most contexts, this is unlikely to be problematic, due to the long life span of slurs and derogatory language. However, given our surprising finding in the previous section that hate speech and white nationalist rhetoric did not increase consistently either over the course of the 2016 election campaign or in its aftermath, we must seriously consider the fact that we have somehow failed to identify a significant subset of hate speech on Twitter.

To give an example, Nikhil Sonnad, writing at *Quartz*, detailed the existence of alt-right code words online, whereby " 'googles' means the n-word; 'skypes' means Jews; and 'yahoos' means 'spic'"(Sonnad 2016). A dictionary-based method that did not contain these code words would therefore be missing the occurrence of hate speech. Perhaps even more problematically, if an event both led to an increase in hate speech *and* to the use of new code words for hate speech, we would completely miss the impact of this event.[26]

With this concern in mind, we repeat our analyses using a semi-supervised text-embedding method of measuring online hate speech and white nationalist rhetoric. The idea underlying this alternative method is to find an example of "hate speech in the wild," or a large

[25] This conclusion also holds when we use any of the alternative measures of prevalence describing previously and reported upon in the online appendix (Section A3).

[26] As an aside, there is an existential question here as to whether hate speech that uses sanitized words ought to be considered hate speech at all. Addressing this issue is beyond the scope of the current paper, although it seems that if everyone knows what is being referred to by"kill all the skypes" then it is difficult to see why this would not be considered hate speech. For now, though, we simply note that the alternative approach described in the remainder of this section is flexible enough to pick up any instances of hate speech that might be missed by dictionary methods, regardless of whether they rely on hateful words that are simply not in the dictionaries we are using or if they represent some sort of new vocabulary—in code or not—for depicting hateful speech.

collection of text that contains the types of hate speech people actually use online. For this task, we rely on publicly available comments posted on Reddit.com for our reference corpus. Reddit.com is a popular news aggregation and discussion web-site, and Reddit entries are organized into forums with specific topics of interest ("subreddits"). Some of these subreddits are infamous for their explicitly racist, hateful and extreme alt-right content. Examples of these subreddits include /r/Coontown, /r/WhiteRights, /rAntiPOZi, /r/european. Many of them were eventually banned or quarantined by the Reddit administration, but the comments that were posted in these subreddits are still available for analysis.

To the extent that these subreddits contain "real world" hate speech, we can measure how much hate speech is present on Twitter by examining the degree to which our tweets resemble the text in these subreddits. Now there will of course be differences in how speech is used across different platforms, so any absolute measure of similarity is going to be difficult to interpret. However, *relative* measures—such as whether tweets after Trump's election were more similar to racist subreddit content than tweets at the beginning of the campaign—can help us assess whether or not hate speech increased on Twitter over the course of the 2016 campaign or after Trump's election. Indeed, we can examine tweets produced each day, just as we did using our machine-learning-augmented dictionary-based methods, but this time comparing not the proportion of classified tweets containing slurs, but rather how much the tweets on a given day resemble the text from hateful or white nationalist subreddits. We can then see if that proportion increases over time and/or shifts significantly following the November 8th election.[27]

Not only does Reddit provide us with naturally-annotated hateful text because the site is organized into "subreddits" that are explicitly devoted to particular topics—in this case communities that are are infamous for their explicitly racist, hateful and extreme alt-right content—but Reddit users frequently up-vote or down-vote posts. By deleting posts that have net negative votes from the data used to train our classifier, we subject our text to two forms of annotation: whether it is posted on a particular subreddit in the first place; and whether users of that subreddit think it belongs there. Because we are not interested in whether the tweets in our collections have a greater resemblance to one type of alt-right subreddit or another, we can we use a two-step classification process. We first train our classifier to group similar subreddits, and then measure the daily probability that the language in our Twitter collections shares features with the language on a collection of alt-right subreddits.[28] The advantage here is that unlike in our first method, we do not need to explicitly provide a dictionary of alt-right terms and phrases. Instead, our model can automatically learn relevant terms from the corpus of subreddit comments.

More specifically, we classify our tweets based on the probability that their text might have been generated by the same generative process as the text found in the hateful (or non-hateful) subreddits. Generally speaking, any kind of supervised classifier can be used to apply this method. Here we use a supervised version of fastText model (Joulin et al. 2016), which is conceptually similar to the skip-gram version of word2vec. Unlike word2vec though,

[27]Moreover, we could also compare whether speech on Twitter in a particular time period is more similar to speech in a anti-minority subreddit than, for example, a pro-minority subreddit, a generic political discussion subreddit, or alternatively a subreddit having nothing to do with politics or identity at all (e.g., a baseball subreddit).

[28]A detailed description of this method is available in the online appendix (Section A2).

instead of learning word representations in an unsupervised fashion, fastText updates these embeddings in a way that optimizes for a particular text classification task. In the current case, it is possible to train the model to predict which subreddit (or group of subreddits) each of the comments in the corpus was posted in. As a result, the model will learn semantic similarities of comments in each subreddit. After training the model to predict which subreddit (or group of subreddits) comments were posted in, it can be used to calculate class probabilities for each tweet in our collection. Finally, average daily probabilities can be calculated for each class, and changes in these probabilities can show us the dynamics of the relative popularity of a particular type of language—in this case hate speech or white nationalist rhetoric—on Twitter. For the sake of parsimony, we relegate the remaining technical details of the method and our extensive validation tests to the online appendix (Section A2.5).

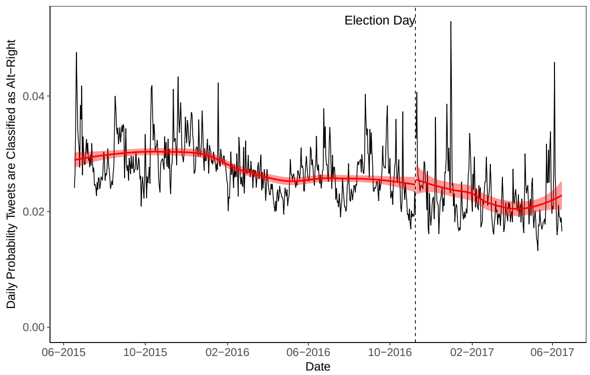
Consistent with our dictionary-based analysis, we do not observe the language in any of our three Twitter collections becoming more similar to content produced on alt-right subreddits over the course of the campaign. Trump's election also has no effect on these probabilities. These findings, the results of our ITSA analyses to assess changes in the likelihood that the text in the tweets in our political and random sample datasets might be found on alt-right subreddits, are displayed in Figure 7.[29]

[29] Regression tables displaying results are provided in the online appendix, Tables A28, A29, and A30.

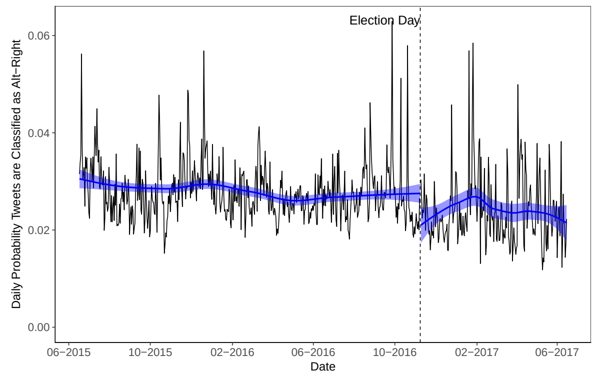
Figure 7: Effect of 2016 Election on Probability that Tweets are Classified as Alt-Right (ITSA)

ITSA (Trump, Clinton, and Random Sample Datasets)

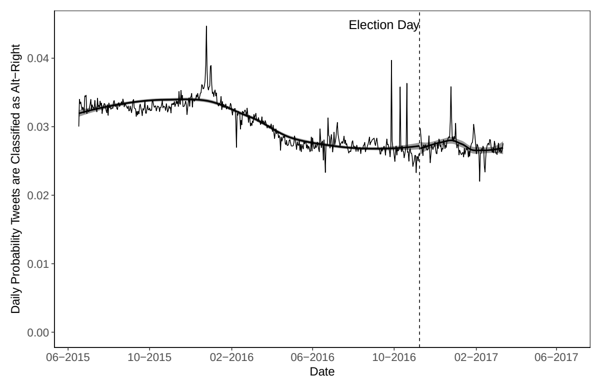
(a) TRUMP DATA



(b) CLINTON DATA



(c) RANDOM SAMPLE DATA



This plot shows the pre and post election trends from our ITSA regression models, plotted as local regression lines with loess smoothing and 95% confidence intervals. These trend lines are plotted against the average daily predicted probabilities that tweets in the Trump (a), Clinton (b), and random sample (c) datasets collected using Twitter's Streaming API between June 17, 2015 and June 15, 2017 are classified as belonging to alt-right subreddits.

6 Conclusions and Steps for Future Research

Contrary to the prevailing narrative that Trump's divisive 2016 campaign and election drove a rise in the popularity of online hate speech, we find little empirical evidence of a persistent "Trump effect" on political Twitter or in a large random sample of American Twitter users during the 2016 US presidential campaign or following Trump's election. Both our dictionary and non-dictionary-based methods reveal no evidence of an increase in hate speech before or after the election across our datasets. By highlighting the shortcomings of the conventional wisdom regarding the rise of hate speech on Twitter over this period, our paper demonstrates the importance of moving beyond short term or small scale datasets when studying online speech.

Precisely because social media platforms like Twitter are so large and diverse, it is easy to find evidence—even relatively large-scale evidence—of the existence of almost any conceivable attitude or behavior. However observing a particular kind of discourse online in a given moment does not necessarily mean it has increased or changed over time. This is particularly true given the bursty nature of Twitter data, where topics can trend briefly in response to events and then re-equilibrate shortly afterwards.[30] There is certainly evidence of thousands of tweets containing hate speech and white nationalist rhetoric on Twitter over the course of the Trump campaign and in its aftermath. And this makes it possible to use snapshots of this data as evidence of a "Trump effect." But when we zoom out and examine the overall and relative popularity of this language over time, we see that such content did not become more common in political comments on Twitter or among American Twitter users in general.

Our analysis therefore offers key innovations for the study of online hate speech and online behavior more broadly that we hope will be adapted by scholars and practitioners alike. Firstly, we employ two different but equally informative datasets: a collection of all tweets referencing the two candidates in the 2016 election, and a random sample of 500,000 American Twitter users. This allows us to study online hate speech both in an explicitly Trump-related political context—where we might expect to see a "Trump effect"—and in a representative sample of American Twitter users. Second, by using both a machine-learning augmented dictionary-based analysis and a non-dictionary approach leveraging data from subreddits to classify hate speech, we avoid drawing conclusions that are biased by one particular classification strategy or definition of speech. Finally, by exploring changes in the volume and proportion of both hateful tweets and unique users producing hateful tweets, we ensure that our findings are robust to a broad set of measures of behavioral trends.

To be clear about the scope of our findings, the application of this method in the present manuscript is limited to Twitter data. While Twitter is of course not the only platform on which hate speech may have proliferated during the election period, our approach enables us to test whether people on a large, popular social media platform—indeed the president's preferred platform—were likely to be incidentally exposed to hate speech, rather than seeking it out on specialized platforms such as Gab, Voat, or particular communities on Reddit. While recent studies have begun to investigate the spread of this language on such alternative platforms (Nithyanand, Schaffner and Gill 2017; Zannettou et al. 2018), this is beyond the scope of our research on the popularization of online hate on Twitter.

Trolling and harassment of journalists on Twitter—particularly anti-Semitic attacks—were frequently reported over the course of the election campaign and may have contributed to the perception of increased widespread online hate in this period. Our approach to measuring broad trends in online hate speech over time does not necessarily allow us to capture these specific incidents if they did not include references to Trump or Clinton or were not perpetrated by users in our random sample of American Twitter users. Thus it is possible that hateful attacks on individuals could have increased over the time period we analyzed, even while hate speech was not increasing generally on Twitter or in discussions of the elections. However, this too would need to be carefully studied, as hateful attacks on individuals on Twitter were taking place before the summer of 2015 as well (Parkin 2014).

[30] Tables of dates with the highest volume of hate speech and white nationalist rhetoric are displayed in the online appendix, Tables A4-A6.

Moreover, our analysis of Twitter data tells us nothing about trends in hate crimes, bias incidents, or other offline events that have also contributed to the popular narrative of a "Trump effect" and deserve further study (Rushin and Edwards 2018; Müller and Schwarz 2018).

Finally, we can of course say nothing about the potentially chilling effects of Trump's political rise on speech of either journalists or ordinary citizens, who might have believed that forays into political discussions were likely to be met with vicious and hateful personal attacks. Our manuscript therefore should not be read as evidence that there were not negative consequences from hateful speech—or hateful acts—during and after the 2016 U.S. election campaign. Nevertheless, the fact that some of these potential negative consequences might have been driven by the *perception* that hateful speech was on the rise itself points to the importance of moving beyond anecdotal reports on hateful speech to rigorous empirical studies of the type presented here.

Despite a growing body of research defining and detecting online hate, the existing scientific literature lacks a systematic approach for tracking the prevalence of this harmful speech over time (Fortuna and Nunes 2018; Gagliardone et al. 2016; Olteanu et al. 2018). Although almost no empirical work has explicitly measured the overall prevalence or temporal dynamics of harmful speech on popular social media sites (Olteanu et al. 2018), governments and online platforms have increasingly proposed and adopted policy interventions to combat online hate speech (Rainie, Anderson and Albright 2017; Marwick 2017; Gagliardone et al. 2016; House of Commons Digital, Culture, Media and Sport Committee 2019). By introducing a new systematic approach to studying the over-time dynamics of hate speech on widely used platforms like Twitter, our work offers a valuable contribution to the study of online hate speech.

Further, our approach could be applied to the study of trends in many other types of online discussion and behavior beyond hate speech and white nationalist rhetoric. Finding consistent results across two different data sets, employing two different means of measuring hate speech, and using several different measures of popularity substantially increases our confidence that we are drawing meaningful inferences about behavior on Twitter over time. Our hope is that by bringing new tools and data sources to the study of online hate speech—and other online discourse—such work will enable academics, policymakers, and everyday citizens alike to better understand and address divisive social and political forces currently at play in the United States and in democracies around the world.

References

ADL. 2017*a*. "Anti-Semitic Targeting of Journalists During the 2016 Presidential Campaign.". Available at: https://www.adl.org/sites/default/files/documents/assets/pdf/press-center/CR_4862_-Journalism-Task-Force_v2.pdf.

ADL. 2017*b*. "Databse of Hate Symbols and Terms." *Anti-Defmation League* . Available at:https://www.adl.org/education/references/hate-symbols.

Barbaro, Michael. 2015. "Pithy, mean and powerful: How Donald Trump mastered Twitter for 2016." *The New York Times* .

Barkun, Michael. 2017. "President Trump and the 'Fringe'." *Terrorism and Political Violence* 29(3):437–443.

Belnik, Sean. 2017. "Racial Slur Database.". Available at: http://www.rsdb.org/.

Benoit, Kenneth, Drew Conway, Benjamin E Lauderdale, Michael Laver and Slava Mikhaylov. 2016. "Crowd-sourced text analysis: reproducible and agile production of political data." *American Political Science Review* 110(2):278–295.

Bernal, James A Lopez, Antonio Gasparrini, Carlos M Artundo and Martin McKee. 2013. "The effect of the late 2000s financial crisis on suicides in Spain: an interrupted time-series analysis." *The European Journal of Public Health* .

Bernal, James Lopez, Steven Cummins and Antonio Gasparrini. 2016. "Interrupted time series regression for the evaluation of public health interventions: a tutorial." *International journal of epidemiology* .

Brader, Ted, Joshua A Tucker and Dominik Duell. 2013. "Which parties can lead opinion? Experimental evidence on partisan cue taking in multiparty democracies." *Comparative Political Studies* 46(11):1485–1517.

Camp, Karin. 2016. "Donald Trump and the Escalation of Hate.". Available at: http://billmoyers.com/story/donald-trump-escalation-hate/.

Cohen-Almagor, Raphael. 2011. "Fighting hate and bigotry on the Internet." *Policy & Internet* 3(3):1–26.

Cohen, Claire. 2016. "Donald Trump sexism tracker: Every offensive comment in one place.". Available at: http://www.telegraph.co.uk/women/politics/donald-trump-sexism-tracker-every-offensive-comment-in-one-place/.

Davidson, Thomas, Dana Warmsley, Michael Macy and Ingmar Weber. 2017. "Automated Hate Speech Detection and the Problem of Offensive Language." *Unpublished Manuscript* . Available at: https://arxiv.org/pdf/1703.04009.pdf.

Duncan, Pansy Kathleen. 2017. "The Uses of Hate: On Hate as a Political Category." *M/C Journal* 20(1).

Eatwell, Roger and Noël O'Sullivan. 1989. *The nature of the right: European and American politics and political thought since 1789*. Burns & Oates.

Filimon, Luiza-Maria. 2016. "Politics and Magical Thinking: How Falsehoods, Showmanshipand Hawkishness Became Trademarks of Republican Presidential Electoral Campaign." *OLD AND NEW DILEMMAS* p. 211.

Financial Times. 2017. "Germany applies the brake to online hate speech." *Financial Times* . Available at: https://www.ft.com/content/75a8c190-08c1-11e7-97d1-5e720a26771b.

Fording, Richard C. 2014. "The Political Origins of Extremism: Minority Descriptive Representation and the Mobilization of American Hate Groups.". SSRN Scholarly Paper ID 3116303, Social Science Research Network, Rochester, NY.

Fortuna, Paula and Sérgio Nunes. 2018. "A survey on automatic detection of hate speech in text." *ACM Computing Surveys (CSUR)* 51(4):85.

Fox, Jesse and Katie M Warber. 2014. "Queer identity management and political self-expression on social networking sites: A co-cultural approach to the spiral of silence." *Journal of Communication* 65(1):79–100.

Gagliardone, Iginio, Matti Pohjonen, Zenebe Beyene, Abdissa Zerai, Gerawork Aynekulu, Mesfin Bekalu, Jonathan Bright, Mulatu Moges, Michael Seifu, Nicole Stremlau et al. 2016. "Mechachal: Online Debates and Elections in Ethiopia-From Hate Speech to Engagement in Social Media.".

Gagliardone, Igninio, Alisha Patel and Matti Pohjonen. 2014. "Mapping and Analysing Hate Speech Online: Opportunities and Cha llenges for Ethiopia.".

George, John and Laird M Wilcox. 1996. *American extremists: Militias, supremacists, klansmen, communists & others*. Prometheus Books Amherst, NY.

Gitari, Njagi Dennis, Zhang Zuping, Hanyurwimfura Damien and Jun Long. 2015. "A lexicon-based approach for hate speech detection." *International Journal of Multimedia and Ubiquitous Engineering* 10(4):215–230.

Giugni, Marco, Ruud Koopmans, Florence Passy and Paul Statham. 2005. "Institutional and discursive opportunities for extreme-right mobilization in five countries." *Mobilization: An International Quarterly* 10(1):145–162.

Glynn, Carroll J, Andrew F Hayes and James Shanahan. 1997. "Perceived Support for One's Opinions and Willingness to Speak Out: A Meta-Analysis of Survey Studies on the" Spiral of Silence"." *Public opinion quarterly* pp. 452–463.

Guynn, Jessica. 2016. "'Massive rise' in hate speech on Twitter during presidential election." *USA Today* . Available at: http://www.usatoday.com/story/tech/news/2016/10/21/massive-rise-in-hate-speech-twitter-during-presidential-election-donald-trump/92486210/.

Hainsworth, Paul. 2000. "The Front National: from ascendancy to fragmentation on the French extreme right." *The Politics of the Extreme Right: From the Margins to the Mainstream* pp. 18–32.

House of Commons Digital, Culture, Media and Sport Committee. 2019. "Disinformation and fake news: Final Report." *Eighth Report of Session 201719* . Available at https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/1791.pdf.

Huber, Lindsay Perez. 2016. "Make America Great again: Donald Trump, Racist Nativism and the Virulent Adherence to White Supremecy Amid US Demographic Change." *Charleston L. Rev.* 10:215.

Joulin, Armand, Edouard Grave, Piotr Bojanowski and Tomas Mikolov. 2016. "Bag of Tricks for Efficient Text Classification." *Unpublished Manuscript* . Available at: http://arxiv.org/abs/1607.01759.

Kaplan, Jeffrey. 2000. *Encyclopedia of white power: A sourcebook on the radical racist right.* Rowman & Littlefield.

Kennedy, Brendan, Drew Kogon, Kris Coombs, Joe Hoover, Christina Park, Gwenyth Portillo-Wightman, Aida Mostafazadeh, Mohammad Atari and Morteza Dehghani. 2018. "A Typology and Coding Manual for the Study of Hate-based Rhetoric.".

Koopmans, Ruud and Jasper Muis. 2009. "The rise of right-wing populist Pim Fortuyn in the Netherlands: A discursive opportunity approach." *European Journal of Political Research* 48(5):642–664.

Koopmans, Ruud and Susan Olzak. 2004. "Discursive opportunities and the evolution of right-wing violence in Germany." *American Journal of Sociology* 110(1):198–230.

Lee, Moon J and Jung Won Chun. 2016. "Reading others comments and public opinion poll results on social media: Social judgment and spiral of empowerment." *Computers in Human Behavior* 65:479–487.

Marwick, Alice. 2017. "Are There Limits to Online Free Speech?". Available at: https://points.datasociety.net/are-there-limits-to-online-free-speech-14dbb7069aec.

Milligan, Susan. 2017. "'A Safe Space for Hate'.". Available at: https://www.usnews.com/news/the-report/articles/2017-03-24/donald-trump-and-the-politics-of-hate.

Montagne, Renee. 2016. "To Make 'Hate Rising,' Jorge Ramos Spent Time With Hate Groups.". Available at: http://www.npr.org/2016/10/21/498804694/to-make-hate-rising-jorge-ramos-spent-time-with-hate-groups.

Müller, Karsten and Carlo Schwarz. 2018. "Making America Hate Again? Twitter and Hate Crime under Trump." *Unpublished Manuscript* .

Nithyanand, Rishab, Brian Schaffner and Phillipa Gill. 2017. "Online Political Discourse in the Trump Era." *Unpublished Manuscript* .

Noelle-Neumann, Elisabeth. 1974. "The spiral of silence a theory of public opinion." *Journal of communication* 24(2):43–51.

Olteanu, Alexandra, Carlos Castillo, Jeremy Boy and Kush R Varshney. 2018. "The Effect of Extremist Violence on Hateful Speech Online." *arXiv preprint arXiv:1804.05704* .

Ott, Brian L. 2017. "The age of Twitter: Donald J. Trump and the politics of debasement." *Critical Studies in Media Communication* 34(1):59–68.

Pang, Natalie, Shirley S Ho, Alex MR Zhang, Jeremy SW Ko, WX Low and Kay SY Tan. 2016. "Can spiral of silence and civility predict click speech on Facebook?" *Computers in Human Behavior* 64:898–905.

Parkin, Simon. 2014. "Gamergate: A scandal erupts in the video-game community." *The New Yorker* 17.

Rainie, Lee, Janna Anderson and Jonathan Albright. 2017. "The Future of Free Speech, Trolls, Anonymity and Fake News Online.". Available at: ttp://www.pewinternet.org/2017/03/29/the-future-of-free-speech-trolls-anonymity-and-fake-news-online/.

Rapaport, Herman. 2016. "Big Data: Communicating Outside the Medium of Meaning." *symploke* 24(1):447–457.

Rushin, Stephen and Griffin Sims Edwards. 2018. "The Effect of President Trump's Election on Hate Crimes." SSRN Scholarly Paper ID 3102652, Social Science Research Network, Rochester, NY.

Scheufle, Dietram A and Patricia Moy. 2000. "Twenty-five years of the spiral of silence: A conceptual review and empirical outlook." *International journal of public opinion research* 12(1):3–28.

Sherrick, Brett and Jennifer Hoewe. 2018. "The effect of explicit online comment moderation on three spiral of silence outcomes." *New Media & Society* 20(2):453–474.

Siegel, Alexandra. 2015. *Sectarian Twitter Wars: Sunni-Shia Conflict and Cooperation in the Digital Age.* Vol. 20 Carnegie Endowment for International Peace.

Siegel, Alexandra A. 2019. "Online Hate Speech: An Overview of the Literature.".

Siegel, Alexandra, Joshua Tucker, Jonathan Nagler and Richard Bonneau. 2018*a*. "Socially Mediated Sectarianism." *Unpublished Manuscript* .

Siegel, Alexandra, Joshua Tucker, Jonathan Nagler and Richard Bonneau. 2018*b*. "Tweeting Beyond Tahrir: Ideological Diversity and Political Intolerance in Egyptian Twitter Networks." *Unpublished Manuscript* .

Silva, Leandro, Mainack Mondal, Denzil Correa, Fabrício Benevenuto and Ingmar Weber. 2016. ""Analyzing the targets of hate in online social media"." *Unpublished Manuscript* . Available at: https://arxiv.org/abs/1603.07709v1.

Sonnad, Nikhil. 2016. "Alt-right trolls are using these code words for racial slurs online." *Quartz* .

SPLC. 2017. "The Year in Hate and Extremism." *Southern Poverty Law Center* . Available at: https://www.splcenter.org/fighting-hate/intelligence-report/2017/year-hate-and-extremism.

Stack, Liam. 2016. "Trump's victory alarmes gay and transgender groups.".

Tuckwood, Christopher. 2014. "The State of the Field: Technology for Atrocity Response." *Genocide Studies and Prevention: An International Journal* 8(3):9.

Tuckwood, Christopher. 2017. "Hatebase: Online Database of Hate Speech." *The Sentinal Project* . Available at: https://www.hatebase.org/.

Van Dijk, Teun A. 1992. "Denying racism: Elite discourse and racism.".

Walker, Hunter. 2015. "Donald Trump just released an epic statement raging against Mexican immigrants and 'disease'.". Available at: http://www.businessinsider.com/donald-trumps-epic-statement-on-mexico-2015-7.

Warner, William and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*. Association for Computational Linguistics pp. 19–26.

Waseem, Zeerak and Dirk Hovy. 2016. "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter." *Proceedings of NAACL-HLT* pp. 88–93.

Wells, Chris, Dhavan V Shah, Jon C Pevehouse, JungHwan Yang, Ayellet Pelled, Frederick Boehm, Josephine Lukito, Shreenita Ghosh and Jessica L Schmidt. 2016. "How Trump drove coverage to the nomination: Hybrid media campaigning." *Political Communication* 33(4):669–676.

Zaller, John. 1992. *The nature and origins of mass opinion*. Cambridge university press.

Zaller, John. 1994. "Elite leadership of mass opinion." *Taken by storm: The media, public opinion, and US foreign policy in the Gulf War* pp. 186–209.

Zannettou, Savvas, Barry Bradlyn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, Haewoon Kwak and Jeremy Blackburn. 2018. "What is Gab? A Bastion of Free Speech or an Alt-Right Echo Chamber?" *Unpublished Manuscript* .